Oversight Board

Public Comment Appendix for

2024-007-IG-UA, 2024-008-FB-UA

Case number

Case description

These cases concern two content decisions made by Meta, one on Instagram and one on Facebook, which the Oversight Board intends to address together. For each case, the Board will decide whether the content should be allowed on Instagram or Facebook.

The first case involves an AI-generated image of a nude woman posted on Instagram. The image has been created using artificial intelligence (AI) to resemble a public figure from India. The account that posted this content only shares AI-generated images of Indian women. The majority of users who reacted have accounts in India, where deepfakes are increasingly a problem.

In this case, a user reported the content to Meta for pornography. This report was automatically closed because it was not reviewed within 48 hours. The same user then appealed Meta's decision to leave up the content but this was also automatically closed and so the content remained up. The user then appealed to the Board. As a result of the Board selecting this case, Meta determined that its decision to leave the content up was in error and removed the post for violating the Bullying and Harassment Community Standard.

The second case concerns an image posted to a Facebook group for AI creations. It features an AI-generated image of a nude woman with a man groping her breast. The image has been created with AI to resemble an American public figure, who is also named in the caption. The majority of users who reacted have accounts in the United States.

In this case, a different user had already posted this image, which led to it being escalated to Meta's policy or subject matter experts who decided to remove the content as a violation of the Bullying and Harassment policy, specifically for "derogatory sexualized photoshop or drawings." The image was added to a Media Matching Service Bank – part of Meta's automated enforcement system that automatically finds and removes images that have already been identified by human reviewers as breaking Meta's rules. Therefore, in this case, the image was already considered a violation of Facebook's Community Standards and removed. The user who posted the content appealed but the report was automatically closed. The user then appealed to the Board.

The Board selected these cases to assess whether Meta's policies and its enforcement practices are effective at addressing explicit AI-generated imagery. This case aligns with the Board's Gender strategic priority.

The Board would appreciate public comments that address:

- The nature and gravity of harms posed by deepfake pornography including how those harms affect women, especially women who are public figures.
- Contextual information about the use and prevalence of deepfake pornography globally, including in the United States and India.
- Strategies for how Meta can address deepfake pornography on its platforms, including the policies and enforcement processes that may be most effective.
- Meta's enforcement of its "derogatory sexualized photoshop or drawings" rule in the Bullying and Harassment policy, including the use of Media Matching Service Banks.
- The challenges of relying on automated systems that automatically close appeals in 48 hours if no review has taken place.

As part of its decisions, the Board can issue policy recommendations to Meta. While recommendations are not binding, Meta must respond to them within 60 days. As such, the Board welcomes public comments proposing recommendations that are relevant to these cases.

Oversight Board

Public Comment Appendix for
2024-007-IG-UA, 2024-008-FB-UA
Case number

The Oversight Board is committed to bringing diverse perspectives from third parties into the case review process. To that end, the Oversight Board has established a public comment process.

Public comments respond to case descriptions based on the information provided to the Board by users and Facebook as part of the appeals process. These case descriptions are posted before panels begin deliberation to provide time for public comment. As such, case descriptions reflect neither the Board's assessment of the case, nor the full array of policy issues that a panel might consider to be implicated by each case.

To protect the privacy and security of commenters, comments are only viewed by the Oversight Board and as detailed in the Operational Privacy Notice. All commenters included in this appendix gave consent to the Oversight Board to publish their comments. For commenters who did not consent to attribute their comments publicly, names have been redacted. To withdraw your comment, please email contact@osbadmin.com.

To reflect the wide range of views on cases, the Oversight Board has included all comments received except those clearly irrelevant, abusive or disrespectful of the human and fundamental rights of any person or group of persons and therefore violating the Terms for Public Comment. Inclusion of a comment in this appendix is not an endorsement by the Oversight Board of the views expressed in the comment. The Oversight Board is committed to transparency and this appendix is meant to accurately reflect the input we received.

2024-007-IG-UA, 2024-008-FB-UA	PC-27001	Latin America & Caribbean
Case number	Public comment number	Region
Withheld	Withheld	English
Commenter's first name	Commenter's last name	Commenter's preferred language
Withheld		No
Organization		Response on behalf of organization

Full Comment

Beyond being a violation of the Bullying and Harassment policy, specifically for derogatory sexualized photoshop or drawings, we see that there is a gap in the Adult Nudity and Sexual Activity policy. Despite including uncovered female nipples, it does not provide guidance regarding digital art; only in sexual activities would the digital art be controlled. There is also a gap in Sexual Exploitation, considering that AI can not only be used to create art but also deepfakes and sexual deepfakes. So, in this sense, we have to keep in mind that these creations through AI can be used for perversion, sexually objectifying the body of the person chosen for this type of content, whether for revenge or not. Generally, this is carried out in a non-consensual and derogatory manner and may cause psychological harm in the short and long term.

Link to Attachment

No Attachment

2024-007-IG-UA, 2024-008-FB-UA	PC-27002	United States & Canada
Case number	Public comment number	Region
Gregory	Stanton	English
Commenter's first name	Commenter's last name	Commenter's preferred language
Genocide Watch		Yes
Organization		Response on behalf of organization

Full Comment

Artificial Intelligence generated images and texts should not be permitted in any form on Facebook, Instagram, WhatsApp, or any other META owned platforms. These platforms are for the use of human beings, not for machines. Incitements to genocide, rape, suicide, sexual abuse of children and other crimes are already a growing problem on these platforms. Incitements to crime are not protected speech. META urgently needs to develop AI tools to instantly remove AI generated fakes.

Link to Attachment

No Attachment

2024-007-IG-UA, 2024-008-FB-UA	PC-27003	United States & Canada
Case number	Public comment number	Region
Lynn	Patsiga	English
Commenter's first name	Commenter's last name	Commenter's preferred language
DID NOT PROVIDE		No
Organization		Response on behalf of organization

Full Comment

Fraudulent AI images which depict an individual in a dishonest manner need to be removed. The individual or group responsible for creating or disseminating the fake image should face a Meta penalty (removal of account for a period of time, etc...).

Link to Attachment

No Attachment

2024-007-IG-UA, 2024-008-FB-UA	PC-27004	Europe
Case number	Public comment number	Region
Per Svein	Hansen	English
Commenter's first name	Commenter's last name	Commenter's preferred language
DID NOT PROVIDE		No
Organization		Response on behalf of organization

Full Comment

Nudity is not to be concidered offensive per s. AI pictures with resemblance of actual individuals, must not be allowed without permission from that individual and under no circumstances engaging in pornographic sexual activity.Pictures of nude persons shall not display underage minors or such persons assumed to appear under legal age of consent.

Link to Attachment

No Attachment

2024-007-IG-UA, 2024-008-FB-UA	PC-27005	United States & Canada
Case number	Public comment number	Region
Withheld	Withheld	English
Commenter's first name	Commenter's last name	Commenter's preferred language
Withheld		No
Organization		Response on behalf of organization

Full Comment

The research already shows deepfake abuse is extremely traumatizing and damaging. Allowing this on Meta's platforms would be detrimental for democracy. This case isn't about celebrities, it's about everyday people. These deepfake sexual abuse applications are now accessible enough that middle schoolers can user them (as they have). If there are no swift policies from platforms like Meta to prevent these abusive materials from being posted publicly, this WILL silence women. A woman running for office or for a corporate position, anyone in or out of the public eye, could be harmed by deepfake abuse. Deepfake abuse is a smoking gun, a weapon that could be used to silence and remove women from public spaces. On a broader scale, women and girls could be afraid to simply exist in the public eye for fear that any random person could create and post deepfake abuse of them.

Link to Attachment

No Attachment

2024-007-IG-UA, 2024-008-FB-UA	PC-27006	United States & Canada
Case number	Public comment number	Region
Withheld	Withheld	English
Commenter's first name	Commenter's last name	Commenter's preferred language
Withheld		Yes
Organization		Response on behalf of organization

Full Comment

To even have to answer such a question explains the state of affairs in mass communication. At no time anywhere is it appropriate to spread AI generated nude pictures, or for that matter, nude pictures period on a public platform. There will be those that say it is an artform and rightly so. But artists are not expressing contempt for a person or attempting to ridicule and adversely affect someone. This not only goes for women. It applies to men also. If these media platforms cannot eliminate the possibility of children encountering nudity meant to denigrate a man or woman, then the platforms need to be censured, fined and, should repeat offenses occur, they need to be shut down long enough for them to learn how to police themselves. Society has to start setting some standards.

Link to Attachment

No Attachment

2024-007-IG-UA, 2024-008-FB-UA	PC-27007	United States & Canada
Case number	Public comment number	Region
Michael	Snell	English
Commenter's first name	Commenter's last name	Commenter's preferred language
DID NOT PROVIDE		No
Organization		Response on behalf of organization

Full Comment

No pornography should be tolerated. I would think this is the current standard. Whether the porn is authentic or not does not matter.

Link to Attachment

No Attachment

2024-007-IG-UA, 2024-008-FB-UA

Case number

PC-27008

Public comment number

United States & Canada

Region

Clifford

Commenter's first name

Leyba II

Commenter's last name

English

Commenter's preferred language

DID NOT PROVIDE

Organization

No

Response on behalf of organization

Full Comment

I feel that the security to our children is most imporantant but pepole should have the right to know whats going on in the background post like these should be on an adult app

Link to Attachment

No Attachment

2024-007-IG-UA, 2024-008-FB-UA	PC-27009	Central & South Asia
Case number	Public comment number	Region
muhammad	iftikhar	English
Commenter's first name	Commenter's last name	Commenter's preferred language
ministery of education		Yes
Organization		Response on behalf of organization

Full Comment

gratitude to the oversight board for giving me an oppertunity to express my views on the given AI issue. Te said issue is highly condemnabe and obnoxuous. women should be esteemed and they should not be humiliated just like that. Such activities must be restricted as it violates individual secrecy and confereracy. cases like these must not be propagated and shared

Link to Attachment

No Attachment

2024-007-IG-UA, 2024-008-FB-UA

Case number

PC-27010

Public comment number

United States & Canada

Region

Withheld

Commenter's first name

Withheld

Commenter's last name

English

Commenter's preferred language

Withheld

Organization

No

Response on behalf of organization

Full Comment

Regulating these sorts of deepfakes is a fools errand, especially when the people depicted are public figures.

Link to Attachment

No Attachment

2024-007-IG-UA, 2024-008-FB-UA	PC-27011	United States & Canada
Case number	Public comment number	Region
Withheld	Withheld	English
Commenter's first name	Commenter's last name	Commenter's preferred language
Withheld		No
Organization		Response on behalf of organization

Full Comment

Deep fake or AI-generated imagery (AI) is unguided and untested territory. AI pornographic images are deeply harmful, especially when they are generated with the features of real persons or by altering digital images to resemble real persons. This is true whether the person is a public figure or a private person. These nonconsensual images serve no value other than to degrade, harass, humiliate, and bully. The images can harm mental health, reputations and physical safety as well as pose risks to friends, families, and career prospects. Sharing of disturbing and volatile images, including pornographic imagery, in a public forum such as FaceBook or Instagram is an appalling misuse of the platforms. If Meta proceeds with the publication of such imagery, then Meta must also agree to publish (at the same time) the legally verified full name and country of the creator AND publisher so the harmed person has reasonable recourse to respond. No hiding behind false names.

Link to Attachment

No Attachment

2024-007-IG-UA, 2024-008-FB-UA	PC-27012	United States & Canada
Case number	Public comment number	Region
Kevin	Hodge	English
Commenter's first name	Commenter's last name	Commenter's preferred language
DID NOT PROVIDE		No
Organization		Response on behalf of organization

Full Comment

Why can Meta not detect A.I.?

Link to Attachment

No Attachment

2024-007-IG-UA, 2024-008-FB-UA	PC-27013	United States & Canada
Case number	Public comment number	Region
Lora	Premo	English
Commenter's first name	Commenter's last name	Commenter's preferred language
DID NOT PROVIDE		No
Organization		Response on behalf of organization

Full Comment

Meta should have a simple rule -- no AI generated graphics or documents. They're easily detected by computers so they could have an automated system set up to crawl through their websites and pull all the AI generated images etc. I don't care if that means a few funny memes just aren't gonna make it out there. The fact that Meta rarely responds to complaints and rarely removes anything based on a complaint means that they need to automate this and there needs to be no exceptions. It's their own fault for showing that they have no interest whatsoever in actually doing the right thing. The number of examples of instances were it took Facebook years to pull down some thing that was obviously against their rules means that they don't care at all about any rules that are established and no matter what rules you establish they're not going to follow them. This is why we need laws to control these technology companies. That's a whole different conversation but I hope and pray Congress Will someday regulate the horror that is Meta and force them to enforce their own rules. I don't care how much META has invested in AI. The fact is most of these deepfake images are harmful even if they aren't pornographic. A 19-year-old boy was telling me about a Facebook image showing Joe Biden appearing to molest a child. Yes indeed. Bet that hasn't been taken down either. And he was fully invested in the idea that it was true -- so there is no such thing as a deepfake image that's not harmful, because it's all lies, bottom line, whether it's an image or a document that's been generated it's complete artificial nonsense and I don't think we should have huge websites full of fake data that people are likely to believe. We've already seen it destroy our democracy when Facebook and Twitter didn't take anything down in regards to all the lies being promulgated by Trump supporters during Covid and all the rest of that. Nobody has ever made any effort to ensure that the information appearing on these websites is in any way accurate and I think that one big step in the right direction would be no AI generated anything. Until AI becomes an established and trustworthy technology, it has no business anywhere but in laboratories and other experimental situations. Since it's perfectly easy to crawl the website and pull all the AI generated stuff down, or prevented from being posted in the first place. Takes no trouble at all, and it would be very quick. Tough luck if people don't like it. Influencers have lived without it all this time, they can live without it now. I think that would be an outstanding way of fixing this problem because there is no other way. Even if Meta Was genuinely interested in pulling down deep fakes which they have shown no sign of being serious about -- It would be impossible to do Manually. Meta Can't possibly hire enough people to adjudicate every individual image. So that's the only choice -- automating it and with all the bragging and all the nonsense I hear constantly from Meta, They shouldn't have the slightest difficulty automating this and letting us get on with all our lives instead of making it into an issue that ends up in front of the Supreme Court or something. They have been allowed to get away with disgraceful behaviors for far too long and it's time somebody reigned them in. Making AI generated images impossible to post or setting up a web crawler to pull them all down would be a very simple solution to all of these problems. I don't care how invested in AI Meta is. The technology is just not ready for public use because it's a powerful weapon in the wrong hands and nobody is monitoring or enforcing any rules regarding it. It's time to start now. It's no different than nuclear weapons were at one time. Everybody stood around while the radiation fell on their head because they didn't know any better and the weapons were made much too large because nobody knew any better. New technologies need to be studied and experimented with For far longer before they become available for public use. The cat is already out of the bag with AI, but that doesn't mean that massive websites should allow unfettered use of AI imagery and AI documents because they're all nonsense. AI makes things up. Since we know this why would we allow any AI generated anything Ever? A fake is a fake is a fake And it doesn't belong on a public website posing as real information.

Link to Attachment

No Attachment

2024-007-IG-UA, 2024-008-FB-UA

Case number

PC-27014

Public comment number

United States & Canada

Region

Withheld

Commenter's first name

Withheld

Commenter's last name

English

Commenter's preferred language

Withheld

Organization

No

Response on behalf of organization

Full Comment

Explicit AI-generated imagery should be banned in all instances. Simply removing the material is not sanction enough. The user(s) who post this material should have their accounts removed and blocked, or publicly identify them and let them deal with the consequences of their actions. I imagine the people who were depicted in this kind of trash should have grounds for a substantial lawsuit.

Link to Attachment

No Attachment

2024-007-IG-UA, 2024-008-FB-UA

Case number

PC-27015

Public comment number

Central & South Asia

Region

Withheld

Commenter's first name

Withheld

Commenter's last name

English

Commenter's preferred language

Withheld

Organization

No

Response on behalf of organization

Full Comment

Deepfake pornography poses significant harms, especially to women, including public figures. These harms include the potential for defamation, harassment, and harm to reputation. The impact on women, particularly those in the public eye, can be severe, leading to emotional distress, damage to personal and professional relationships, and even threats to physical safety. In countries like India, the issue of deepfakes is particularly concerning, given cultural sensitivities and the potential for widespread harm.To effectively combat deepfake pornography, Meta must adopt a comprehensive strategy that goes beyond mere enforcement. Yes, policies and algorithms play a crucial role, but equally important are the human elements empathy, understanding, and a commitment to protecting users' dignity and rights. Proactive measures to detect and remove such content, as well as mechanisms for users to report and appeal decisions are important. Education and awareness campaigns can also help users identify and mitigate the impact of deepfake content. Clear guidelines and training for moderators are essential to ensure consistent and fair enforcement of already existing policies. Relying solely on automated systems for content moderation poses significant challenges, as is well attested by the automatic closure of appeals in the case presented. This is limiting and should be acknowledged. This can further result in errors and injustices, particularly in cases involving sensitive content like deepfake pornography. Meta should invest in improving its review processes to prevent wrongful removals and closures.Overall, addressing deepfake pornography requires a multifaceted approach, combining technological solutions with policy reforms and community engagement. Meta must prioritize the protection of users, especially women, from the harmful effects of deepfake content while upholding principles of free expression and privacy.

Link to Attachment

No Attachment

2024-007-IG-UA, 2024-008-FB-UA	PC-27016	Europe
Case number	Public comment number	Region
Jurriaan	Daems	English
Commenter's first name	Commenter's last name	Commenter's preferred language
DID NOT PROVIDE		No
Organization		Response on behalf of organization

Full Comment

It is understandable if we know that our governments loves to sexually confuse our children in schools, and even promote abortion and euthanasia.

Link to Attachment

No Attachment

2024-007-IG-UA, 2024-008-FB-UA	PC-27017	United States & Canada
Case number	Public comment number	Region
Withheld	Withheld	English
Commenter's first name	Commenter's last name	Commenter's preferred language
Withheld		No
Organization		Response on behalf of organization

Full Comment

I have a very simple but sure solution ban all nudity on the Meta platform, even real paintings. You are a private platform and this broad application solution would be easier to enforce than to pick and choose what is or isnt appropriate. Just ban nudity from the dcollet down. If people want to see nudity, there are other all ready established venues and publications offering it.

Link to Attachment

No Attachment

2024-007-IG-UA, 2024-008-FB-UA	PC-27018	United States & Canada
Case number	Public comment number	Region
Lynn	Hinkins	English
Commenter's first name	Commenter's last name	Commenter's preferred language
DID NOT PROVIDE		No
Organization		Response on behalf of organization

Full Comment

Our group is a Roller Skating group, when I set up a chat page Meta fills it with Porn . This is a family group where adults share what they did growing up & a chance to bring old friends together. Meta should not put porn anywhere they want but rather a special page as the children dont need to see these cartoons or photos.Mimicombo Memories is the group these photos are showing up on & I would like to see it STOP . I had to close the chat both times as meta kept putting it up .

Link to Attachment

No Attachment

2024-007-IG-UA, 2024-008-FB-UA	PC-27019	United States & Canada
Case number	Public comment number	Region
Withheld	Withheld	English
Commenter's first name	Commenter's last name	Commenter's preferred language
Withheld		No
Organization		Response on behalf of organization

Full Comment

The AI generated deep fakes shouldnt just focus on especially celebrity women but ALL women.

Link to Attachment

No Attachment

2024-007-IG-UA, 2024-008-FB-UA	PC-27020	Europe
Case number	Public comment number	Region
Carolina	Are	English
Commenter's first name	Commenter's last name	Commenter's preferred language
DID NOT PROVIDE		No
Organization		Response on behalf of organization

Full Comment

Explicit AI Images of Female Public Figures Public comment response (Carolina Are) This public comment addresses the cases of explicit AI images of women in the public eye raised by the Oversight Board. In line with my area of expertise platform governance of sex, nudity, and sex work, as well as online harms my public comment will raise three key issues arising from this case in connection with Metas content moderation: 1) inaccurate grouping and enforcement 2) double standards in content moderation and 3) authorship. #1 Inaccurate grouping and enforcementThe cases in question highlight the failures of Metas obsession with removing pornography, nudity, sex work and sexual expression post-FOSTA/SESTA (see: Are & Paasonen, 2021). Indeed, the fact that a user had to report the content for pornography and not for image-based abuse or impersonation shows the limited options users have to defend themselves when it comes to issues generated by deepfake or the non-consensual sharing of intimate images. Non-consensual deepfakes or leaked nudes are not pornography, which is a form of entertainment and/or expression that is consensually made and shared. Non-consensual deepfakes or leaked nudes are abusive material, and Meta should have specific avenues and categories to distinguish them from consensual modes of entertainment and sexual expression, since one-size-fits-all approaches harm victims as well as over-censoring those who share this material consensually. #2 Double standards in content moderationA second, crucial issue raised by these cases is the double standard in content moderation of nudity, showing discrepancies in moderations between when this content is posted consensually by sex workers or sex positive accounts and when this is posted by malicious actors as a form of abuse.The fact that reports for pornography had not been actioned against an account that repeatedly posted explicit, abusive images of women mirrors the experiences reported by the sex workers and sex positive accounts I have interviewed as part of my research (see: Are, 2023; 2024; Are et al., 2023; Are & Paasonen, 2021). While users and particularly women and LGBTQIA+ folks - consensually posting nudity, sex work and sexual expression post-FOSTA/SESTA are immediately de-platformed and shadowbanned, celebrities, or even accounts stealing their images and impersonating them, are often left up by Metas moderation even after repeated reports. It's of course difficult to uncover the reason behind this if community guidelines truly apply to all users, but one cant help but wonder if certain profile elements trigger algorithms differently e.g., in 2019/2020 users told me that changing their profile gender to male helped them face less strict moderation of nudity (Are, 2022). Because of these user conjectures, and because of these content moderation double standards, more transparency about internal moderation guidelines and their application is needed. Further, Meta should invest in dedicated customer service for victims of deepfake and/or image-based abuse, in order to apply nuance and care in a situation where automated appeals are clearly falling short of understanding the context.#3 AuthorshipLast but not least, a less discussed but vital issue to consider in the realm of deepfakes is that their victims are not solely the (too often) women whose faces are used to create these images. Indeed, the sex workers whose online content is often used to train AI models and/or to create this imagery also find their consent being violated, unwillingly becoming involved in schemes to harm other women. Too often, when discussing violence against women and girls, sex workers are not included in the conversation: as argued by Decrim Now and the UK Sex Working Union during the virtual Challenging Sex Worker Invisibility: Actioning the Case for Sexual Services Protections in the Online Safety Act 2023 conference, the violation of sex workers consent and authorship is missing from discussions around deepfakes. As such, impersonation and copyright violations should be created and enforced towards a better-rounded strategy to tackle deepfakes, given that sex workers have a right to earn an income from their online images, and should have authorship and ownership rights to their content so that it isnt used to harm others. More resources: Alptraum L (2020)

Deepfake Porn Harms Adult Performers, Too. Wired. https://www.wired.com/story/deepfake-porn-harms-adult-performers-too/. Are C.(2024) Flagging as a silencing tool: Exploring the relationship between de-platforming of sex and online abuse on Instagram and TikTok. New Media & Society, 0(0). https://doi.org/10.1177/14614448241228544. Are C (2023) The assemblages of flagging and de-platforming against marginalised content creators. Convergence: 0(0). https://doi.org/10.1177/13548565231218629Are C (2022) The Shadowban Cycle: an autoethnography of pole dancing, nudity and censorship on Instagram. Feminist Media Studies 22: 20022019.Are C, Collingham H, Carrothers AM & Fox E (2023) Co-designing platform governance policies Tackling malicious flagging and de-platforming with impacted social media users. Centre for Digital Citizens. https://digitalcitizens.uk/blog/platform_governance_inequalities/ Are C, Paasonen S (2021) Sex in the shadows of celebrity. Porn Studies Forum 8(4): 411419.Sojit Pecha (2023) Deepfake porn isnt just a consent issue, its a labor issue. Document. https://www.documentjournal.com/2023/02/twitch-streamer-deepfake-controversy-ai-porn-sex-work-labor-technology/.

Link to Attachment

PC-27020

2024-007-IG-UA, 2024-008-FB-UA	PC-27021	United States & Canada
Case number	Public comment number	Region
Withheld	Withheld	English
Commenter's first name	Commenter's last name	Commenter's preferred language
Withheld		No
Organization		Response on behalf of organization

Full Comment

To Whom,   I believe you all are wasting your time with these so called corporate meetings. My suggestion" Give us the ability to PERMANENTLY delete what we decide is an offensive photo. I personally woulddelete the whole dang thing being every one of them pics is offensive as heck. What is really offensive to me is that I don't have a choice to delete or not to delete which should be my choice. :0(

Link to Attachment

No Attachment

2024-007-IG-UA, 2024-008-FB-UA	PC-27023	United States & Canada
Case number	Public comment number	Region
Tina	Grantham	English
Commenter's first name	Commenter's last name	Commenter's preferred language
DID NOT PROVIDE		No
Organization		Response on behalf of organization

Full Comment

They shouldn't be using bots to investigate claims of indecent posts, computers can't determine what is and isn't porn, they need humans to determine that, but the will continue using bots because they don't have to pay for their time. Links with explicit images and spam should never be allowed, I have ads in my fb stories of literal porn and nothing is done about it, because the image is linked to some website that's a phishing website, spam shouldn't be allowed on these sites, I post a comment without an emoji and my whole account gets flagged as spam even though I've never posted spam content in my whole time using social media. But these phishing accounts continue to get away with it, look at the comments of any type of public news story, it's full of phishing links that look like YouTube links but they're not. I got a 30 ban for posting a picture of my SON playing in water during the summer and they stood by their decision claiming I posted child porn, because HIS chest was exposed, make it make sense.

Link to Attachment

No Attachment

2024-007-IG-UA, 2024-008-FB-UA	PC-27025	United States & Canada
Case number	Public comment number	Region
Billie	Maquet	English
Commenter's first name	Commenter's last name	Commenter's preferred language
DID NOT PROVIDE		No
Organization		Response on behalf of organization

Full Comment

All your use of AI has been a disaster. It flags things it shouldn't and prevents sharing of things that are approved. Go back to using humans as AI is a failure.

Link to Attachment

No Attachment

2024-007-IG-UA, 2024-008-FB-UA	PC-27026	Europe
Case number	Public comment number	Region
David	Wright	English
Commenter's first name	Commenter's last name	Commenter's preferred language
SWGfL		Yes
Organization		Response on behalf of organization

Full Comment

DID NOT PROVIDE

Link to Attachment

PC-27026

2024-007-IG-UA, 2024-008-FB-UA	PC-27028	Asia Pacific & Oceania
Case number	Public comment number	Region
Withheld	Withheld	English
Commenter's first name	Commenter's last name	Commenter's preferred language
Withheld		No
Organization		Response on behalf of organization

Full Comment

Public Comment1: AI-Generated Nude Image of Indian Public Figure on InstagramThis case involves the non-consensual creation and sharing of an explicit AI-generated image depicting a public figure from India in a nude and sexualized manner.Such content poses significant harms to the individual's privacy, dignity, and reputation, potentially leading to psychological distress, harassment, and real-world consequences.The targeting of a public figure from India is particularly concerning, as deepfake pornography is an increasing problem in the country, exacerbating existing gender-based inequalities and perpetuating harmful stereotypes.The account sharing these AI-generated images primarily targets Indian women, indicating a concerning trend of non-consensual sexual exploitation through deepfake technology in the Indian context.Therefore the majority of users engaging with the content having accounts in India further highlights the localized nature of this issue and the need for region-specific interventions.Strategies for Meta:Meta should enforce robust content moderation policies that explicitly prohibit the upload and sharing of non-consensual deepfake pornographic content, prioritizing the protection of individuals' privacy and consent.- Proactive detection and removal mechanisms, including advanced machine learning algorithms and human review processes, should be implemented to identify and remove such content before it can spread widely.- User reporting mechanisms should be streamlined, and prompt review of reports should be ensured to address instances of deepfake pornography in a timely manner.Challenges with Automated Appeal Closure Systems:- In this case, the initial report and subsequent appeal regarding the deepfake image were automatically closed without proper review, leading to the content remaining up despite violating Meta's policies.- This highlights the limitations of relying solely on automated systems and short timeframes (e.g., 48 hours) for appeal closures, as deepfake pornography cases often require nuanced human review and consideration of contextual factors.- A balanced approach involving both automated systems and dedicated human review teams is recommended to ensure fair and accurate decisions while effectively addressing the harms of deepfake pornography.2: AI-Generated Nude Image of American Public Figure on Facebook- This case involves the non-consensual creation and sharing of an explicit AI-generated image depicting an American public figure in a sexualized and demeaning manner, with a man groping her.- Such content constitutes a severe violation of the individual's privacy, dignity, and consent, potentially causing significant psychological distress and reputational damage.- The targeting of a public figure in this manner normalizes and perpetuates the objectification and exploitation of women, reinforcing harmful gender-based stereotypes and power dynamics.- The presence of this content on a Facebook group dedicated to AI creations highlights the need for heightened awareness and regulation surrounding the misuse of AI technology for non-consensual purposes.- The majority of users engaging with the content having accounts in the United States indicates the broader societal implications and the need for effective measures to combat deepfake pornography in the American context.Strategies for Meta:- Meta's decision to remove the content for violating the "Bullying and Harassment" policy, specifically the "derogatory sexualized photoshop or drawings" provision, is a positive step in addressing such harmful content.- The use of Media Matching Service Banks (MMS Banks) to automatically detect and remove known instances of non-consensual intimate imagery is a valuable tool in Meta's arsenal against deepfake pornography.- However, continuous updating and expansion of these databases, as well as robust human review processes, are crucial to ensure comprehensive coverage and accurate enforcement.Challenges with Automated Appeal Closure Systems:- While the content was initially identified and removed correctly, the subsequent automatic closure of the user's appeal without proper review raises concerns about due process and the potential for legitimate appeals to be overlooked.- As with the previous case, relying solely on automated systems for appeal closures can lead to inadequate consideration of nuanced factors, potentially undermining the effectiveness of Meta's efforts to combat deepfake pornography.

Link to Attachment

PC-27028

2024-007-IG-UA, 2024-008-FB-UA	PC-27029	Central & South Asia
Case number	Public comment number	Region
Rakesh	Maheshwari	English
Commenter's first name	Commenter's last name	Commenter's preferred language
DID NOT PROVIDE		No
Organization		Response on behalf of organization

Full Comment

DID NOT PROVIDE

Link to Attachment

PC-27029

2024-007-IG-UA, 2024-008-FB-UA	PC-27031	Central & South Asia
Case number	Public comment number	Region
Himanshu	Gupta	English
Commenter's first name	Commenter's last name	Commenter's preferred language
Karuna Shakti Foundation		Yes
Organization		Response on behalf of organization

Full Comment

Regarding : Explicit AI Images of Female Public Figures. This could be decided by brainstorming on the following pointers1. How would I feel and what would i do if come across such images of females(mother, sister, wife, girl friend etc) from my family?2. What could be the impact on the children and youth (age 6 to 35 yrs) if they come across such images of females from their or outside their family?3. What positive change it could bring in the society if such images are floated in the public?4. Would allowing such images be of benefit to anyone except anti social elements?I think answering these questions would help us/you in coming to the decision.

Link to Attachment

No Attachment

2024-007-IG-UA, 2024-008-FB-UA	PC-27032	Central & South Asia
Case number	Public comment number	Region
Siddharth	Pillai	English
Commenter's first name	Commenter's last name	Commenter's preferred language
RATI Foundation for Social Change		Yes
Organization		Response on behalf of organization

Full Comment

- The nature and gravity of harms posed by deepfake pornography including how those harms affect women, especially women who are public figures.Till date Mumbai based RATI foundation has reported 96 cases to primarily META where AI Tool clothoff.io was used to generate deepfake CSAM/NCII.Victims: Most victims were young female local language influencers or micro-influencers. There were a few mainstream actresses and some contestants of the popular reality Indian TV show 'Big Boss'. We have at least one confirmed victim who was a minor. Many of the victimswere very young. We have 2 victims who were male. The youngest victim was a 15 year old school going girl followed by a 16 year old boy. The oldest was in her late 30s.Victims feel shamed and violated. Victims risk social stigma and frequently the ask to ensure that the content is removed before family member witness or become aware of the content. Victims fear that parents will curtail thier online presence if the content is found by them. They also panic when they find reuploads and copies. In a sense, the effect of the violation on the victim is not very different from encountering genuine nudes.The term public figure as being used in this case is vague. The 15 year old victim who reported our first case of AI manipulated CSAM/NCII had few followers. But her reels were starting to get popular and one of the popular reels was manipulated.Perpetrators: Perpetrators were typically young males. They self-identified as 'trolls' or meme-ers' (internet edgelords). They branded their account using a specific logo that would be the account display picture and even double as a thumbnail. This made it difficult to spot which post was the one that featured deep fake nudity. However, our teams soon realized that the post with the most number of opens/clicks were likely to be ones featuring nudity.There was a pattern of misogyny and shaming across all their posts, even the ones that did not feature nudity. The AI posts were an attempt to sexually shame the victims as opposed to sexual voyeurism. Even the non AI content sometimes featured the same women where theywere being ridiculed and discredited.The Content: The content would be a reupload of the victim's popular reels. The content would be overlayed with a logo in the middle. A similar logo would drop down in slow motion from the top part of the reel. When the two logos would overlap each other, there would be aquick flash of the deep fake nude content. The viewer can only be exposed to the nudity by clicking on the reel at the right second. In most cases this would require sustained engagement and multiple clicks with the reel. Off the reels reported, there were some previouslyreported copies and duplicates. We have also come across a few cases where a deepfake was made and uploaded on group chat to bully a victim as well as on DMs to scare and coerce a victim.Problematic Content we came across but did not report: Completely fabricated AI imagery in Incest/Cuckold fantasies (featuring mothers, sisters) as it was disturbing but was difficult to say that it needed actioning.- Contextual information about the use and prevalence of deepfake pornography globally, including in the United States and India.Keywords: Key words such as "up down troll" and "pause challenge" will tend to throw up deep fake content. The manner in which deep fake AI has proliferated NCII/CSAM on meta platforms is unprecedented.Deepfake Generator: Clothoff.io is the most popular deep fake generator as is evidenced from the watermark. Both the telegram bot and website are being used.- Strategies for how Meta can address deepfake pornography on its platforms, including the policies and enforcement processes that may be most effective.Disrupt Organized Activity: There is a need to identify and disrupt evident patterns such as 'pause challenge' and 'up down troll'. Up down troll was reported in December 2023. Pause Challenge in early April. No concerted action has been taken to curb such organized efforts.Actioning Collaborator Accounts: A continuation of the above point. Many of these videos are posted in collaboration with another account. However, when the post is actioned only one of the accounts is penalized. The other account which is an alt account of the offender survives and it resumes posting.Strengthening Prevention of Reupload & Removal of Copies: Many of reels uploaded were copies. Reels that are already actioned are currently still live at other links. Strengthening Prevention of Reupload, Removal of existing copies, (Youtube does this), and building tolerance for manipulation in hash detection.Strengthening Detection of Nudity within Reels: Perpetrators are increasingly placing nude footage or AI deepfakes 5 or 10 seconds into the reel. The offending footage is visible only for a second before the reel resumes. Systems must detect such content. Else please give reporters the ability to report timestamps.Create Reporting Category for Digitally Manipulated Sexual Abuse Imagery about myself/about others that covers nude/erotic/partially-nude deepfake contentCreate an option for Creators to Self-Declare Ai assisted Enhancements/Manipulations & Disincentivize creators who are caught not attempting to evade decalaration.Faster Actioning of Reports & Widening of Meta's most severe category to include Digitally Manipulated resulting in a zero tolerance approach.- Metas enforcement of its derogatory sexualized photoshop or drawings rule in the Bullying and Harassment policy, including the use of Media Matching Service Banks.We have not reported under this rule ever. So we are unaware of this provision. Our standard email for reporting deepfake imagery check the following community guidelines:"We dont allow nudity on Instagram. This includes photos, videos, and some digitally-created content that show sexual intercourse, genitals, and close-ups of fully-nude buttocks. It also includes some photos of female nipples"criminal content: We have zero tolerance when it comes to sharing sexual content involving minors or threatening to post intimate images of others.& community: We remove content that contains credible threats or hate speech, content that targets private individuals to degrade or shame them, personal information meant to blackmail or harass someone, and repeated unwanted messages.It also goes against the wildlife exploitation guidelines which state that: "...identify and take action on photos or videos that violate our community guidelines, such as posts depicting animal abuse."- The challenges of relying on automated systems that automatically close appeals in 48 hours if no review has taken place.This is an unfairsystem that creates an element of luck and lottery in a system which should ensure equitability. Suppose I report content about a deepfake while a riot is taking place nearby. The riot related reports should take precedence but once the riot passes, my reportshould be considered. In case something cannot be looked at within 48 hrs, some system may be considered that shows reporter the likelihood of timeline between reporting and receiving a response.

Link to Attachment

No Attachment

2024-007-IG-UA, 2024-008-FB-UA	PC-27033	Asia Pacific & Oceania
Case number	Public comment number	Region
Shailja Vikram	Singh	English
Commenter's first name	Commenter's last name	Commenter's preferred language
CyberVista Insights - Regulatory Policy & Compliance Advisory		Yes
Organization		Response on behalf of organization

Full Comment

The cases concerning AI-generated nude images on Meta's platforms reveal troubling inconsistencies in the company's content moderation practices. While similar instances of potentially harmful content were handled differently, with one being removed only after intervention by the Oversight Board, the other was swiftly taken down based on Meta's policy. This inconsistency raises concerns about Meta's commitment to fair and consistent moderation, especially regarding sensitive issues like nudity and harassment. It also underscores the need for Meta to ensure transparency and accountability in its content moderation processes, aligning them with local laws and community standards.Moreover, the automatic closure of user appeals without review demonstrates a lack of accountability on Meta's part and neglects the responsibility to protect users from abusive or inappropriate material. This practice erodes user trust and fails to address legitimate concerns raised by users regarding potentially harmful content. To address these issues, Meta must improve its grievance redressal mechanisms, ensuring timely review of user reports and appeals, with transparent communication about decisions made.Regulatory bodies also play a crucial role in addressing these challenges. They should investigate Meta's content moderation practices to ensure compliance with local laws and regulations, particularly regarding the removal of obscene or harmful content. Regulators must enforce accountability and transparency, holding platforms like Meta to legal and ethical standards in protecting users from harmful content. Additionally, developing guidelines or regulations specific to AI-generated content will help address its potential for abuse and harm, ensuring a safer online environment for all users.

Link to Attachment

No Attachment

2024-007-IG-UA, 2024-008-FB-UA	PC-27034	United States & Canada
Case number	Public comment number	Region
Kavita	Mehra	English
Commenter's first name	Commenter's last name	Commenter's preferred language
Sakhi for South Asian Survivors		Yes
Organization		Response on behalf of organization

Full Comment

The use of AI generated nude images of women, men or any individual, without their consent is a form of violence and should be treated as such. It is incumbent upon Meta to review its adjust its policies to ensure the safety and security of its users.

Link to Attachment

No Attachment

2024-007-IG-UA, 2024-008-FB-UA	PC-27035	Central & South Asia
Case number	Public comment number	Region
Blaise	Crowly	Hindi
Commenter's first name	Commenter's last name	Commenter's preferred language
DID NOT PROVIDE		No
Organization		Response on behalf of organization

Full Comment

Two points only -1. First of all am dumbstruck that a reporting event that involved an image with nudity went without review for over 48 hours. And further that it was just closed just for that reason. Irrespective of the reporters selection of option when reporting META needs to urgently setup an automated flagging system that prioritise reports of images with nudity (as it can very easily fall into Child sexual abuse material, bullying or revenge porn ...even when the reporter did not identify it as such) and prevent its auto closing. Basically no auto closing for reports on content with nudity, Use AI to detect that. 2. The responsibility in current political environment is very serious for technology pioneers who hope to establish themselves as credible platforms into the future. Hence META needs to put in effort into using a database of public figures and a capable AI model to scan through at least in reported content for public figures. If such is detected a further check for sexual tone could also be done making use a model. If both are true the ticket needs to be prioritised and not be closed. Please note that while these may sound like too much effort, the technology stack that is capable of this will only be a long term asset and only enhance METAs credibility as a trustworthy partner in the social space.

Link to Attachment

No Attachment

2024-007-IG-UA, 2024-008-FB-UA	PC-27036	Europe
Case number	Public comment number	Region
Withheld	Withheld	English
Commenter's first name	Commenter's last name	Commenter's preferred language
Withheld		No
Organization		Response on behalf of organization

Full Comment

Firstly, I'd like to highlight the importance of the language we use describing various forms of abuse facilitated by technology. So called deepfake pornography does not accurately capture the true nature of the non-consensual creation of sexualised images and videos featuring the likeness of someone or intended to represent someone. That's image-based sexual abuse. Calling it otherwise lacks clarity in distinguishing it from videos that feature fictional characters and prevents victim-survivors the opportunity to refer to their abuse without referring pornography, which suggests a level of consent or creation to cause sexual excitement. Our language matters: It informs what we understand to be right and wrong in society. Let's not minimize image-based sexual abuse in all its forms by misnaming it as pornography. Secondly, the harm image-based sexual abuse, including that created or maninpulated using AI, causes those victimised, who are mostly women, is often underestimated. I am a researcher and recently spoke to a woman who had a sexual video depicting her likeness and featuring her name circulated on Facebook. It is fair to say that this experience has ruined her life. Family members told her not to make a fuss, Meta did nothing to remove the video, law enforcement said it wasn't a crime, her college shrugged their shoulders, her friends told her to get over it, and her boyfriend remained friends with the man who created it, resulting in her ending their 3 year relationship. She is isolated, alone, and feels utterly powerless. If only this video had been immediately removed, perhaps she could have continued with her life at that time. But it wasn't removed, it continued to haunt her, and over the course of a year every door she went through to access support and help was closed in her face. We must do better, you, Meta, must do better. By taking a stand against all forms of image-based sexual abuse, including that created or manipulated using AI, Meta can send a clear message to their users and wider society: This is not ok, we will not accept it, it should not happen. And to those victimised, it says we believe you when you tell us this is harmful. We believe you.

Link to Attachment

No Attachment

2024-007-IG-UA, 2024-008-FB-UA

Case number

PC-27037

Public comment number

Asia Pacific & Oceania

Region

Samson

Commenter's first name

Selladurai

Commenter's last name

English

Commenter's preferred language

Im safe Australia Pty Ltd

Organization

Yes

Response on behalf of organization

Full Comment

In addressing deepfake pornography, Meta should prioritize robust, transparent enforcement mechanisms coupled with AI-driven detection tools that are sensitive to nuances in global contexts. It is crucial to implement stringent policies that specifically target and define deepfake content to prevent the spread of non-consensual imagery. These policies should be clear on the repercussions for violations and offer a streamlined, efficient reporting process.Furthermore, Meta should foster collaborations with AI ethics researchers and civil society organizations to keep pace with evolving technologies and societal norms. This includes ongoing updates to their algorithms and detection methods to stay ahead of deepfakes increasing sophistication. Additionally, public education campaigns about the impact and recognition of deepfakes could empower users to better identify and report such content.Effective policy enforcement must also include prompt human review of flagged content, ensuring that automatic closures do not hinder necessary actions against harmful posts. Ensuring transparency in these processes and decisions can help build trust and accountability, crucial for user cooperation in policing the platform. This comprehensive approach will support Meta's commitment to safeguarding user rights and maintaining the integrity of its platforms.

Link to Attachment

No Attachment

2024-007-IG-UA, 2024-008-FB-UA	PC-27039	United States & Canada
Case number	Public comment number	Region
Michelle	Neville	English
Commenter's first name	Commenter's last name	Commenter's preferred language
DID NOT PROVIDE		No
Organization		Response on behalf of organization

Full Comment

Meta does not do enough to curtail the explicit AI content at all. I just had a profile suggested to me with a photo of nude princesses and characters on my profile. They also do absolutely nothing to get rid of fake military profiles. They used to be better at catching these and now it seems like they dont care.

Link to Attachment

No Attachment

2024-007-IG-UA, 2024-008-FB-UA	PC-27040	Central & South Asia
Case number	Public comment number	Region
Withheld	Withheld	English
Commenter's first name	Commenter's last name	Commenter's preferred language
Withheld		No
Organization		Response on behalf of organization

Full Comment

Deepfake pornography is another vein of sexual crimes against women. As times evolve to integrate AI into every sphere of human evolution and convenience, feminist groups foresaw discrimination and sexism evolving in parallel. Pornography itself is built upon the exploitation of womens bodies by reinforcing that the naked female body is meant to be objectified and used for the pleasure of men. If there werent inherent ideologies that deemed womens bodies to be hidden, pure, virginal and sexually precocious there would be no stigma surrounding pornography and sex. In this, deepfake pornography simply makes it easier to produce images that dont even require women to show their bodies, just their faces. In an Indian setting, sex is still stigmatized and is a topic mired with taboos; constantly spoken in whispers or by female doctors in medical settings majorly in the context of family planning or childbirth. The perception of female itself can be defined by the binary psychological concept The Madonna-Whore complex. Women who are thought to follow the archetype that is in harmony with patriarchal expectations are Madonnas (pure, nurturing, maternal) and are revered as goddesses and mothers, their most important role being building families. Whereas the whores are immoral and seductive temptresses focused on their aim to indulge in sexual activity with many men while remaining ignorant to their role in crumbling families and exist only to be objects for men.The sexualization of female public figures is way to discredit the complex human beings they are, since once they are objectified, they remain objects whose opinions hold no water. It turns them from Madonnas to Whores. Their accomplishments, ambitions, talents and essence are automatically devalued. Deepfake pornography essentially communicates that the most important thing a woman can do is exist for male pleasure and the only thing she can be is a sexual object. That she will possess no other dimension to her person once she is sexualized. For public figures, this leaves a massive digital footprint that follows them throughout their lives and bleeds into every aspect of their careers and personal lives. With more of the world digitalizing and networking, it is much easier for deepfake pornography to spread and remain accessible, untraceable due to data privacy laws or because the internet is too vast to wade around in and retrieve the photo of every victim. With accessibility to AI tools becoming easier, Meta seems to be inundated with requests to review content that are hate crimes, bullying, sexualizing women without consent, violent and inappropriate, thus automated systems automatically close appeals that havent been reviewed in an attempt to accommodate new ones. Meta needs a separate category outside bullying and harassment to target sexist content and streamline the process of addressing it. The appeals made under this category could have a longer window for review before being closed and the members reviewing this content should be specialized in the respective field or even those personally victimized by Deepfake Pornography.

Link to Attachment

PC-27040

2024-007-IG-UA, 2024-008-FB-UA

Case number

PC-27041

Public comment number

Europe

Region

Ulf

Commenter's first name

Haeussler

Commenter's last name

English

Commenter's preferred language

DID NOT PROVIDE

Organization

No

Response on behalf of organization

Full Comment

My comment addresses the following aspect:"The challenges of relying on automated systems that automatically close appeals in 48 hours if no review has taken place. "1. General RemarksFrom a rule of law perspective, automatically closing appeals after a predetermined period of time is a doubtful practice.a) This practice is not equitable inasmuch as the likelihood of an appeals being closed may depend on the overall number of appeals filed, i.e., the more appeals in the system, the bigger the probability of a given appeal's being automatically closed.b) In addition to this empirical aspect, this practice defies legal certainty, which is equally important from a rule of law perspective. Doubts pertaining to this category may arise from the fact that the parameters for selecting appeals for further processing are not known or publicized. From an appellant's perspective it is neither predictable in advance nor understandable after the fact why his or her, rather than someone else's, appeal was not selected and hence closed automatically. 2. Stand-Alone AnslysisIn and of itself, the rule whereby appeals are automatically closed after the expiry of a predetermined period of time does not reflect an appropriate balance of all interests relevant to the creation of an oversight system. This rule seems to implement a mechanism exclusively designed to handle the volume of appeals. Whilst it is accepted practice that, e.g., senior courts apply a volume management methodology, volume management is not the sole concern of any such methodology, no matter where it is employed. On addition to this legitimate interest of the reviewers, those of appellants also need to be part of the equation. Appellants' relevant interests may comprise, without being limited to, the scale and gravity of the conduct appealed against, and the appeal's prospect for success on the merits. More specifically, some appeals may deserve to be selected for further processing regardless of their prospect for success because the questions they rise are important beyond the individual case, and some appeals may deserve to be selected for further processing regardless of their importance beyond the individual case because of their prospect for success - these elements reflecting the generalistic and individualistic aspects of case-by-case justice. Inasmuch as automatically closing appeals completely ignores case-by-case justice, a rule implementing such mechanism fails to deliver material justice.3. Analysis in Context The rule causing appeals to be automatically closed may have seriously adverse cumulative effects on an appellant's interests and vis--vis the broader community if applied in conjunction with different rules or practices that do not mitigate the permissible outcome of volume management (i.e. of a volume management mechanism that balances the interests of reviewers and appellants in line with the rationale discussed at para 2). In the situation at hand, such different rules or practices may be reflected by the oversight board's decision to accept the case on further appeal. Being unaware of the process and parameters underlying this decision, I would hesitate to state whether - leaving aside the obvious positive effects in the case at hand - seriously adverse cumulative effects are sufficiently well excluded, or whether the oversight board is equipped with sufficient resources to exclude said effects.

Link to Attachment

No Attachment

2024-007-IG-UA, 2024-008-FB-UA	PC-27042	United States & Canada
Case number	Public comment number	Region
Jared	Weisinger	English
Commenter's first name	Commenter's last name	Commenter's preferred language
DID NOT PROVIDE		No
Organization		Response on behalf of organization

Full Comment

As someone that has had various images taken down that were not anyore explicit than what is commonly posted and accepted, I think Metas policies are too strict.While I understand that deep fakes can constitue as harassment, they are not entirely preventable or enforceable. It will forever be a cat and mouse game. Personally, unless words are included that deem it harassment, I don't think there should be an issue with deepfakes. Let everyone eventually come to understand that not everything you see online is real, and that includes explicits. It should be treated no different than other art created depicting someone, despite how "real" it may seem. AI is artwork, it's just the computer science version.So for example, sending the photo to ones friends claiming its them, that is harassment, but merely posting it? Even with a name, that is art, as long as it's mentioned that it is AI generated and not real. Trying to pass it off as real could deem it harassment.

Link to Attachment

No Attachment

2024-007-IG-UA, 2024-008-FB-UA

Case number

PC-27043

Public comment number

Europe

Region

Anastasia

Commenter's first name

Karagianni

Commenter's last name

English

Commenter's preferred language

DID NOT PROVIDE

Organization

No

Response on behalf of organization

Full Comment

DID NOT PROVIDE

Link to Attachment

PC-27043

2024-007-IG-UA, 2024-008-FB-UA	PC-27044	Central & South Asia
Case number	Public comment number	Region
Barsha	Chakraborty	English
Commenter's first name	Commenter's last name	Commenter's preferred language
Breakthrough Trust		Yes
Organization		Response on behalf of organization

Full Comment

Context for the usage of deepfake pornography against women in India: Non consensual images/pornography/deepfakes etc are used to defame, discredit, shame/humiliate women and push them out of online spaces. Lack of access and control is the point. We have no legislation to directly deal with AI/deepfake related GBV, therefore it becomes more difficult to deal with. It is currently treated as a theft of identity or along those lines rather than gender based violence which often trivialises the nature of the crime.Women especially find it difficult to report such cases because there is no clear legal framework here and the police dont understand/cant help. They often face secondary victimisation while reporting such cases in police stations / courts (why did you put your picture out etc. even when its not their pictures such as deepfakes) Once on the internet, the picture goes beyond the source platform very fast and merely taking it down on the source platform is not enough because it quickly spreads to other platforms (such as from Facebook to WhatsApp or Telegram) and can also be shared in the form of screenshots, with little to no control. In countries like India, while pornography is illegal yet remains in use, it becomes difficult for organisations to track down cases where such images or videos have made their way onto pornographic websites. What needs to be done: Based on the above, its important that Meta adopt a multi-stakeholder approach to dealing with deepfake pornography as its not enough to simply delete the image on their respective source platform. - Meta platforms should invest in building mechanisms that can prevent the uploading of Deepfake/AI-generated pornography on this platform. It should also put a disclaimer on susceptible content for the users. - 48-hour redressal time should not be uniform for all cases. If any case requires multiple reviews, it should be considered. Meta should consider not closing those appeals. - Meta platforms should also lay out a proper set of guidelines on their respective platforms which are easy to access and can read in local languages so that there is no uncertainty about what kind of content is allowed/not allowed on the platforms and more importantly, the steps to be taken in case someone finds content such as deepfakes etc being shared on these platforms. The tools for reporting should be often advertised on the platform and easily accessible. - Under Metas current platform policy on platforms such as Facebook and Instagram, there is no space to report AI-generated content or deep fakes. This should be added. - Prioritise human-centric intervention rather than AI-centric intervention because AI is likely to miss out on contextual/cultural intricacies.

Link to Attachment

No Attachment

2024-007-IG-UA, 2024-008-FB-UA	PC-27045	Asia Pacific & Oceania
Case number	Public comment number	Region
MARIA	OSULLIVAN	English
Commenter's first name	Commenter's last name	Commenter's preferred language
DID NOT PROVIDE		No
Organization		Response on behalf of organization

Full Comment

Meta Oversight BoardInquiry into Explicit AI Images of Female Public FiguresSubmission by Dr Maria OSullivan, Associate ProfessorDeakin Law School, AustraliaOverviewI make this submission in my role as an Associate Professor in the Deakin Law School in Melbourne, Australia and as a scholar with expertise in international human rights law. My submission focuses on two aspects of the current investigation:1.How the use of deepfakes impacts womens right to participation in public life; and2.The compliance of the automated review and appeal system with the right to remedy under international human rights law. I note that Meta is not, strictly speaking, a party to international human rights law treaties (as it is not a state). However, Metas Corporate Human Rights Policy and other frameworks indicate that it has agreed to abide by core human rights principles. Thus, it is important that Meta policies on Explicit AI Images of Female Public Figures reflects relevant international human rights norms. The aim of my submission is to bring these international human rights law principles to the attention of the Oversight Board.1.Deepfakes and the right of women to participate in public lifeRelevant Human Rights PrinciplesArticle 25 of the International Covenant on Civil and Political Rights recognizes the right to participate in public affairs, including the following three elements: (a) the right to take part in the conduct of public affairs; (b) the right to vote and to be elected; and (c) the right to have access to public service. This should be read in conjunction with Article 7 of the Convention on the Elimination of All Forms of Discrimination against Women 1979 (CEDAW) which calls on State Parties to eliminate discrimination against women in the political and public life and in particular to ensure womens equal rights: (a) to vote in all elections and public referenda and to be eligible for election to all publicly elected bodies; (b) to participate in the formulation of government policy and the implementation thereof; and to hold public office perform all public functions at all levels of government, and (c) to participate in non-governmental organizations and associations concerned with the public and political life of the country. Deepfakes and Political ParticipationThere is evidence from academic and civil society commentary that publication of deepfakes of women acts as a chilling effect on the participation of women in public life. For instance, human rights and technology specialist Vandinika Shukla has noted that:'Online violence against women includes aggression, coercion, and intimidation that seeks to exclude women from politics simply because they are women. It targets individual women to harm them or drive them out of public life, but also sends a message that women dont belong in politics as voters, candidates, office holders, or election officials'. In a 2023 report on Technology-Facilitated Gender-Based Violence as an Attack on Womens Public Participation, scholars found that:'Online violence reinforces inequality and maintains discriminatory norms, maintains and reinforces patriarchal gender hierarchies, and can result in WIPPL [Women in Politics and Public Life] choosing not to engage in public life or similar roles, for fear of abuse. As a result of TFGBV [Technology-Facilitated Gender-Based Violence], women in public life can feel compelled to withdraw from online as well as offlinepublic spaces. 'This is significant given that women are under-represented in political and other public roles a problem which has been recognised in a number of UN reports. For instance, the UN Human Rights Committee has raised concerns about the under-representation of women in senior positions in the public service, in political life, the judiciary and other sectors and frequently recommends affirmative action where necessary. The UN Human Rights Council has also highlighted the impacts of discrimination on the right of women to participate in public affairs:'The adverse impact of discrimination, including multiple and intersecting forms of discrimination, on the effective exercise of the right to participate in public affairs should be recognized, in particular for women and girls, young people, persons with disabilities, indigenous peoples, older persons, persons belonging to minority groups, persons with albinism, lesbian, gay, bisexual, transgender and intersex persons and other groups that are discriminated against.'Lucy

Purdon, an expert on gender justice and technology has also highlighted that women candidates do not typically have the funds to counter sexualised disinformation such as deepfakes. As she states: 'Online harassment will have a higher cost for female politicians because that harassment manifests in not just attacks on political competency but a cultural rejection of women. Women candidates are already too underfunded to challenge sexualised and gendered disinformation and will always risk stronger retaliation. 'My submissionI therefore urge the Oversight Board to consider deepfakes of female public figures in the context of these broader, systemic problems with womens participation in public life.I also urge the Board to frame deepfakes as a form of gender-based violence against women. In this regard, I call your attention to the comments of the 2020 Expert Group Meeting on Womens full and effective participation and decision-making in public life, as well as the elimination of violence, for achieving gender equality and the empowerment of all women and girls which stated that:'Violence against women in political and public life is internationally recognized as a violation of womens political rights. Violence against women in politics (VAWP) is a form of gender-based violence against women. It is any act, or threat, of physical, sexual or psychological gender-based violence against women that prevents women from exercising and realizing their political rights and a range of human rights. It manifests in specific, gendered ways that men do not experience, including but not limited to physical violence, sexual violence and psychological violence.' 2.Meta Processes and the Right to a RemedyThe right to a remedy under international human rights law is raised by the reliance by Meta on automated systems that automatically close appeals in 48 hours if no review has taken place. The right to an effective remedy is set out in Article 2(3) of the International Covenant on Civil and Political Rights (ICCPR). This states that signatories must commit themselves to ensuring that any person whose rights or freedoms are violated under that treaty shall have an effective remedy and that any claims shall be determined by a competent authority and remedies enforced by those authorities. The UN Human Rights Committees General Comment 31 on the right to a remedy emphasises that Article 2(3) of the ICCPR places obligations on signatories to ensure that individuals have accessible and effective remedies to vindicate their rights. The General Comment also underlines that [s]uch remedies should be appropriately adapted so as to take account of the special vulnerability of certain categories of persons... My SubmissionWhilst I recognise that the provision of remedies for users of high-volume online systems must be balanced with the need for review procedures to be efficient and timely, I urge the Board to consider the importance of deepfakes to womens human rights, including its use as a form of gender-based violence against women (as noted above).The use of a blanket closedown of appeals, without consideration of the vulnerability of certain users, raises the risk of improperly denying a remedy to an individual. I urge the Board to consider the need to adapt the online review and appeal process so that it is calibrated to allow for some flexibility. One solution could be for users to identify that they are suffering from a particular vulnerability (as noted above in the UN General Comment on the Right to a Remedy). Further, there should be some flexibility built into the process when the case involves violence (including online gender-based violence such as the use of female deepfakes).Contact informationShould you have any questions arising from this submission, please contact me at:Email: m.osullivan@deakin.edu.au   Mobile: +61 415585708. Dr Maria OSullivan

Link to Attachment

PC-27045

2024-007-IG-UA, 2024-008-FB-UA	PC-27046	Central & South Asia
Case number	Public comment number	Region
Salonee	Mistry	Hindi
Commenter's first name	Commenter's last name	Commenter's preferred language
DID NOT PROVIDE		No
Organization		Response on behalf of organization

Full Comment

Whether public figure or not, deep fakes by their very nature are damaging of one's reputation and usually created with the knowledge that it will cause a certain amount of reputational harm to the individual whose image has been deep faked. In both the above mentioned cases, there was knowledge that sharing the deep fake images would cause a certain degree of harm or shame to the public figures involved. Intention in this regard is inconsequential, as no deep fake can result in anything good and this is the universal truth. The 'fake' part of the word, takes away any autonomy from creator or publisher to claim that they posted the deep fake with 'good intentions'. Whether they were shared to gain popularity or with malice, in both these instances the fact that sharing a deep fake will adversely affect the individual is known and as such both should have been taken down instantly and the publisher reprimanded for them to truly realise the damage they have done. The onus of replying on time lies on Meta, as in the first case. If they have too many complaints to get through because of which 48 hours is also not enough to respond to complaints then they need to recheck their policies as they are getting too many complaints to begin with. Deep fakes in a country like India, where the respect and honour of people, especially women is tied to how society perceives her, images like this, no matter her stature in society, can be extremely harmful and must be taken that much more seriously.

Link to Attachment

No Attachment

2024-007-IG-UA, 2024-008-FB-UA	PC-27047	Central & South Asia
Case number	Public comment number	Region
ARNIKA	SINGH	English
Commenter's first name	Commenter's last name	Commenter's preferred language
Social & Media Matters		Yes
Organization		Response on behalf of organization

Full Comment

We strongly condemn the proliferation of any form of inappropriate images, especially those targeting women. The use of deepfakes and AI-generated images can cause significant mental trauma and tarnish the image of the targeted person. The virality of these images exacerbates the problem, as the time taken to address reports allows perpetrators to circulate them widely. This distribution without consent is deeply problematic and harmful.The nature and gravity of harm posed by deepfake pornography, especially towards women who are public figures, cannot be overstated. Such content not only violates privacy but also has serious implications for the professional and personal lives of the individuals targeted. It can lead to harassment, reputational damage, and psychological distress.The use and prevalence of deepfake pornography globally are on the rise, impacting individuals in countries like the United States and India. The ease of creating and sharing deepfakes has contributed to their proliferation, making it crucial for platforms like Meta to implement robust measures to address this issue.During India's election period which is right now as we submit the comment, there is an increase in the abuse of AI-generated deepfakes, particularly targeting women. The heightened national sentiment during this time, coupled with the misuse of social media platforms, exacerbates the issue of tarnishing the images of leaders. India, as a nation, becomes emotionally charged when a woman's image is involved, leading to wider circulation, derogatory conversations, and the topic becoming highly sensationalized. You may encounter posts, graphics, or videos featuring manipulated images of leaders superimposed onto movie or song posters, often with derogatory intent. Understanding the context and narrative behind these images is crucial to recognizing their inappropriate nature. India witnessed a lot of deepfakes that were targeted at film actresses and politicians. The bigger issue is the time taken for action and the detection of source to immediately end the circulation. We need to understand that the ease of creating these videos is whats leading to their proliferation. To address deepfake pornography on its platforms, Meta should consider implementing stricter policies and more effective enforcement processes. Implement a strict zero-tolerance policy for deepfake content. Utilize advanced AI detection technology to proactively identify and remove deepfake videos and images. Collaborate with trusted fact-checkers and trusted partner networks to quickly verify and flag potential deepfake content. Provide clear guidelines to users on what constitutes a deepfake and the consequences of sharing such content.Enhance Reporting Mechanisms: Improve reporting tools for users to easily flag suspected deepfake content. Provide clear guidelines on what constitutes a deepfake to assist users in reporting accurately.Educational Campaigns: Launch educational campaigns to raise awareness among users about deepfakes, including how to identify them and the potential harms they pose.Regular Audits and Reviews: Conduct regular audits and reviews of content moderation processes to identify areas for improvement and ensure effectiveness in combating deepfakes. The gray zones in the content moderation policies often give a free pass to inappropriate content to circulate online. Social & Media Matters recommends that Meta should leverage its trusted partner networks more effectively. Reports flagged by these partners should be given prime importance, and a streamlined mechanism should be developed for quick communication between partners and the platform to discuss and resolve issues promptly. This would help in addressing harmful content more efficiently and prevent its spread.Furthermore, Social & Media Matters highlights the need to address confusion in the appeal process. For example, if a trusted partner has flagged content and it is taken down, there should be clarity and consistency in the decision-making process. It can be confusing if content is reinstated upon appeal after being taken down based on partner flags. Meta should ensure that the appeal process is transparent and that decisions are made based on clear and consistent guidelines.Meta's enforcement of its "derogatory sexualized photoshop or drawings" rule in the Bullying and Harassment policy, including the use of Media Matching Service Banks, is a step in the right direction. However, more proactive measures may be necessary to prevent the spread of deepfake pornography.One of the challenges Meta faces is relying on automated systems that automatically close appeals in 48 hours if no review has taken place. This can lead to delays in addressing reports and allow harmful content to remain online longer. Meta should review and improve its appeals process to ensure timely and effective responses to reports of deepfake pornography.In conclusion, addressing deepfake pornography requires a multi-faceted approach, including robust policies, effective enforcement mechanisms, and collaboration with experts and stakeholders. By taking decisive action, Meta can help mitigate the harms posed by deepfake pornography and protect the privacy and dignity of individuals, especially women who are public figures.

Link to Attachment

No Attachment

2024-007-IG-UA, 2024-008-FB-UA	PC-27048	Latin America & Caribbean
Case number	Public comment number	Region
Yasmin	Curzi	English
Commenter's first name	Commenter's last name	Commenter's preferred language
Center for Technology and Society FGV Rio Law		Yes
Organization		Response on behalf of organization

Full Comment

Submission to the Meta Oversight Board about sexually explicit AI-generated imagesBy Prof. Dr. Yasmin Curzi (Center for Technology and Society at FGV Law Rio de Janeiro) Harms posed by non-consensual intimate images and deep nudesThe exposure of personal images in online platforms is a recurring concern on the Internet since the beginning of the Web 2.0. In Brazil, during the development of our Civil Rights Framework for the Internet (MCI), from 2009 to 2014 , cases of "revenge porn", or, in the most appropriate way, "dissemination of non-consensual intimate images (NCII)", were highlighted, due to the suicide of two teenagers after the sharing of their images online by their ex-boyfriends. Due to these cases, civil society and academia were pushing for more expeditious removal and a specific notice-and-takedown regime to avoid the harms posed by the sharing of such contents.In addition to the MCI, Law No. 12,737/2012, known as the Carolina Dieckmann Law, is a significant milestone in combating the non-consensual dissemination of intimate images in Brazil. Introduced after actress Carolina Dieckmann had her personal photos released without authorization, the legislation amended the Penal Code to explicitly criminalize the invasion of electronic devices with the intention of obtaining, tampering with or destroying personal data, including intimate images. The law sought not only to provide a more effective remedy for hacking and other violations of privacy, but also to be a tool for reparation of the harms posed by NCII. Concerns about the dissemination of personal images and videos reached a new level with the advent and dissemination of Generative Artificial Intelligence (Gen-AI) and Deep Learning tools. These new technologies allow us to create content with simple prompts and/or alter not only static images, but also videos, audios and even live broadcasts, challenging social capacities to discern real from fake. Regarding NCII, such new tools create space for new forms of harm, such as deep nudes (or fake nudes), which are already threatening and violating rights of girls and women worldwide in multiple ways. Some of the rights violations are:Regarding privacy, honor and reputation (Article 17, ICCPR): by allowing that their images are synthetically altered to depict nudity, IIGT not only harms women's reputations, but endangers all online activity that rely on their image sharing.The constant and pervasive fear of having their images altered to depict themselves nude also significantly endangers freedom of expression (Article 19, ICCPR). There is a potential and permanent chilling effect that might cause generational harms to girls and women all over the world. Reasonably, women and girls would prefer not to share their images online or actively participate in the digital sphere, risking to be targets of image-based abuse.   Participation of women in politics is also threatened by the use of deep nudes and image-based abuse. Cheap fakes/shallow fakes and deep fakes to attack campaigns, as tactics of gendered disinformation are already causing significant harms for women in politics worldwide The economic and social rights of women and girls, as laid out in the International Covenant on Economic, Social and Cultural Rights (ICESCR), are also at stake. The ability of women to safely and confidently participate in digital commerce and the broader digital economy, such as content creators, can be severely hindered by the threat of image-based abuse. Evidently, deep fake and Gen-AI tools can be employed in the creation of artistic content, which could be encompassed by free speech rights. Nevertheless, when a third-party uses an image of another person, without taking into account their authorization, to depict nudity, they are acting unlawfully. In this scenario, balancing free speech rights with integrity rights and also the free speech rights of potentially affected users, there is no reasonable argument that would justify a lighter measure than completely forbidden synthetic content that depicts nudity on Meta's platform. Following Robert Alexy's methodology for weighing rights, this measure (1) is suitable as the only means available to stop and avoiding harms; (2) it is necessary to protect users from the chilling effects and other harms derived from the mere existence of deep nudes; and (3) considering the proportionality in a narrow sense, potential free speech rights in the producing of such content in the case of the person depicted had previously authorized the usage of their copyrights and image rights to an artist , the disproportionality between the two cases is obvious. 2. Current Policy by MetaRegarding Metas current policy enforcement mechanism, the automated systems that close appeals automatically if no review has taken place within 48 hours, can lead to several issues such as: Lack of fairness, as users whose content has been wrongly flagged might feel unjustly penalized if their appeal is not reviewed timely;The closures without review can erode user trust in the platforms commitment to fair and accurate content moderation.In addition, heavy reliance on automation can be problematic, especially for complex decisions involving context and intent that AI may not yet fully comprehend.Furthermore, there is no data available from Meta regarding the enforcement of the protection given to all people from being attacked by "derogatory sexualized photoshop or drawings", as provided for in its Bullying and Harassment policy.There is also no information on the use of "Media Matching Service Banks" (MMSB) in this context. MMSB is a mechanism used by Meta to automatically identify and remove images that have already been identified by humans as violating its guidelines. The OversightBoard has already acted in a case about the use of this mechanism ("Colombian police cartoon"/2022-004-FB-UA). In the case, the OB highlighted how the use of the MMSB can amplify the impact of errors by human reviewers, and ordered the Meta to develop correction mechanisms for the instrument, including publicizing its error rates.3. RecommendationsRecent UNESCO report (2023) highlights the need for reporting mechanisms, proactive methods of identifying artificially created content and transparency in access to third-party controls as crucial measures. In this sense, the StopNCII (Stop Non-Consensual Intimate Image Abuse), which offers tools for the protection of intimate images, is a reference in initiatives for victims of NCII. The tool works by creating a hash of the intimate content, which will be used to search for similar files in the databases of partner companies for removing automatically and preventing them from being shared. Stemming from this, we offer the following recommendations for the Oversight Board in order to address deep nudes adequately and promoting the respect for for women and girls rights worldwide:Encompass deep nudes into the NCII existing framework, not allowing the circulation of any synthetically altered content that depicts nudity in the platform; Expand the use of media matching technologies to include a database specifically for reported and removed deep fake content. Update the Community Policy to include explicit references to deep nudes, clarifying the rules for users and set a clear legal ground for enforcement.Recruit more specialized and local content moderators, providing them with sufficient resources and training, taking into account the complexity of TFGBV. Increase the quantity of human moderators and reviewers trained specifically to handle cases involving complex content, ensuring that nuanced decisions are made.Implement a policy guaranteeing that all appeals are reviewed by a human moderator before any closure decision is made, possibly extending the 48-hour window if necessary.Collaboration with academia, non-profits, and other tech companies to improve detection technologies and share best practices for content moderation.Publishing detailed reports on the types and volumes of deep nude content detected, the actions taken, and the outcomes of those actions, including appeals.Regularly audit the performance of automated systems and make adjustments based on effectiveness and user feedback.

Link to Attachment

PC-27048

2024-007-IG-UA, 2024-008-FB-UA	PC-27049	Asia Pacific & Oceania
Case number	Public comment number	Region
Ysrael	Diloy	English
Commenter's first name	Commenter's last name	Commenter's preferred language
Stairway Foundation		Yes
Organization		Response on behalf of organization

Full Comment

DID NOT PROVIDE

Link to Attachment

PC-27049

2024-007-IG-UA, 2024-008-FB-UA	PC-27051	United States & Canada
Case number	Public comment number	Region
Gregory	Angelo	English
Commenter's first name	Commenter's last name	Commenter's preferred language
New Tolerance Campaign		Yes
Organization		Response on behalf of organization

Full Comment

April 29, 2024Meta Oversight Board[Submitted electronically via this link]Members of the Oversight Board:The New Tolerance Campaign (NTC) is a watchdog organization whose mission is to ensure institutions consistently apply their stated policies and values, particularly when it comes to fostering free speech and open dialogue. As such, we welcome the opportunity to submit comment in response to Facebooks request seeking guidance from the Oversight Board regarding the Explicit AI Images of Female Public Figures case.Before addressing the specific issues for which the Oversight Board is seeking guidance, it must be noted that NTC agrees with the decision to remove the images off its platforms. It should also be noted that our concurrence has nothing to do with the artificial intelligence (AI) aspect of the cases in question.To be sure, questions and controversies regarding deepfakes and AI-generated images will become more numerous and complex in the near future. A Bloomberg Law report by Isaiah Poritz noted recent viral deepfakes that included [i]mages of former President Donald Trump hugging and kissing Dr. Anthony Fauci, his ex-chief medical adviser and [p]ornographic depictions of Hollywood actresses and internet influencers as well as [a] photo of an explosion at the Pentagon. As Indian AI creator Divyendra Singh Jadoun told The Washington Post, The only thing stopping us from creating unethical deepfakes is our ethics. In instance of the cases at hand, Meta already had guardrails in place. Policies governing [d]erogatory, sexualized photoshopped images and drawings were applied to this case consistent with Meta rules. Of note and concern however, is another Meta policy that states the display of digitally created sexual content is restricted unless it is posted for educational, humorous, or satirical purposes. In regard to the cases herein, those criteria do not apply, but these qualifiers will need to be confronted sooner rather than later, as sexualized AI images of public figures could be considered exempt under those guidelines.As for the questions for which the Meta Oversight Board is soliciting comment at present, NTC offers the following input:The nature and gravity of harms posed by deepfake pornography including how those harms affect women, especially women who are public figures.Deepfake pornography reduces women to mere objects of sexual gratification and ridicule. For public figures, the circulation of such depictions can severely damage their reputation and credibility. Deepfake pornography also opens the door to extortion and blackmail, where the target can be threatened with the release of realistic but phony content unless certain demands are met. This presents a damaging prospect to public figures and everyday Americans alike: the potential compromising of their personal and professional lives.Strategies for how Meta can address deepfake pornography on its platforms, including the policies and enforcement processes that may be most effective.The threshold of what constitutes harmful AI must be higher for public figures than everyday Americans. Being a target of criticism and attacks (even vulgar ones) comes part-and-parcel with notoriety. As mentioned above, educational, humorous, and satirical exemptions to Meta bans on sexual content exist and will need to be tackled in the future.AI can also be used for good in cases such as these. AI images and video should be labeled as such, either automatically via Media Matching Service Banks or by users confronted with a prompt that would allow them to manually confirm the image or video they are uploading was created in whole or in part using AI.Meta should establish clearer policies prohibiting the distribution, creation, and sharing of AI generated pornography on their platforms. Artificial Intelligence must explicitly be added to Meta content policy so that there is no question about where rules governing AI apply. The photoshop rule rightfully applied in these cases may be insufficient in future adjudications of AI-related content.Meta should launch public awareness campaigns highlighting the dangers of AI and deepfake pornography. Such campaigns can underscore the growing prevalence of deepfake pornography, state the actions Facebook and Instagram are taking to combat it, and recommend users approach unusual posts with skepticism.The challenges of relying on automated systems that automatically close appeals in 48 hours if no review has taken place.NTC questions how the 48-hour rule was created was the time span for open appeals arbitrary? Was it sufficient to address concerns about content moderation in the past but is now inadequate? Meta must consider these questions in revisiting its window of appeals, increasing the window of time appeals remain active and/or adding additional staff to handle the growing and increasingly complex caseload AI presents.By implementing these recommendations, Meta can take significant steps towards mitigating the harm of deepfake pornography on its platforms while upholding user safety and trust.NTC once again thanks the Oversight Board for the opportunity to submit this comment for your consideration and would be glad to engage with you further on this matter should you see fit.Sincerely, Gregory T. AngeloPresident

Link to Attachment

PC-27051

2024-007-IG-UA, 2024-008-FB-UA	PC-27053	Europe
Case number	Public comment number	Region
Withheld	Withheld	English
Commenter's first name	Commenter's last name	Commenter's preferred language
Withheld		No
Organization		Response on behalf of organization

Full Comment

As someone with over 8 years of experience in combating child sexual exploitation and advocating for marginalised groups like people living with HIV and the LGBTQ+ community, I understand and have witnessed the profound impact of deepfake pornography or forged videos, as they were previously addressed, on individuals' fundamental rights and dignity, particularly for women and children.Beyond being a violation of privacy, deepfake content inflicts deep emotional trauma, perpetuates harmful stereotypes, and can severely damage reputations. For women and historically marginalised group, especially those in the public eye, the consequences can be devastating, leading to harassment, threats, and even physical harm. Children, being particularly vulnerable, face profound risks from deepfake videos, which can cause significant psychological distress. These manipulated videos often depict children in compromising or inappropriate situations, leading to confusion and emotional turmoil. The inability of young minds to discern between real and fake content exacerbates their distress, leaving them feeling helpless and exposed.Moreover, the dissemination of such videos can result in social isolation, bullying, and lasting stigma, compounding the emotional impact . Perpetrators of child exploitation may exploit deepfake videos to manipulate and coerce children, perpetuating cycles of abuse and harm. As a result, children may endure long-term psychological consequences, including anxiety and depression. The pervasive nature of digital media facilitates the rapid spread of these videos, amplifying the harm inflicted on vulnerable children.Protecting children from deepfake videos requires proactive measures, including comprehensive education on media literacy and online safety. Swift intervention by platforms and law enforcement is crucial to remove and mitigate the dissemination of harmful content. Empowering children with the skills to critically evaluate digital content can help mitigate the impact of deepfake videos on their mental well-being. However, the onus is on adultsparents, educators, policymakers, and technology companiesto take responsibility for protecting children from exposure to harmful content.Creating safe spaces for children to express their concerns and seek support is essential in addressing the emotional fallout from encountering such content. Collaboration between stakeholders is vital in safeguarding children from the detrimental effects of deepfake videos. Upholding children's rights to privacy and dignity must be a central consideration in all efforts to combat the spread of deepfake content. By prioritising child protection and well-being, we can work towards creating a safer online environment for the next generation.Addressing this issue requires a multi-faceted approach. Meta must establish clear policies prohibiting the creation, distribution, and sharing of non-consensual explicit content, rigorously enforced through a combination of advanced AI detection technology and human moderation. Collaboration with child protection organizations, human rights advocates, and academic experts is essential to develop targeted strategies and interventions.Moreover, investing in user education and awareness campaigns can empower individuals to recognize and report deepfake content, fostering a culture of digital literacy and responsible online behavior. By prioritizing the protection of human rights and child safety, Meta can play a crucial role in combating deepfake pornography and creating a safer online environment for all users. Regarding automated systems for handling appeals, relying solely on them poses significant challenges, especially concerning child protection and human rights. These systems may lack the context and nuance required to accurately assess complex situations, potentially resulting in wrongful removals or dismissals of legitimate appeals.To address these challenges, Meta should prioritise human oversight and review processes, ensuring trained moderators handle appeals promptly and effectively while incorporating safeguards to protect fundamental rights principles.

Link to Attachment

No Attachment

2024-007-IG-UA, 2024-008-FB-UA	PC-27055	Central & South Asia
Case number	Public comment number	Region
Kaustubha	Kalidindi	English
Commenter's first name	Commenter's last name	Commenter's preferred language
Tattle Civic Technologies Pvt Ltd		Yes
Organization		Response on behalf of organization

Full Comment

In India, with increased access to the internet, there has also been an increase in accessing pornographic material. Excessive watching of porn is also beginning to be considered a public health concern. Many communities in India to this day penalise women for being born, and for being visible. The gender ratio is skewed, cross-gender interaction is restricted, nudity is normalised and sex is still considered a taboo. Deepfake pornography is being introduced in a specific cultural context and societal norms which differ across geographies, and is now one of the ways in which men are introduced to the opposite sex; it is deeply problematic and presents harms in an immediate sense through the expectations and perceptions that it sets for real world engagement with women. Deepfake pornography of women public figures is intended to degrade, and dehumanise them to an extent where they are objects to be manipulated digitally to a persons satisfaction. It takes away the agency and dignity of an individual to present themselves in a manner they wish to do so.India stands 3rd in the list of countries for most porn-watching, and 4th in highest rape crimes amongst countries. It has been approximated that 93 women in India are raped everyday.

Deepfakes are widely used in India, including for pornographic purposes. Considering the nature and gravity of harm caused to women public figures, it is all the more important that any fallout is minimised to the extent possible. It may require re-prioritisation and a re-assessment of current escalation protocols at Meta.Detailed comments addressing the issues as presented by the Oversight Board are available in the attachment.

Link to Attachment

PC-27055

2024-007-IG-UA, 2024-008-FB-UA	PC-27056	Europe
Case number	Public comment number	Region
Ellen	Judson	English
Commenter's first name	Commenter's last name	Commenter's preferred language
Global Witness		Yes
Organization		Response on behalf of organization

Full Comment

DID NOT PROVIDE

Link to Attachment

PC-27056

2024-007-IG-UA, 2024-008-FB-UA	PC-27057	Europe
Case number	Public comment number	Region
Withheld	Withheld	English
Commenter's first name	Commenter's last name	Commenter's preferred language
Withheld		Yes
Organization		Response on behalf of organization

Full Comment

We would like to thank the Oversight Board for the ability to provide comments on this emerging and exceptionally important topic for consideration. We are a company who deal predominantly in the provision of services within the Domain Name System (DNS). We have chosen to make this comment anonymously so as not to seek to cause any perception of connection with our services, or our clients services with these comments. As a service provider in the DNS, we understand and appreciate the shared goal in the reduction of victimization and in seeking to lessen the impact of online harms on end users. Our goal however is tempered by knowledge that we all play a distinct role, and in this context, we wish to ask the question if not Meta, then who will act on such matters? Speaking with a distinct expertise and knowledge of dealing with allegations of abuse, both technical and content related within the DNS (i.e the suspension of domain names) we know very well the consequences of suspension and the potential and massive collateral damage that may occur to entities, such as Meta, who rely on the domain name as a critical backbone to their online presence. As such, in the context of the Case before the Oversight Board, our comment is intended to establish clearly, that Meta, and its constituent platforms, rank as the most appropriate place to seek the primary disruption of harms occurring on, or through their platforms.

This is fueled by the now persistent attempts of many key parties, through the Internet Corporation of Assigned Names and Numbers (ICANN) multistakeholder process, and through governmental lobbying (e.g. at the European level) who seek to force intervention on such content related matters, by operators on the other layers of the internet stack, such as the DNS. In this context, we believe a strong and consistent application of the terms and conditions of Meta (and its platforms) remains the most proportional means by which content, provided through and on your platforms, as well as other similar, albeit smaller platforms and services, should be primarily tackled. The challenges of relying on automated systems that automatically close appeals in 48 hours if no review has taken place & Metas enforcement of its derogatory sexualized photoshop or drawings rule in the Bullying and Harassment policy, including the use of Media Matching Service Banks.From the outset, we must note that purely content related claims/reports of harm, as well as an observed otherwise genuine use of a domain name (e.g. as an identifier for a Platform such as Facebook or Instagram), are usually a defining factor that would prevent registry or registrar intervention. This being noted, in cases where there exists an evidenced, objective, credible and immediate harm to end users, several industry players do agree there remains enhanced potential for domain name level intervention. Although this would usually be isolated to domains which either seem dedicated to such a harm, the accumulative effect of factors such as the lack of effective and consistent review of complaints, continued evidence of ongoing harms, and a lack of recourse for those affected, remain key considerations. Ultimately, balancing the potential consequences of such an intervention (e.g. a suspension of a domain name) vs. the gravity/impact of the harm must be very carefully considered. Application of the Meta Terms and Conditions. Considering the Case before the Oversight Board, the reluctance of those empowered to take the decision in the first instance (on one hand permitting continuation of the harm complained of, and the other seeking to prevent it) undermines all claims that such harms are carefully considered. This infers that the terms of service are anything but thoughtfully and consistently applied. Regardless of whether the Oversight Board concludes that observed issues are rightly considered abusive or not abusive, a major issue lies in this inconsistent application of the terms and conditions. Primarily, there was clearly a failure to intervene appropriately in at least one of these cases. Meta (Facebook and Instagram), inconsistently applied, or had no clear uniformity or process that apparently sought to expeditiously address similar complaints, in a satisfactory and consistent manner, despite actual notice of the issues in both. Should Meta intervene? Although not the point of our submission, we do feel compelled to note, that given at least one of your teams in their assessment believe the content to be contrary to the bullying and harassment policy, this was the right call. This seems to be easily reconcilable. When one considers the act of creation and of sharing a sexualized image, which depicts, or is intended to deceive a reasonable person into believing it to be another person, it is hard to consider a genuine purpose. Given the technology now routinely available, immediate, strong and predictable enforcement should be sought. Considering the issue at its simplest, the clear potential that such an image would harass, embarrass, bully, (perhaps sextort) or was simply created to degrade or demean (or was reckless as to this outcome), should lead to the simple conclusion of a credible and immediate likelihood of harm to the person depicted. Additionally, the more legal considerations of data privacy and data accuracy should only compound this. The decision to remove such content, with reasoning aligned with the actual policy expectations of the platform, should be supported. Relying on automated systems that automatically close appeals. In this case, the inconsistent application of the terms and conditions were then clearly compounded by reliance on a flawed automated decision-making process. This process seems to equate inability to carry out a timely review (by the recipient of the appeal), with the actual validity of the underlying issue of the appeal itself. This does not make sense. The decision to close both cases automatically by not responding to an appeal withing 48 hours, tends to confirm a clear lack of any subjective consideration. The Board should consider whether such a decision approaches a level of arbitrariness repugnant to Metas relevant obligations under legislation such as the Digital Services Act (DSA) within the EU.If Meta does not adequately address harms on their platform, who will?Like Metas own platforms, use of a domain name comes with expected standards of acceptable use of the domain name. These are contained in the Acceptable Use Policies of registries and registrars. Where a registrant uses a particular domain and fails to remedy a harm being perpetrated in a timely manner, then an expectation persists that the registry or registrar would intervene, where appropriate. Strong voices in the domain name multistakeholder community have consistently sought obligatory registry and registrar intervention i.e. suspension of entire domain name in such circumstances. This is less of an issue in circumstances where evidence shows a domain was likely registered for purposes that appear to be singularly abusive. The situation becomes exponentially more complex where such abuses arise, not as a primary purpose, but through an abuse of an otherwise legitimate use of the domain e.g. a service or platform that uses the domain to anchor their online presence. Many in our community feel that intervention in such circumstances, has too much of an impact, thus escalation to the platform, or service should be sufficient. In cases where the platform consistently and competently deals with such abuses, then the relationship is well maintained. Where no such consistency or reliability exists however, and persistent instances of reported, but unanswered, or ignored abuse are recorded, there remains strong calls for intervention by the DNS provider. One such recent call specifically sought an enhanced contractual obligation on all generic Top Level Domain (gTLD) Registries and Registrars to specifically deal with such persistent and systemic abuse by way suspension, stating registrars should be required to suspend or terminate a customer account after a defined number of reported and verified cases of abuse. . Such an obligation, were they successful, would have very serious consequences for all platforms, not just Meta. Tracing responsibilityIn the context of clear advocacy for enhanced domain name level intervention, it should be noted that the registrations of both Instagram.com and Facebook.com are registered at the Meta owned registrar, Regsitrarsafe, LLC. Changes in ICANN policy, if applied would place that entity at a difficult crossroads, as it would seem unlikely that Meta themselves would suspend their own domains, regardless of their additional contract expectations. In a connected vein, the registry operator, Verisign Inc, has stringent restrictions in their Cooperative Agreement with the US government, that requires them to operate the .com registry in a content neutral manner . Intervention in matters relating to content present on the Meta services, would not be compatible with such actions. ICANN themselves are not permitted regulate either the services using the DNS, or the content that such services carry or provide . Although these several layers of protection exist, unmistakable pressure from regulators , vehemently supported by some within the ICANN community, continues to seek to force registry/registrar level action in most, if not all, content related matters. This would certainly also encompass the incidents contained in the Case before the oversight Board. We encourage the Oversight Board to see the huge benefit that continued consistent and demonstrable action by your platforms, to address abuse to prevent the continued push for fundamental changes to the responsibilities. Platforms such as Facebook and Instagram should work in concert and not against other layers of the internet. Should such decisions be forced on DNS operators, the effect on the use of domain names for your platforms would be greatly affected. ConclusionExpecting consistency in application of Metas decision (or indeed the Oversight Board)s decision in all such matters, right or wrong, continues be critical. In this instance, there was an apparent failure to apply a policy across two major platforms, only to be compounded by a purely performative appeals process. In this context we submit that the Meta Oversight Board should therefore decide this matter in a clear and consistent manner and ensure that the action in such instances continue to be being taken at this appropriate level. Whereas the intervention of others in the DNS stack may sometimes be necessary, it remains the exception to the rule, and would prove costly and detrimental to Meta and its platforms. Persistent failures of the anti-abuse process of platforms such as Facebook and Instagram, as services which are delivered via a specific domains name, creates significant openings for those who challenge this status quo. We therefore would appeal to the Oversight Board to ensure that Meta does see fit to ensure the outcome of these cases support a strong anti-abuse action at the platform level that supports a collaborative approach to working with other online entities to ensure timely and appropriate action at the right level across the eco-system.

Link to Attachment

PC-27057

2024-007-IG-UA, 2024-008-FB-UA	PC-27058	United States & Canada
Case number	Public comment number	Region
Nina	Jankowicz	English
Commenter's first name	Commenter's last name	Commenter's preferred language
The American Sunlight Project		Yes
Organization		Response on behalf of organization

Full Comment

As a researcher of disinformation and technology-facilitated gender-based violence, the weaponization of deepfake pornography against women in public life is an issue that has concerned me for the better part of a decade. Deepfake technology has now become democratized, and applications that help users nudify any woman with a few clicks are widely available, affecting women of all walks of life, from government officials, to celebrities, to preteen girls. That Meta has any uncertainty whether non-consensual, AI-generated, explicit images of female public figures are prohibited by its Community Standards belies the platforms flimsy commitment to womens online safety and right to self-expression. These images should be removed, and Meta must improve its policies so that such images are not allowed to proliferate in the future. International ExamplesI first encountered fabricated non-consensual intimate imagery (NCII) in my work in 2017, while conducting research about Russian disinformation in Ukraine. In an interview with Svitlana Zalishchuk, a member of the Rada (parliament) elected along with a wave of other reformers after the 2013 Revolution of Dignity, I learned how cheap fakesimages faked with basic editing toolshad undermined her authority as an elected official: A screenshot began appearing on posts about [a speech

Zalishchuk gave at the United Nations] showing a faked tweet claiming that she had promised to run naked through the streets of Kyiv if the Ukrainian army lost a key battle. To underline the point, the message was accompanied by doctored images purporting to show her totally naked. It was all intended to discredit me as a personality, to devalue me and what Im saying, says Zalishchuk.Though the images were clearly fake, they followed Zalishchuk throughout her term in parliament as she continued her official duties. She suspected the Russian Federations involvement, as the image first appeared at the height of the early years of Russias invasion of Ukraine on pro-Russian message boards. Both Russia and its predecessor state, the Soviet Union, had a long history of employing kompromat (state-sponsored blackmail containing compromising material, sometimes of a sexual nature) to undermine regime critics. Particularly in patriarchal, traditional societies, faked photos and videos of this nature can end a womans career. After fake sex tapes were released in the Republic of Georgia, for instance, a politically-active woman depicted in the tapes all but retreated from public life. Through my research, I became convinced that mitigating gendered disinformationof which deepfake pornography and faked non-consensual intimate imagery is a subsetis an imperative not only for gender equality, but for the health of democracies worldwide. Fabricated Explicit Images in 2020 ElectionsUnfortunately, it is not only authoritarians who exploit entrenched misogyny to undermine the reputations of women with the temerity to speak out and stand up. In research I led investigating online abuse and disinformation against women running for office in the United States, Canada, the United Kingdom, and New Zealand in 2020, my team identified fake, sexualized images of a range of politicians. For example, in attempting to smear then Vice Presidential candidate Kamala Harris, users who opposed Harriss candidacy edited photos to make it appear that Harris was engaged in or about to engage in sexual activity, including depicting her on her knees in front of men, in revealing clothing, or with her legs spread. Such allegations and evidence of an allegedly scandalous sexual past are frequently targeted at women in politics in an attempt to humiliate and discredit them. Additionally, Harris, former New Zealand Prime Minister Jacinda Ardern, and Michigan Governor Gretchen Whitmer were all depicted in faked photographs alleging that they were secretly transgender. These photographs were digitally altered to make a subjects facial features appear more male or to add evidence of male genitalia to a subjects clothing.The secretly transgender narrative is a longstanding fixture of gendered online abuse. This rumor targeted Michelle Obama throughout and beyond the Obama Administration, asserting that she was formerly a man named Michael. At their foundation, these narratives tap into the trope of the duplicitous woman, implying that not only are transgender individuals inherently deceptive, but that this deception is responsible for the power and influence that these women hold. To this end, the narrative is also deeply misogynistic in its assumption that women cannot gain power without trickery. Proponents of these disinformation campaigns appear to assume that transgender identities, especially hidden ones, are so abhorrent that once the truth is revealed these women will lose all credibility and power.The 2020 study made another important finding related to image-based abuse online: while social media platforms frequently monitor text-based posts for abusive keywords, images are frequently disregarded as potential abusive content. Image-based abuse was part of what we dubbed malign creativitythe use of coded language; iterative, context-based visual and textual memes; and other tactics to avoid detection on social media platforms. Since this study was published, Meta has invested in tools to aid in detection and stop the amplification of image-based abuse. Considering the explosion of readily available AI-powered tools to generate NCII, Meta should ensure that detection tools are well-resourced, continually updated, and informed by ongoing contact with targets of deepfake pornography around the world. Personal TestimonyIn addition to studying the effects of deepfake pornography on womens online participation, I myself have been depicted in it. As a result of a widespread hate campaign against me in response to my appointment to lead a counter-disinformation body within the Biden Administration in 2022, I was targeted by anonymous individuals who created deepfake pornography of me and posted it to well-known deepfake forums. As if to underscore video makers compulsion to punish women who speak out, one of the videos[]depicts me with Hillary Clinton and Greta Thunberg. [] Users can also easily find deepfake-porn videos of the singer Taylor Swift, the actress Emma Watson, and the former Fox News host Megyn Kelly; Democratic officials such as Kamala Harris, Nancy Pelosi, and Alexandria Ocasio-Cortez; the Republicans Nikki Haley and Elise Stefanik; and countless other prominent women. By simply existing as women in public life, we have all become targets, stripped of our accomplishments, our intellect, and our activism and reduced to sex objects for the pleasure of millions of anonymous eyes.This targeting has persisted. Since launching The American Sunlight Project, a counter-disinformation advocacy organization that I co-founded, an internet user sarcastically posted about me, tacitly encouraging others to target me with explicit deepfakes again: Would be a real shame if someone deep faked her face onto some porn and spammed her Twitter account with it, he wrote. It would be absolutely terrible. I hope no one does that. That would be reprehensible.The Question of IntentI am lucky that the deepfake pornography in which I star has not yet proliferated beyond prominent deepfake forums, where men gather to exchange tips and perfect their art. But as much as they may assert they mean no harm in creating explicit deepfakes, those who engage in the creation of such non-consensual explicit images and videos are not engaging in artistic expression; they are engaging in attempted repression. As scholar Sarah Sobieraj argues, digital misogyny, including deepfake pornography, is aimed at protecting and reinforcing a gender system in which women exist primarily as bodies for male evaluation and pleasure and extends the history of attempts to curtail womens freedom to use public spaces as equals. She further explains that aggressors repeatedly draw upon three overlapping strategiesintimidating, shaming, and discreditingto silence women or to limit their impact in the digital public. The creation and amplification of deepfake pornography and other NCII serves all three strategies. Women depicted in deepfake porneven those of considerable resourcesfeel intimidated by their privacy being invaded. They are shamed; the titles of the videos themselves often expressly describe women being pounded, or humiliated. And they are discredited; prominent womens accomplishments, professionalism, and intellect are erased when they are reduced to sex objects. Creators of deepfake pornography often argue that they are not causing harm because their videoslike one of those in the case before the Meta Oversight Boardare clearly labeled as fake. This does not negate the malign intent of the content, nor the fact that these unilateral, non-consensual invasions of privacy and assaults on dignity have been viewed thousands of times. Similarly, some argue that deepfake pornogrpahy is simply sexual fantasy that causes no real harm. Professor of Law Clare McGlynn writes that creating deepfake pornography is neither art nor sexual fantasy, it is creating a digital file that could be shared online at any moment, deliberately or through malicious means such as hacking. She continues:Its also not clear why we should privilege mens rights to sexual fantasy over the rights of women and girls to sexual integrity, autonomy and choice. This is non-consensual conduct of a sexual nature. Neither the porn performer nor the woman whose image is imposed into the porn have consented to their images, identities and sexualities being used in this way.Toward Policy SolutionsMetas existing policies against Bullying and Harassment and Adult Sexual Exploitation appear to universally prohibit the content under review in this Oversight Board case. The Bullying and Harassment policy offers universal protections, regardless of public profile, against: unwanted contact that is directed at a large number of individuals with no prior solicitation, derogatory sexualized photoshop or drawings, and severe sexualized commentary. Similarly, the Adult Sexual Exploitation Policy prohibits: any form of non-consensual sexual touching, necrophilia, or forced stripping, including depictions. This policy unfortunately includes an exception for real-world art contexts, while confusingly prohibiting:Sharing, threatening, stating an intent to share, offering or asking for non-consensual intimate imagery that fulfills all of the 3 following conditions:Imagery is non-commercial or produced in a private setting.Person in the imagery is (near) nude, engaged in sexual activity or in a sexual pose.Lack of consent to share the imagery.The cases at hand appear to meet all of the criteria of both policies presented above. The Oversight Board should recommend to Meta the clarification of its Community Standards to include the express prohibition of the sharing of any non-consensual deepfake or AI-generated pornography or explicit imagery, regardless of public figure status or intent. Additionally, Meta should invest more resources in and prioritize NCII-related content moderation. The fact that one of the cases before the Oversight Board was raised simply because the report was automatically closed because it was not reviewed within 48 hours and remained up even after appeal is unacceptable, particularly during volatile election periods. In the time a victim is waiting for Meta to exercise its duty of care, such content could receive hundreds of thousands of views, be reported on in national press, and sink the public perception of a political candidate, putting her on uneven footing when compared with her opponents. In some countries, including India, where this case took place, it could even endanger her life. Meta might consider establishing special escalation routes for public figures, particularly those involved in politics, so that this content will undergo expeditious human review. Ultimately, when female public figures are targeted with AI-generated explicit imagery, it is not only the subject who is intimidated, undermined, and violated: it is all women. The message is not only one of repudiation of the public figure. It is a warning to any woman who adds her voice to public debate that she should think twice before doing so. If Meta seeks to create an environment where women can freely express themselves, it must ban this harmful content and ensure the ban is muscularly enforced.

Link to Attachment

PC-27058

2024-007-IG-UA, 2024-008-FB-UA	PC-27059	Asia Pacific & Oceania
Case number	Public comment number	Region
Lisha	Chheda	English
Commenter's first name	Commenter's last name	Commenter's preferred language
Rubaroo Breaking Silences Foundation		Yes
Organization		Response on behalf of organization

Full Comment

Deepfake pornography harms people and society in a multitude of ways. The ease of access to public figures photographs leaves them particularly vulnerable to their images being used non-consensually. In a culture where womens sexuality is often used to shame, silence,control, objectify or dehumanise them, deepfake pornography becomes a potent tool that is easily accessible and protects perpetrators from the accountability for causing the sexualviolation of another human being. Its use perpetrates and maintains the harmful systems thatviolate womens rights, privacy, dignity, agency and autonomy to how their images, bodies and being are used. It would deter individuals especially women from taking up space in any capacity online as there would be constant fear and anxiety around someone nonconsensually using their images, videos, voice and data to create pornography. It may leave the individual vulnerable to trauma and anguish to see their intimate selves non-consensually put up for public consumption. Further, the economic burden of the reversal of damage and redressal may have to be borne by the woman. Further, while we are grappling with protection of womens right to privacy, safetyand agency, imagine the harm deepfake pornography could wreak upon even more vulnerable populations like teenagers, children and infants. In September, more than 20 girls aged 11 to 17 came forward in the Spanish town of Almendralejo after AI tools were used to generate naked photos of them without their knowledge. (Matt Burgess, Wired, October, 2023) Women who are public figures may still have some means to better protect their rights and fight against their violation, but looking at it as a womens issue instead of a human rights one, maybe gravely erroneous.Several celebrities in India and the US have been targeted by the same and what is so disturbing is the short amount of time in which millions of views and thousands of shares are garnered and by the time the situation is flagged off and redressal mechanisms can kick into place, the harm is already done, without any guarantee that some content may have slipped through the gaps anyway. Only 4 US states have passed laws the specifically cover deepfakes (Arwa Mahdawi, The Guardian, April 2023). Sensity AI, a research company that has tracked online deepfake videos since December of 2018, has consistently found that between 90% and 95% of them are non-consensual porn. About 90% of that is non-consensual porn of women (Karen Hao, MIT Technology Review, February 2021). India ranks 6th among nations vulnerable to deepfake adult pornography (Singh Rahul Sunilkumar, The Hindustan Times, November 2023). Typically work needs to be done to cohesively work within legal frameworks of the countries and keeping in mind evolving global perspectives on the same. A proactive approach that does not just focus on restrictions but also inclusion, keeping in mind to build democratic spaces for people to exist online may be a holistic approach to combat the issue long-term. Having a survivor and gender-centric approach could be helpful. References:Nonconsensual Deepfake Porn is an Emergency that is Ruining Lives, Arwa Mahdawi, April 2023. (The Guardian)- https://www.theguardian.com/commentisfree/2023/apr/01/ai-deepfake-porn-fakeimagesArticulating A Regulatory Approach to Deepfake Pornography in India, Siddharth Johar,December 2023, Indian Journal of Law and Technology- https://www.ijlt.in/post/articulating-a-regulatory-approach-to-deepfake-pornographyin-indiaDeepfake Porn Is Out of Control, Matt Burgess, October 2023 (Wired)- https://www.wired.com/story/deepfake-porn-is-out-of-control/Deepfake Pornography is Ruining Womens Lives. Now the Law may Finally Ban it, Karen Hao, February 2021 (MIT Technology Review)- https://www.technologyreview.com/2021/02/12/1018222/deepfake-revenge-porncoming-ban/94% deepfake adult content videos targets entertainment industry celebs: Survey, Singh Rahul Sunilkumar, November 2023, The Hindustan Times. - https://www.hindustantimes.com/technology/94-deepfake-pornography-videostargets-entertainment-industry-celebs-survey-101699276557517.ht

Link to Attachment

PC-27059

2024-007-IG-UA, 2024-008-FB-UA	PC-27060	Central & South Asia
Case number	Public comment number	Region
nan	nan	English
Commenter's first name	Commenter's last name	Commenter's preferred language
Point of View		Yes
Organization		Response on behalf of organization

Full Comment

DID NOT PROVIDE

Link to Attachment

PC-27060

2024-007-IG-UA, 2024-008-FB-UA	PC-27061	Central & South Asia
Case number	Public comment number	Region
Vaishnavi	Sharma	English
Commenter's first name	Commenter's last name	Commenter's preferred language
The Dialogue		Yes
Organization		Response on behalf of organization

Full Comment

Oversight Board Public Comments: Explicit AI Images of Female Public FiguresThe Dialogue welcomes the opportunity to comment on the Oversight Boards Call for Public Comments on Explicit AI Images of Female Public Figures. Our submissions for the Call are listed below.The nature and gravity of harms posed by deepfake pornography, including how those harms affect women, especially women who are public figures. Nature and Gravity of HarmsDeepfake pornography, while not strictly a case of non-consensual dissemination of intimate images, may still bring about similar impacts. Non-consensual dissemination has been known to cause anxiety, panic attacks, severe emotions of humiliation and shame, potential unemployment,lower self-esteem, verbal and physical harassment, and stalking, among other consequences. To mitigate these intense emotions, victims have been known to resort to detrimental coping strategies, resulting in the display of various extreme behaviours like avoidance, denial, excessive alcohol consumption, fixating on their victimisation, and self-medication, etc.As is the case in one of the cases flagged by the Oversight Board, victims who are public figures often suffer a reputational loss and get chilled online. Recently, in November 2023, an Indian actress deepfake video went viral on social media. The incident highlighted the regulatory and enforcement gaps concerning information technology laws in the extant regulatory and policy landscape. In November 2023, the aggrieved actress responded on her social media account, I feel really hurt to share this and have to talk about the deepfake video of me being spread onlineSomething like this is honestly, extremely scary not only for me but also for each one of us who today is vulnerable to so much harm because of how technology is being misusedFurthermore, it is critical to note that such conversations are more complex in India because of issues of caste, class, gender, faith, and race. The complexity of the issue arises from three main factors: firstly, in instances of violence targeting marginalised communities, where breaches of privacy and consent occur as a byproduct, these violations are rooted in notions of humiliation and aggression; secondly, as a result, breaches that could be classified as clear-cut infringements of privacy and consent often go unnoticed, unreported, or unacknowledged as such; and thirdly, there exists a reluctance to engage with legal authorities or face outright obstruction, particularly in cases involving issues of caste, faith, race, and gender.Court adjudication resulting in additional harmsRecent legal cases involving online gender-based violence in India reveal several troubling trends in the judicial handling of these issues. Firstly, there is a tendency among courts to treat online violence as less severe and impactful than physical violence. Secondly, patriarchal notions still permeate the judicial discourse around online abuse. Lastly, there appears to be a limited understanding among courts of the nuances of privacy in the digital environment, particularly how it should be protected online.Similar to trials for sexual harassment and assault in India, pursuing criminal or civil cases against perpetrators in the courts may only aggravate the harm, both in terms of exposing oneself to the legal machinery (which cannot be anonymised entirely), repeatedly being forced to relive the initial shock and pain, and in terms of chasing a court case still seeped in patriarchy.Contextual information about the use and prevalence of deepfake pornography globally, including in the United States and India. Historical relevance of obscenity, indecency, and moral policingThe prevailing socio-cultural norms in India, often steeped in patriarchal attitudes and gender biases, contribute significantly to the online harassment of women, and create an environment where misogynistic behaviours and attitudes are normalised and perpetuated. These dynamics prioritise male dominance and control over women's bodies and voices, either via written laws or policy. To understand the prevalence of deepfake pornography online, we feel one of the most relevant jurisprudence to consider is that of obscenity and indecency in India. The obscenity and indecency jurisprudence, deeply rooted in the colonial legacy, significantly shaped and influenced Indian perceptions of morality regarding sexually explicit depictions. Influenced by Victorian and European ideals, India's post-independence pursuit of a civilised and regulated society led to increased regulation of domains related to obscenity and purity. Within this context, representations of women, often subjected to the male gaze in Hindi literature, underwent substantial transformations to align with evolving social and moral standards, aiming to establish a new hierarchy of power and integrate chastity with middle-class identity.Today, the continuation and formulation of numerous laws regarding obscenity and sexually suggestive representation remain strongly influenced by the concept of purification and moralising the diverse Indian society to create a governable and homogenous identity of society, by repressing and condemning behaviours and attitudes perceived as deviant. Often, women face the brunt of these laws - directly or indirectly. Considering the legal and constitutional history behind such laws (which are superimposed onto information technology laws), it is understandable that deepfake pornography, although a new method of shaming women, is gendered in nature, and disproportionality impacts women and members of other marginalised populations in India. Inadequacies in the Legal and Policy Landscape Lawmaking and policymaking tend to be either slow and/or reactive and consequently struggle to keep pace with the rapid evolution of deepfake technology. In November 2023, the government and policymakers intervened and called for social media intermediaries to implement proactive measures to mitigate the proliferation of deepfakes across online platforms. While diverse legislative provisions indirectly address issues concerning deepfakes, particularly in the context of pornography, their coverage is somewhat limited, addressing various fundamental rights only partially. Additionally, the current legal resources and grievance redressal mechanisms may not respond sufficiently promptly to these concerns. As a result, harmful content might continue circulating online and spread across platforms in various forms. Strategies for how Meta can address deepfake pornography on its platforms, including the policies and enforcement processes that may be most effective. Need to Have Clear Metrics: There is a need to establish clear criteria for what constitutes deepfake pornography, distinguishing clearly between content that uses the imagery of natural persons without consent and wholly synthetic creations that do not repurpose the images of specific real individuals but are generated from scratch to look convincingly lifelike. While the former type of deepfake pornography may amount to an infringement of an individuals privacy and sexual autonomy, the latter type may not necessarily aim to direct any personal harm but could have broader detrimental effects, such as the perpetuation of unhealthy sexual norms and the psychological impact on societal perceptions of privacy and consent.Prohibition of Promotion of Harmful Apps: Meta should consider actively looking into advertisements on Meta platforms that facilitate and promote the creation of non-consensual sexual images, such as nudification apps. Meta could consider enforcing disclosure norms for partnerships between content creators and synthetic media platforms to ensure consumers are fully informed about the nature of the service being promoted and potential risks. Safety Measures: Meta should consider implementing a default opt-out safety shield for images, particularly profile pictures, to prevent their misuse in the creation of deepfake content. Similar measures, like blurring AI-generated images and displaying warnings that such content is AI-generated on such content, can also be helpful. Labelling of AI-generated Content: Ensuring clear labelling on any AI-generated content across Meta platforms to ensure that users are always aware when they are interacting with or viewing content that has been artificially created or altered can also be considered. This labelling should be prominently displayed and easily understandable to ensure that all users can recognise the contents nature at first glance.Media Literacy and User Empowerment: Meta should consider promoting media literacy by educating users about the nature of deepfakes and their potential impact and empowering users to recognise and report deepfakes.Human Moderation and Oversight: Meta should ensure that content flagged by automated systems as a potential deepfake is reviewed by human moderators to consider the context. Collaborative Partnerships: Meta should consider engaging in global initiatives and collaborating with industry leaders to share knowledge and advance AI safety research specific to deepfake detection.

Link to Attachment

PC-27061

2024-007-IG-UA, 2024-008-FB-UA	PC-27062	Central & South Asia
Case number	Public comment number	Region
Dr. Shruti	Khare	English
Commenter's first name	Commenter's last name	Commenter's preferred language
DID NOT PROVIDE		No
Organization		Response on behalf of organization

Full Comment

1.Deepfake technology is an unavoidable unpleasant component of AI that may have been beneficial, but because so many people are abusing it, the consequences of fake crimes and illicit activities are severe. Victims, particularly women celebrities, face a variety of hardships, including emotional, social, psychological, and financial consequences. AI has been developed to become more capable at creating authentic photos and videos that are not true but are deep fakes of well-known female personalities. These changes can persuasively show people saying or doing things they never did. The people involved in doing so do not understand the gravity of the situation and the problems they are causing these celebs, because of the widespread usage of the internet and public attention, they are frequently the major targets of malicious exploitation. These photos, being highly digital and convincing, can be exploited to create potentially destructive scenarios or promote misleading narratives or stories or rumours, destroying reputations, damaging image, weakening credibility and even blackmailing. The horrible mix of sexism and technology causes these concerns by forcing damaging stereotypes and objectifying women. These kinds of activities, which should be strictly banned, are very much prevalent. The deep fakes are not just about visual damages, it's way too much beyond that. They trespass on women's private life, endangering them to frequent observation and unwanted speculation. The psychological significance of such attacks cannot be understated. Women who are public figures have a higher psychological weight. Tenacious attacks increase emotional breakdowns, emotional vulnerability and distrust, lowering the confidence with self-esteem and overall mental well-being. The endless attempt to clarify lies and avoid damage, exhausts the emotional energy, leaving little room for personal growth or professional fulfilment for these women. Victims experience sadness, depression, fear, hopelessness, and panic as they navigate with no idea of what may happen next. Loss of control over one's image and narrative can result in emotions of hopelessness and loneliness. The offensive act of deep fake Technology represents a very problematic crossover of mental health, technology, society and gender. Celebrities who are women who are in the public face a very brutal attacks of manipulation where they keep on fighting for their reputation. At this age of digital era proactive measures are needed of the hour, to protect privacy with strong upholding dignity and positive mental health. 2.Deep fake Technology has been widely used in the politics as well as the entertainment Industries particularly in the United States. The entertainment industry has seen the rise of the fake videos and photos of famous people/celebrities, which are digital limitations of public figures. It causes a lot of problems regarding social, financial, and moral problems with the concern regarding the privacy and the consent of the celebrity. Not only in the entertainment or music industry even in the political area malicious individuals use the fakes videos or images to tarnish the image of multiple political personalities, which inevitably raises the concern about the authenticity of their work and further more misleading the public about their political leaders.India, like many other countries, has faced difficulties related to deep fake pornography and derogatory content. The fast development of social media and digital platforms, has fast-tracked the spread of inappropriate information. Recently multiple deep fake images and videos have been noticed in India in which case individuals, especially the celebrities, have been the target with digitally altered videos and photos, to harm the reputation of those celebrities especially women. This has created a sense of insecurity and fear among not only celebrities of all genders but also among the public. Women, especially, are now more concerned about privacy and are scared regarding the potential harmful effect of the deep fakes. As India is culturally diverse and with most religious societies and cultures, hence more prone to face the repercussions of malicious work of deep fakes creating societal tension and spreading misleading information. The malicious activities that can be done using the deep fake Technology can be very diverse like gender bias, objectification of women, manipulation regarding the elections, discord among the community or even incite hatred.3.Deepfakes and offensive content should be explicitly banned by Meta. Users should not be allowed to publish, share, or modify content in any way to produce offensive or deepfake content. Clearly specify what constitutes prohibited content, including deepfakes, in Meta's terms of service. Ensure that users are made aware of these conditions when joining. Meta should use advanced (exclusively used by Meta) AI-driven strategies to precisely identify modified content. These technologies can identify questionable information and flag it up for further scrutiny by human moderators. Hiring skilled human moderators to carefully review reported content. Human judgment is critical for eliminating highly sophisticated deepfakes from authentic content. Educating the users regarding the identification of the malicious contain of deep phase is absolutely necessary. As soon as any content is flagged by any user regarding the deep fake malicious content, immediate removal or banning of the particular content should be done until proved authentic and acceptable content by the review Team. The content should remain band temporary until and unless the team has reviewed it manually and the 48 hours time limit should be discredited. Because of that 48 Hour timeline there is a huge possibility of offensive content being shared as it is falling between the cracks of scrutiny. And force and reinforce penalties ban those who violate the guidelines of clean content, even if it is required to permanently ban the users who are repeat offenders. Meta can also take help from the government and the law enforcement, after tracking down the users who repeatedly upload deep fake content. 4.The rules and policies given by any company, in this case meta, will eventually become ineffective if we are not educating the users enough. The company needs to educate users on community standards and satisfactory policies and should promote responsible uploading and reporting of inappropriate content, especially if they are unsure of its origin. I am pretty sure that meta is using a lot of strict automated filtering for any of the derogatory or sexualized or inappropriate content, but that's not enough, evidently. Multi layers of protection which includes automated filtering with strict human reviews are the need of the hour. Employing a team of experienced moderators to manually review and report the content flagged as inappropriate. This process of human review is much more capable of identifying the minute nuances or they can understand the indirect inappropriate remark that an automated computer programs system may overlook or miss out on. But as humans are fallible, meta needs to conduct regular audit over the content checking manually done by the team to ensure the compliance.5.The company is going to face a lot of challenges and problems if they are going to only rely on the automated systems that automatically close the appeal in 48 hours if no review has taken place. Firstly, prompt closure of the appeal made by the user without any proper scrutiny will definitely eventually result in poor User experience. Users may tag the site as unfair or irresponsible regarding the sensitive information, especially closure or solution. The computerised systems that are responsible for this kind of appeal can face a lot of challenges to handle this kind of context dependent sophisticated content. The automated system can mark inappropriate data as acceptable or authentic due to its computational limit in detecting the subtleties of expression. These automated systems can also flag out any harmless and authentic content as inappropriate. Meta should extend the review period, because 48 hours may be too short for this kind of scrutiny of complex problems of inappropriate data content uploaded on the websites. There should also be multilayer analysis with the combination of computerised automated systems and human level scrutiny by the teams. For the clear-cut concern, the automated filtering will get out and furthermore it will be reviewed by the team of experienced executives. Also, meta can prioritize the specific content flagged out by the users. This will give them the feedback of the users and the company can use that information and utilise the feedback from the user for improving automated systems. One of the other ways can be providing the very clear-cut Apparent guideline to the uses as well as for the company. This includes providing clear and explicit examples of what is considered as acceptable content and how the appeal processes work.

Link to Attachment

No Attachment

2024-007-IG-UA, 2024-008-FB-UA	PC-27063	Asia Pacific & Oceania
Case number	Public comment number	Region
Saurav	Bhattarai	English
Commenter's first name	Commenter's last name	Commenter's preferred language
Digital Rights Nepal		Yes
Organization		Response on behalf of organization

Full Comment

DID NOT PROVIDE

Link to Attachment

PC-27063

2024-007-IG-UA, 2024-008-FB-UA	PC-27064	Central & South Asia
Case number	Public comment number	Region
Withheld	Withheld	English
Commenter's first name	Commenter's last name	Commenter's preferred language
Withheld		No
Organization		Response on behalf of organization

Full Comment

DID NOT PROVIDE

Link to Attachment

PC-27064

2024-007-IG-UA, 2024-008-FB-UA	PC-27065	Europe
Case number	Public comment number	Region
Jeffrey / Beatriz / Sarah	Howard / Kira / Fisher	English
Commenter's first name	Commenter's last name	Commenter's preferred language
UCL Digital Speech Lab		Yes
Organization		Response on behalf of organization

Full Comment

Oversight Board Public CommentExplicit AI Image Cases <See attached PDF for full text, including footnotes and references>This public comment addresses the following aspects of the two cases on explicit AI images of female public figures selected by the Oversight Board:1.The nature and gravity of harms posed by deepfake pornography including how those harms affect women, especially women who are public figures;2.Strategies for how Meta can address deepfake pornography on its platforms, including the policies and enforcement processes that may be most effective;3.Metas enforcement of its derogatory sexualised photoshop or drawings rule in the Bullying and Harassment policy, including the use of Media Matching Service Banks.1. The nature and gravity of harms posed by deepfake pornography including how those harms affect women, especially women who are public figuresThe cases involve AI-generated images of nude women that were created to resemble public figures one in India and one in the US. Both cases should be conceived as instances of non-consensual intimate deepfakes (NCID). NCID is a distinctive form of harmful deepfake that falls under the umbrella of image-based sexual abuse, i.e. the non-consensual creation, distribution or threatened distribution of private sexual images. The harms of NCID are no different from the well-established harms of non-synthetic image-based sexual abuse, a problem that existed long before AI. NCID inflicts the very same harms as abusive images created without AI. Like traditional image-based sexual abuse, NCID inflicts both individual and collective harms. Individual harms include violations of mental and physical integrity, dignity, privacy, and sexual expression. Social, collective harms include the risks of normalising non-consensual sexual activity and contributing to a culture that accepts rather than reprimands creating and/or distributing private sexual images without consent. The technology used to create the abusive media (AI or not) is irrelevant when it comes to these harms. Conceiving NCID as a form of image-based sexual abuse provides better understanding of and terminology for this phenomenon, than that suggested by terms such as deepfake pornography or AI-generated pornography. The term deepfake pornography connotes a form of erotic material, rather than a form of abuse, and risks conflating legal pornography with properly prohibited content. This parallels the inadequacy of using the term pornography to describe child sexual abuse material (CSAM), or revenge porn to describe the malicious dissemination of non-synthetic forms of non-consensual intimate content. Assigning the label pornography to NCID trivialises the content and shifts the focus to its usage or interpretation, rather than the victim-survivors experience. In contrast, the term non-consensual intimate deepfakes (NCID) accurately describes the harmful nature of the content and the impact on victim-survivors. Like other forms of image-based abuse, non-consensual intimate deepfakes disproportionately impact women. An industry report based on the analysis of 14,678 deepfake videos online indicates that 96% of them were non-consensual intimate content and that 100% of examined content on the top five deepfake pornography websites were targeting women. Image-based sexual abuse harms both private and public figures. In fact, NCID targeting public figures risks a further, distinctive harm to democracy: incentivizing women not to run for public office. 2. Strategies for how Meta can address deepfake pornography on its platforms, including the policies and enforcement processes that may be most effectiveMeta should treat non-consensual intimate deepfakes on its platforms the same way it treats non-synthetic image-based sexual abuse. Meta should prohibit all forms of image-based sexual abuse, regardless of origin or creation method. Because the harms posed by NCID are not different from any other forms of image-based sexual abuse (as noted above), there is no need or justification for a standalone policy targeting only synthetic abusive content. This approach eliminates loopholes based on content type and avoids the difficult task of consistently differentiating real from synthetic content.We pressed a similar point about Metas manipulated media policy in our public comment on 2023-029-FB-UA (Altered Video of President Biden); the Oversight Board echoed this concern, and we are gratified that Meta is now proposing to replace its standalone policy on Manipulated Media (that we previously criticised for being too narrow and not fit for purpose), clearly stating that we will remove content, regardless of whether it is created by AI or a person, if it violates our policies against voter interference, bullying and harassment, violence and incitement, or any other policy in our Community Standards. What this means is that the effectiveness of Metas policies to address NCID hangs on how well it defines and enforces its general rules against image-based sexual abuse.Meta should revise its Bullying and Harassment policy to include an explicit prohibition on image-based sexual abuse. Specifically, Meta should prohibit the posting of unwanted intimate media depicting the likeness of an individual, whether real or synthetic. The rationale for such a prohibition already exists in the policy, which states that Meta will remove content thats meant to degrade or shame, including, for example, claims about someones personal sexual activity. The revision is necessary because the existing Bullying and Harassment policy is insufficient to capture all image-based sexual abuse. For instance, the tier-one rules, designed to protect all users, prohibit unwanted contact that is sexually harassing; yet image-based sexual abuse is often not communicated directly to the persons depicted, and so may not qualify as harassment. Likewise, it is doubtful that image-based sexual abuse straightforwardly falls within existing attempts to protect everyone against attacks based their experience of sexual assault, sexual exploitation, sexual harassment, or domestic abuse, statements of intent to engage in sexual activity or advocating to engage in sexual activity, or severe sexualized commentary. The tier-one rules come closest to prohibiting image-based sexual abuse in protecting against derogatory sexualized photoshop or drawings; but this policy is drawn far too narrowly. It should prohibit all forms of image-based sexual abuse, including synthetic non-consensual intimate content. The tier-two rules, which apply to minors, private adults, and limited-scope public figures, are somewhat broader, prohibiting content sexualizing another adult. But it is unclear what constitutes such content, and it is not a perspicuous way of describing image-based sexual abuse.Metas policy against image-based sexual abuse should not distinguish between private and public adult figures. We stress this point because the Bullying and Harassment policy currently provides weaker protection for public figures. This differential treatment is arguably misguided in general; but for present purposes our narrower point is that it is clearly misguided in the case of image-based sexual abuse. Accordingly, the prohibition on image-based sexual abuse should appear in Metas tier-one rules (applicable to all), rather than its tier-two rules. While the current tier-two rules might encompass such abuse under content sexualising another adult, the removal bar for public figures is high, limited to severe cases or and those in which the public figure has requested removal. So even if the current policy were sufficient to protect private adults, making removal more difficult for content depicting public figures would be a mistake, given the harm of such speech for women in public life.Finally, Meta should clarify its Adult Nudity and Sexual Activity policy, which should have justified removal in both cases under discussion here. The current policy involves a qualified ban on imagery of real nude adults and imagery of sexual activity. It should clarify that synthetic depictions of nude adults, or sexual activity (e.g., through AI-generated deepfakes) are encompassed under this rule. The chief rationale for this clarification is that it serves the existing aims of the Adult Nudity and Sexual Activity policy; a subsidiary rationale (relevant to the discussion here) is that it provides a second layer of protection against image-based sexual abuse. Were such a clarification in place, this policy could have been used to remove the deepfakes of the Indian public figure and the American public figure (especially since the policy includes a clear prohibition of images depicting someone [s]queezing female breasts). 3. Meta's enforcement of its "derogatory sexualised photoshop or drawings" rule in the Bullying and Harassment policy, including the use of Media Matching Service BanksThis rule is drawn too narrowly; as noted above, Meta should prohibit all forms of image-based sexual abuse, regardless of origin or creation method. Manipulating existing images with traditional photoshopping software, or manually drawing images, are swiftly becoming the least common modes of image-based sexual abuse. Again, Meta should prohibit the posting of unwanted intimate media depicting the likeness of an individual, regardless of the technology deployed to produce the relevant image. Metas current enforcement, which relies on automated systems that automatically close appeals in 48 hours if no review has taken place, is flawed and unable to offer victim-survivors adequate protection or redress. Metas response in the first case (involving a depiction of an Indian public figure) illustrates the limitations of reactive tools in tackling image-based sexual abuse. (It is unclear if this automatic closure applies for all policy areas, or only some.)The use of Media Matching Service banks is justified, given how swiftly this content can cause harm and the limits of relying on ex post complaints. Ex post removal of violating content, after it is reported by victim-survivors, will often be too late; much of the harm will have already been done. As Meta has itself noted with respect to victims, the damage they experience increases the longer the images remains online. The literature has examined the frequent failure of notification and takedown requests, notably because victim-survivors often dont report abuse. When it comes to fully synthetic content, individuals being depicted may be completely unaware of the materials existence since they likely were not involved in its creation, further hindering their ability to report it. In addition, the ease of resharing digital media makes complete removal after distribution nearly impossible and violating images often persist online, especially when hosted outside the victim-survivors jurisdiction. The use of the Media Matching System in the second case (involving the American public figures depiction) resulted in better protection than in the first case (involving the Indian public figures depiction). Undeniably, automated and algorithmic content moderation tools raise valid concerns. These concerns stem from AIs documented biases and, crucially, its struggles with context, subtlety, sarcasm, and subcultural meaning, which can lead to the disproportionate silencing of certain voices. However, image-based sexual abuse is a category of harmful content involving less nuance in its designation, we venture, than other categories. Given the potential of the Media Matching Service Bank in countering NCID, Meta should expand its use, while continuing to study its effectiveness and collateral costs.Submission Prepared By:Beatriz Kira is Lecturer in Law at the University of Sussex, member of the Sussexs Law and Technology Research Group, and Fellow in Law & Regulation at the Digital Speech Lab at University College London.Jeffrey W. Howard is Director of the Digital Speech Lab and Associate Professor of Political Philosophy & Public Policy at University College London, and Senior Research Associate at the Oxford Institute for Ethics in AI. Sarah A. Fisher is Fellow in Philosophy at the Digital Speech Lab at University College London, and (from autumn 2024) Lecturer in Philosophy at the University of Cardiff.About the UCL Digital Speech LabThe Digital Speech Lab at University College London hosts a range of research projects on the proper governance of online communications. Its purpose is to identify the fundamental principles that should guide the private and public regulation of online speech, and to trace those principles concrete implications in the face of difficult dilemmas about how best to respect free speech while preventing harm. The research team synthesizes expertise in political and moral philosophy, the philosophy of language, law, social science, and computer science.About the University of Sussexs Law and Technology Research GroupAn international hub for research, teaching, and engagement in law and technology, the University of Sussexs Law and Technology Research Group houses leading scholars with expertise in technology and information regulation, global governance of technology, intellectual property, and legal innovation. We conduct cutting-edge research through collaborative projects with policymakers, civil society organisations, and industry leaders.

Link to Attachment

PC-27065

2024-007-IG-UA, 2024-008-FB-UA	PC-27066	Central & South Asia
Case number	Public comment number	Region
Aparajita	Bharti	English
Commenter's first name	Commenter's last name	Commenter's preferred language
The Quantum Hub		Yes
Organization		Response on behalf of organization

Full Comment

TQH Submission to Metas Oversight Board on Cases involving Explicit AI ImagesOverview and Executive SummaryBased in India, The Quantum Hub (TQH) works extensively on issues relating to technology and intermediary regulation as well as gender-responsive policymaking. Since this case lies at the intersection of our work, we make the following submission by drawing on insights from Indian social-cultural realities. Given our location and work, we have accorded primacy to Indian contextual analysis and the Indian case amongst the two identified in the problem statement. Though our analysis of this case, we recommend that Meta:Have a clear position on AI generated sexualised or derogatory imagery not being permitted on its platform, and prioritize this objective in content moderation practices;Improving the design of reporting tools available on Instagram (in this case) to allow users to better navigate the tool and make more specific reports. Specifically, users must be provided ways to describe the reason for their report, and add additional context and information; Consider slowing the spread of content that starts to be reported by users (particularly in this category) as an interim protective measure; and,Reconsider and discontinue its policy of automatically closing appeals within 48 hours. Response to the Boards

QueriesQuery 1The nature and gravity of harms posed by deep fake pornography including how those harms affect women, especially women who are public figures. Query 2Contextual information about the use and prevalence of deep fake pornography globally, including in the United States and India. (addressed together below)Deepfake pornography or synthetically generated sexualised or derogatory content (for brevity, referred to collectively as SDC in this Submission) is the latest facet of a long standing online gender based violence (OGBV) pandemic that women around the world have been combating. Data supports that these digital creations are used most often to target women, often resulting in severe repercussions. Different studies independently arrive at the same conclusion that an overwhelming majority of deep fake content (upwards of 90%) targets women by generating SDC. In the misuse of synthetic content generators, there exists a troubling capacity to diminish and intimidate individuals, particularly women. Creation of SDC in concerning volumes therefore presents not just ethical dilemmas but also credible threats of real world harm and stifling women's expression, particularly in Indian settings. Contextual factors: Dominant social norms in India place a premium on womens modesty, reputation and honour. Bad actors often target notions of decency as a means to attack women, ranging from threats to sexual violence, disclosure of intimate details or imagery, aspercions of promiscuity, etc. As a result, individuals targeted on this front are more likely to incur social censure. SDCs form a new weapon to target these vulnerabilities and damage womens social capital while living in Indian realities. It is plausible that women targeted by SDCs experience ostracisation, shaming, and secondary harassment. For this reason, Indian laws (under the Indian Penal Code, Information Technology Act, and to an extent the Indecent Representation of Women Act) criminalize content that fits the SDC description. In fact, the gendered harms of deep fakes have also been expressly recognised by the Indian government, which is exploring regulatory solutions to prevent deep fake proliferation. Perpetrators can exploit deep fake technology to threaten, blackmail, and manipulate victims, exacerbating the harm inflicted. This technology poses a significant threat as perpetrators can leverage deepfakes to instigate and perpetuate cycles of abuse, similar to other forms of non-consensual intimate image sharing. The social pressure, especially outside cosmopolitan or urban contexts, can often be so strong as to result in discrimination in professional or social settings, and severely damage personal relationships. This places SDC squarely on the OGBV continuum, representing not only an act of violence itself but also a catalyst for escalating threats against women.SDC drives up online toxicity: In online contexts, SDCs can be a potent kernel that attracts and fuels sexist narratives and harmful online engagement. The tendency for toxicity to take over even benign content must be well understood and appreciated in crafting appropriate policy responses. Deepfake videos featuring Swift on Twitter(X) accumulated over 27 million views and over 260,000 likes within a span of 19 hours before the account responsible was suspended. Subsequently, X even blocked searches for 'Taylor Swift' on the platform. Bad actors can use their networks to widely disseminate SDC and use reactions and comments to bully the subject. When coupled with users ability to identify the subject and their place of residence and/or work, the online harassment can quickly translate to threats of real world harm. Overall, SDCs contribute heavily towards a climate of apprehension for women both online and offline. SDC should therefore be considered high-risk content and a form of OGBV that must be proscribed.Heightened vulnerabilities of public figures and politically active/opinionated women: These risks are exacerbated in the case of women who are public figures (such as politicians, activists, journalists, entertainment sector celebrities), who are already at the forefront of sexist attacks and reprisals when commenting on subjects in a charged social-political environment. There is a documented track record of Indian women public figures experiencing threats of aggravated violence (death, rape, etc.) and toxic abuses as a direct response to the expression of their political or social views online. Abusive online behavior against an Indian journalist has also resulted in police action. In the context of alarming levels of online toxicity directed at vulnerable groups, AI tools and SDC fan the flame by providing inauthentic material that spurs further harassment. For instance, images of an Indian filmmaker were doctored by swapping his image with that of a female model in order to trigger homophobic and derogatory comments aimed at his identity.In this context, SDC serves as a formidable barrier for women with dissenting views or those from marginalized communities, preventing them from speaking up without fear of manipulation or retaliation thus pushing women away from the political or public arena. Without the strong assurance of countermeasures against SDC, women (especially young women) can be forced to consider leaving the public arena and stop airing their views to avoid reputational harm. A policy on SDCs envisioned by Meta should be crafted from the perspective of an ally to women and other victims of SDCs, rather than an impartial observer of the abusive potential of SDC. As a first step, this involves developing a clear and normative understanding of SDCs to be critically harmful, and prioritizing the prevention of SDC on Meta as the objective of any content moderation policy. Traditional GBV is historically underreported on account of survivors having to bear the cost of seeking justice or redressal that is often disruptive to their lives. Online platforms have the unique opportunity to change this paradigm as it applies to certain forms of OGBV, including SDC abuse. Changes in Metas reporting and content moderation practices that result from this case should lead to increased ease of reporting SDC and enhanced efficiency in SDC removal. Query 3Strategies for how Meta can address deepfake pornography on its platforms, including the policies and enforcement processes that may be most effective. Metas content moderation efforts should aim to prevent the discovery of SDC on its platforms. At a time where most deep fakes fall within this category, achieving this objective requires a combination of effective user reporting, Metas timely reviews, and automated moderation techniques. We find that the way Instagram allows users to report SDC does not help users identify the most appropriate reason for their report, nor does it allow users to provide any details or explanations that can add much needed context to their report. Though the problem statement clarifies that the content was found to violate the Bullying and Harassment Standard, the corresponding reporting label does not indicate that Meta removes "derogatory sexualised photoshop or drawings" under this policy. Rather, the tool only refers to threats to post intimate images of others. Understood normally, this does not describe the SDC content at-issue. On the other hand, when a user begins to report SDC, at least three broad categories (Nudity or sexual activity Bullying or harassment, and arguably Hate speech or symbols) all appear to be viable heads to describe the content report, with none of these containing descriptions that correctly describe the content. Furthermore, it is concerning that the Bullying label enumerates content of differing risk potential (posts that shame other people are relatively low risk by comparison to NCII, that is featured in the same list.) Even so, users have no ability to specifically identify the reason that most accurately describes the reason for their report. This is unlike other popular platforms (like X) that allow users to be more specific while reporting content. Finally, users have no ability to provide additional details, information or context to their report. Allowing users to briefly describe their reasons for reporting content will provide Meta crucial context that could help action contextually high-risk content before it can create further harm. Today, there is also technical capability in automated systems to scan user-written descriptions and assess the category, nature, and urgency of a report, that can benefit both human moderators as well as automated moderation systems. By limiting users from providing such information, there is lost potential in building robust content reporting and assessment frameworks. Since Meta relies significantly on user reports to action bullying content, design inefficiencies in its reporting tool compromises the first layer of defense in preventing the damaging content of this nature from reaching a wider audience. Recommendation: Empowering users to provide the greatest level of detail (as they are able or willing to) in their reports will help Meta speed up and appropriately prioritize its content moderation efforts. Given that the overwhelming majority of synthetic content is of SDC description, Meta should provide a specific way to report such high priority content in a manner that is clearly labeled, well understood by users, and with cross links to other plausible categories (even if the user inadvertently chooses another reporting head).Query 4Meta's enforcement of its "derogatory sexualised photoshop or drawings" rule in the Bullying and Harassment policy, including the use of Media Matching Service Banks. Metas inability to review and action the content in the first case of the problem statement reveals gaps in its enforcement of the Bullying and Harassment Standard. Aside from a flawed appeal mechanism and ineffective reporting options (covered elsewhere in this Submission), the omission to review the complaint/appeal could be attributed to the improper prioritization of reviewer resources and failure to implement temporary or interim restrictions on the content pending review.SDC bears the risk of doing real world harm to users that is identical or closely resembles the harmful effects of NCII (non-consensually shared intimate imagery). Therefore, in the broader gradation of priorities for content moderation, SDC should occupy a high-priority position, ideally, around the same level of attention that is paid to NCII reports. Instituting interim measures to prevent SDCs harm even until it reaches a content moderator can be a useful approach. Neglecting to slow content that received community notes on X has been assessed as a limitation in its crowdsourced content moderation efforts. While one should always be cautious of recommending interim measures that can restrict users from expressing themselves to as large a group as they desire, we believe there are exceptions to this rule. SDC is an apt candidate for content that should ideally be slowed down pending review.Content that is wrongly flagged under a label meant for SDC is most likely to be found to fall within benign themes of artistic expression or health or educational related content. Normally, there is no urgency associated with speech for such expression, and therefore, there is no countervailing interest that is impacted if its circulation is limited, and subsequently restored. Further, SDC is likely to attract engagement that exacerbates violence and harassment intended towards women. So long as the content remains freely available on the platform, reposts (on IG stories), comments, user-to-user sharing are all avenues by which bad actors can maximize the reach of the material which all contribute to the ultimate harm borne by the user/individual in question. Therefore, delayed action or poor practices in response to reports under this label serves to increase the likelihood of toxic discourse on Meta platforms - which is better addressed if temporary restrictions are placed on the velocity of the spread of the post. Therefore, interim or protective measures that automatically kick in upon receipt of a predetermined level of reports under this flag should be considered. Without slowing suspicious content down, countermeasures such as fact checks (in case of misinformation) were not found to be effective. This coupled with our suggestion of improving the quality of user reporting by giving additional details can help Meta appropriately prioritize content for review. Finally, Meta has disclosed that a bulk of its enforcement of the Bullying and Harassment policy activity is proactive. As covered in response to the preceding query, we believe that Metas proactive enforcement of the Bullying and Harassment policy must include automated tools. Metas automated content moderation tools have shown some promise in correctly identifying intended content on the platform. More sophisticated versions of such tools can be effective in proactively identifying content that fits the SDC description. Incorporating AI in the identification process based on existing banks of SDC reported content could increase the accuracy of such exercises, with diminished risk of legitimate speech being censored. We caution against over-reliance on Media Matching Service Banks (MMSB) as a way for Meta to proactively identify SDC. While MMSB can be effective in preventing secondary transmission of offending content (like in the second case of the problem statement), they are likely ineffective at preventing SDC from Meta unless the exact same image has been reported and found to be offending. AI generated sexual imagery, by its nature, is likely to be novel and plenty given that it is easy and cheap to generate at scale. Query 5The challenges of relying on automated systems that automatically close appeals in 48 hours if no review has taken place.We strongly advocate for Meta to reconsider and discontinue the policy of automatically closing appeals, especially when Meta itself fails to address the user complaint. From a user perspective, automatic closures do not contribute to resolution of complaints, but rather increase the burden on affected users to repeatedly track and report triggering content. The system does not bring with it any benefits that are immediately clear, when viewed from the perspective of healthy content moderation practices. Having said that, we appreciate considerations of resource limitations that come with having to review a large number of reports that need appropriately trained human resources to address. Rather than closing the complaints/appeals, Meta should consider apportioning reviewer time to close out high priority user reports, and have automated tools / technical aids that can work to streamline reviewers burden. Even considered from a regulatory perspective, automated closure of complaints (when no review is conducted) is more likely to be treated as Metas inaction. In the context of content which is likely illegal or harmful, regulators are less likely to be sympathetic to Metas inaction, as compared to even slower review and content action. Founded in 2017, The Quantum Hub (TQH) is a multi-sectoral public policy research and consulting firm based out of New Delhi, India. Within our technology policy practice, we have been working on various digital economy and governance issues with a variety of stakeholders, and have closely tracked discussions around data protection and online safety. Within our gender practice, we work on womens labour force participation, womens representation in private and public sector leadership and womens overall health and wellbeing. We often work at the intersection of these two practices to track and study womens role and participation in an increasingly digital world.

Link to Attachment

PC-27066

2024-007-IG-UA, 2024-008-FB-UA	PC-27067	United States & Canada
Case number	Public comment number	Region
Vrinda	Marwah	English
Commenter's first name	Commenter's last name	Commenter's preferred language
DID NOT PROVIDE		No
Organization		Response on behalf of organization

Full Comment

Women, sexual minorities, and other marginalized groups are soft targets for AI generated deepfakes. This makes deepfakes the latest in a long series of sexual harassment and bullying modalities, including cyber stalking and abuse. But AI generated deepfakes are not "more of the same" when it comes to harassment and misinformation. It is now widely accepted that platforms are loaded dice, and that algorithm-generated echo chambers dangerous undermine democracy. Data suggests that women are the overwhelming targets of deepfakes, but that politically motivated deepfakes get the most media and policy attention. However, these two categories are not mutually exclusive. Women are often targeted even by politically motivated deepfakes. In India, women from opposition parties and the minority Muslim community have been especially targeted by AI deepfakes-- unsurprising for a country descending into Hindu majoritarianism and cult-of-personality authoritarianism, orchestrated by a ruling party, the BJP, that has notoriously deep pockets and a zealous IT cell. If the BJP does its job right, then they will no longer need to source numerous deepfakes: with enough poison, people will do it for them. What this means for platforms like facebook is that the problem is set to grow. And that they can rely on governments to be short-sighted and neglectful, or malicious and opportunistic. To both protect and advance itself in such a milieu, facebook will have to GENUINELY do better. Tech platforms should lead the way. They cannot fall back on "self-regulation" because that has been shown, time and again, to be a cop out. Facebook ought to take leadership of a situation that they both bear responsibility for and have expertise in. This means taking immediate and long-term steps towards putting in the necessary resources: such as enough people who can check content, at least flagged content. In the India case under consideration here, so much harm would have been averted if the case was not just automatically closed again and again, but actually investigated. This is neglect after a problem has been reported. But facebook also needs to take preventative steps. Some jurisdictions now require all tech-modified content to be clearly marked as such. This is necessary for ALL jurisdictions-- and not just the ones with enough clout to seek these changes. In fact, this needs to be accompanied by a public education campaign about tech literacy. And facebook needs to put serious R&D dollars into developing an efficient, counter-balanced, and somewhat flexible system for monitoring, and eradicating deepfakes. This system will require researchers and content moderators as much as algorithms. Another way for tech to lead the way with this is to investigate other elements that would go into creating a desirable online climate. The private sector cannot be expected to tackle problems as complex and evolving as AI deepfakes alone. We also need context appropriate criminal laws, state capacity to implement those laws, and the political will to genuinely defend internet freedoms while protecting the most vulnerable. There is no reason tech should not sponsor (without influencing) an inter-disciplinary study of what ALL sectors of society should do to work with tech companies to combat technological crises. This can take the form of a best practices framework that would carry considerable moral if not regulatory weight. There is urgent need for leadership that can take everyone along, instead of what we have right now, which is a lot of passing the buck. Even for its own PR, tech ought to be able to say: "This is what we need to do, this is what the govt needs to do, and this is what people need to do. And we have taken the necessary steps but we are not being supported by XYZ."

Link to Attachment

No Attachment

2024-007-IG-UA, 2024-008-FB-UA	PC-27068	Asia Pacific & Oceania
Case number	Public comment number	Region
Withheld	Withheld	English
Commenter's first name	Commenter's last name	Commenter's preferred language
Withheld		No
Organization		Response on behalf of organization

Full Comment

Hi team OSB, very glad that you have selected this case! Im looking forward to seeing the Boards assessment and outcome. Sharing a few top of mind thoughts on how such harms can affect women in India. There could be a range of negative consequences because of deep fake pornography or nudity - financial, emotional, and physical consequences. Culturally, a significant portion of India is conservative when it comes to nudity - more so when it comes to female nudity. Apart from the emotional harm to the victim itself, allowing deep fakes of celebrities or public figures involves the risk of normalising the practice of generating such media - amplifying the risk of such generated media of private individuals also circulating at a larger scale. Below are two examples illustrating how even partial nudity of a male celebrity sparked outrage in parts of the country.(https://www.vogue.in/culture-and-living/content/ranveer-singh-nude-photoshoot-paper-magazine-sparked-nationwide-outrage-fir, https://timesofindia.indiatimes.com/readersblog/voices-from-within/nudity-vs-morality-44057/)The non-consensual publication of any nude, partially nude, or intimate photos of a woman, whether real or generated has the potential to cause severe emotional and physical harm to the victim. There have been instances where women in India died by suicide after their morphed intimate images were published online. See examples below.https://www.newindianexpress.com/states/tamil-nadu/2016/Jun/28/girl-commits-suicide-after-morphed-pics-appear-on-facebook-885793.htmlhttps://timesofindia.indiatimes.com/city/delhi/suicide-cousin-held-for-morphing-photos/articleshow/81767713.cmsYoung women in India also share that bullying using morphed media or deep fakes can crush their self-esteem: https://timesofindia.indiatimes.com/city/bengaluru/deepfakes-no-laughing-matter-can-crush-self-esteem-say-young-women/articleshow/105054087.cmsFurthermore, morphed photos are often used for blackmail by scammers, resulting in both financial and emotional harm to the victim and sometimes even their immediate family.https://organiser.org/2023/07/13/183574/bharat/bhopal-family-of-4-committed-suicide-after-cyber-hacking-and-theft-asked-for-a-collective-cremation/#google_vignetteIts critical that such media is removed as quickly as possible, and that there are ample Trust & Safety measures in place to prevent its distribution. The accounts that post such content should be considered for being disabled with fewer strikes than the standard requirement (i.e., treat it the same as accounts involved in extreme severe abuse types like Revenge Porn, Sextortion - because the intent to humiliate, harass, or threaten a person by sharing non-consensual intimate imagery is a common behavior among these abuse types and deep fake pornography).Lastly, standard treatments regarding AI generated images (e.g labels) should apply to such deep fakes too until they are identified as policy-violating and get removed from the platform. https://about.fb.com/news/2024/02/labeling-ai-generated-images-on-facebook-instagram-and-threads/

Link to Attachment

No Attachment

2024-007-IG-UA, 2024-008-FB-UA	PC-27069	Central & South Asia
Case number	Public comment number	Region
Withheld	Withheld	English
Commenter's first name	Commenter's last name	Commenter's preferred language
Withheld		No
Organization		Response on behalf of organization

Full Comment

The Information Technology (Intermediary Guidelines and Digital Media Ethics Code) Rules, 2021[4] under Rule 3(2) speaks to the grievance redressal mechanism for intermediaries. The proviso to Rule 3(2)(a)(i) states that complaints regarding obscene, pornographic, paedophilic, invasive of anothers privacy including bodily privacy, insulting or harassing based on gender material should be acted upon as expeditiously as possible and should be resolved within seventy-two hours of such reporting. Further, Rule 3(2)(b) states that the intermediary shall, within twenty-four hours from the receipt of a complaint made by an individual or any person on his behalf under this sub-rule, in relation to any content which is prima facie in the nature of any material which exposes the private area of such individual, shows such individual in full or partial nudity or shows or depicts such individual in any sexual act or conduct, or is in the nature of impersonation in an electronic form, including artificially morphed images of such individual, take all reasonable and practicable measures to remove or disable access to such content which is hosted, stored, published or transmitted by it. Such automatic closure of the reports is in contravention of the IT Rules, and Meta policy must be amended to be in compliance with Rule 3 of the IT Rules. The amended policy should ensure Rule 3 also demonstrates the need for pornographic material (real or deepfake) to be regulated under a common policy, which reviewers may take up on priority given the gravity of the same, and the compliances given under Indian Law.

Link to Attachment

No Attachment

2024-007-IG-UA, 2024-008-FB-UA	PC-27072	Central & South Asia
Case number	Public comment number	Region
Aleena	Afzaal	English
Commenter's first name	Commenter's last name	Commenter's preferred languagebr
Digital Rights Foundation		Yes
Organization		Response on behalf of organization

Full Comment

DIGITAL RIGHTS FOUNDATION PUBLIC COMMENT ON OVERSIGHT BOARD CASES 2024-007-IG-UA, 2024-008-FB-UA (EXPLICIT AI IMAGES OF FEMALE PUBLIC FIGURES)Submission: Research Department - Digital Rights FoundationAleena Afzaal - Sr. Research Associate Abdullah B. Tariq - Research AssociateSubmission Date: April 30, 2024 Legal Context:Given the borderless nature of digital content, Meta should consider international legal developments as a framework for its policies. The European Unions Digital Services Act and specific statutes from the U.S. state of California, such as AB 602, provide precedents for regulating digital content and protecting individuals against non-consensual use of their images. Irregular responses in two different cases (How such cases affect people in different regions):It is important to note that the two cases relating to deepfake videos of women public figures were approached and dealt with differently potentially due to difference in ethnicity and identity: one being from the Global North and the other belonging to the global majority identity. The American public figure case received a relatively immediate response whereas the case of resemblance to a public figure in India was not highlighted or amplified as quickly. Despite the technical discrepancies, it cannot be ignored that in the latter case, an Instagram account with several similar images remained unflagged for a long time. Additionally, one question that arises continuously from a string of these cases is why have tech platforms not adopted technological mechanisms that can flag sensitive content, particularly deepfakes circulating on different platforms. The harms prevailing due to the emerging technologies particularly generative AI content need to be viewed through a more intersectional lens. Women and marginalized groups in the global majority particularly from South Asia are more vulnerable to attacks online with a significant impact on their online and offline safety rather than individuals from the global North. While female security and inclusion is crucial, the potential otherization of the community is concerning and needs to be revisited. Moreover, taking cultural context into account, the level of scrutiny and criticism a South Asian female is subjected to in such events is higher as compared to a woman of American descent. In India, a woman is viewed as good only if she is able to maintain the respect and honor of her family. Female bodies are sexualized and any attack on them is considered to be an attack on men and the community's honor. Several cases have come forward in the past where women and young girls in India have taken their own lives as a result of leaked photos. In the wider Indian subcontinent region, cases have arisen where women have been subjected to honor killing as a consequence of being romantically involved with a man, their explicit photos being leaked and more. Such cases in the region showcase an underlying problem where women and honor are used as interchangeable terms and need to be taken into consideration when handling issues of similar nature. Public figures or not, women are more prone to being targeted by AI-generated content and deepfakes. Recently, incidents have come forward where deepfakes of two female public figures in Pakistan have been made widely available across different social media platforms. As far as Metas platforms are concerned, these deepfakes have been uploaded with nudity being covered with the use of stickers and emojis however in the comments section, users have offered and/or asked to share the link to view the originally created content. It is crucial that platforms like Meta have mechanisms in place where content and comments amplifying technology-facilitated gender-based violence are also flagged. Considering the higher probability combined with the societal consequences, it is essential for Meta to give greater consideration to cases involving deepfakes and AI-generated content showcasing characteristics of technology-facilitated gender-based violence more importance on the platform, particularly with countries from the global majority where the risk of potential harm is higher than others. Human reviewers should also be made aware of the language and cultural context of the cases under consideration. Trusted partners of Meta should be entrusted with the task of escalating the cases, where the response time of prioritized cases is expedited and addressed at the earliest.   Clarification and Expansion of Community Guidelines:Metas current community standards need to be more explicit in defining violations involving AI-generated content. There is an urgent need to develop a specific section for public-facing community guidelines on the platform to address deepfakes. Detailing examples and outlining repercussions would clarify the company's stance for users and content moderators alike. Public figures are at a higher risk of being victims of deep fake content due to their vast exposure (reference imagery) in online spaces. Thus, the policy rationale and the consequent actions need to be similar in the case of public figures and private individuals considering the sensitivity of such content regardless of an individuals public exposure. It is equally important that Meta revises its policy regarding sensitive content where the person being imitated is not tagged. The policy needs to be inclusive of such content as the potential harms remain. Regular updates to these guidelines are crucial as AI technology evolves.Technical Mechanisms for Enhanced Detection and Response: Implementing cutting-edge machine learning techniques to detect deepfake content (image, video and audio) can significantly reduce the spread of harmful content. These algorithms should focus on detecting common deepfake anomalies and be regularly updated to keep pace with technological advancements. A two pronged approach can be utilized for detecting and flagging harmful content on their platforms. Larger investments should be placed in automated detection systems to efficiently categorize and identify generative AI content and be adaptable to future advancements. Detected Gen AI content should be marked on Meta platforms to avoid confusion or the spread of misinformation. Meta needs to reassess its appeals pipeline and allow for extended review times, especially for content that contains any human likeness. Moreover, Meta needs to reassess its appeals pipeline and allow for extended review times, especially for content that contains any human likeness.Collaborating with AI developers to embed watermarks in AI-generated content can help automatically identify and segregate unauthorized content. This would bolster Meta's ability to preemptively block the dissemination of harmful material. Expanding this database to include international cases and allowing for real-time updates can enhance its effectiveness in identifying and removing known violating content swiftly.Meta should build on and enhance the capacity of its trusted partners particularly in terms of escalating content to the platform and having a robust and quick escalation channel in case of emergencies or content that is life-threatening. Meta needs to have emergency response mechanisms in place and have policy teams who are sensitized to deal with matters of utmost urgency particularly when it relates to marginalized groups and vulnerable communities.The current challenges faced by

Meta in managing AI-generated content are largely due to its lack of specificity in its policies to encapsulate generative AI content. The community standards in their current state fail to address the complexities of AI-generated content and the adverse impacts it can have on people and communities. Metas clear differentiation in its policy application rationale for two different cases raises concerns over irregular and inefficient content moderation policies. While we acknowledge that content in both these cases is no longer on the platform, the urgency displayed in taking down content from the second case compared to the delay in the removal from the first case highlights the dire need for stringent and equitable response of social media platforms on gen-AI content. Moreover, in the second case the deepfake video of an American woman public figure was removed under the policy Bullying and Harassment, specifically for "derogatory sexualised photoshop or drawings" Greater discourse is required over what classifies as derogatory in this context. In the absence of a derogatory element, will an AI-generated image that involves sexualisation and nudity be available to view on the platform? If so, then how is Meta perceiving the consensual privacy and dignity of public figures on its platforms? These are the questions that need to be addressed and outlined in Metas content moderation policies, especially in terms of tech-facilitated gender-based violence.Metas Media Matching Service Banks are restricted by the database of known images, which renders them highly ineffective against newly generated deepfake content. With tools to create generative AI content becoming increasingly accessible, the technology to flag and address such content needs to catch up as soon as possible. It is essential for Meta to expand its database to encompass a wider array of AI-generated content types and implement real-time updates. In conclusion, Metas automated detection systems struggle to keep pace with rapidly advancing sophisticated technologies used in deepfake content. For Meta to ensure safety on its platforms for marginalized groups and communities, it is essential for them to revisit their content moderation policies pertaining to generative AI content while enhancing and investing in its trusted civil society partners to escalate content towards the platform.

Link to Attachment

PC-27072

2024-007-IG-UA, 2024-008-FB-UA

Case number

PC-27073

Public comment number

Central & South Asia

Region

Withheld

Commenter's first name

Withheld

Commenter's last name

English

Commenter's preferred language

Withheld

Organization

No

Response on behalf of organization

Full Comment

1.The nature and gravity of harms posed by deepfake pornography including how those harms affect women, especially women who are public figures.According to the Institute of Development Studies, between 16-58 per cent of women have experienced technology-facilitated gender-based violence . The Economist Intelligence Unit found that 38 per cent of women have had personal experiences of online violence, and 85 per cent of women who spend time online have witnessed digital violence against other women . The EIU study further revealed that unwanted images or sexually explicit content, which includes digitally created content, is experienced by 43% of the women in the study.Deep fakes exacerbate the objectification of women and perpetuate the belief that men are entitled to control womens bodies. Additionally, when deep fakes manipulate public figures into making statements they did not intend, it causes both reputational and illocutionary harm. This can lead to individuals being misrepresented and coerced into accepting false actions, undermining their self-respect. Ultimately, deep fakes undermine the autonomy and self-worth of their survivors, sometimes unknowingly leaving them anothers mercy. Regina Rini and Leah Cohen alarm us about the detrimental impact of deep fakes in their paper Deep Fakes, Deep Harms.
2.Contextual information about the use and prevalence of deepfake pornography

globally, including in the United States and India. Women are overwhelmingly affected by deepfake technology. Few studies have been conducted in this area but the ones that exist share a similar outcome. A study conducted by Deeptrace Labs, a company that makes tools to identify deepfakes, revealed in 2019 that the vast majority of the subjects of these fake videos across the internet a full 96 per cent are of women, mostly celebrities, whose images are being used without their consent .In that study, Deeptrace analyzed the gender, nationality, and profession of subjects in deepfake videos from the top 5 deepfake pornography websites, as well as the top 14 deepfake YouTube channels that host non-pornographic deepfake videos. The study found that deepfakes had increased 100% from the previous years and the subjects are overwhelmingly women. While Western women made up most of the subjects, there was a global increase in the usage of such technology. Deepfake technology has been used to victimize children as well. Multiple cases have led to at least a dozen states in the US working on bills, or pass laws, to combat A.I.-generated sexually explicit images of minors . In the UK, a law has been passed to criminalize the creation and sharing of deepfake content created without the consent of the subject . Internationally, we are moving towards the criminalization of the usage of deepfake technology without consent. Deepfake technology should be treated akin to revenge pornography, as it allows the targeting of women to affect reputation and dignity in a similar manner. The measures taken by META to combat revenge pornography in the U.S. such as the creation of helplines and bodies like National Center of Missing and Exploited Children where all sexually explicit images detected are reported, should be replicated in the Indian context to facilitate reporting of digitally created content in the appropriate territorial jurisdictions. A study conducted by iScience observed how confident people felt with detecting deepfakes in contrast to their actual ability to be able to detect deepfakes. In the study, it was seen that the confidence that the sample group, which was 210 participants, had to detect deepfakes was much higher on average than their actual ability of being able to detect deepfakes. The study stated most participants were overconfident with their ability to detect deepfakes but when the study was conducted, the accuracy of being able to detect was much lower. The study concludes that the inaccuracy of not being able to detect these deepfakes isnt a reflection of the inability of participants but rather is a sign of the fast-paced growing technology of deepfakes. This study classifies a major issue with the growing prevalence of deepfakes which is that consumers of media aren't in a position to be able to judge whether the content they are consuming is a deepfake or not. The prevalence of deepfakes grows at a sparring rate, especially their use being a form of attack towards women mostly. In a study done by Deeptrace Labs, it was stated that 96% of the deepfakes present are of women and have been sexualised with no consent from the part of the women. The rise of deepfakes is at an all-time high, and since the use of the content mostly impacts women and is also heavily politically charged, the consequences of such content are very grave and need to be regulated. 3.Strategies for how Meta can address deepfake pornography on its platforms, including the policies and enforcement processes that may be most effective. In the Child sexual exploitation, abuse and nudity policy by Meta, on becoming aware of content on child exploitation, the same is reported to the National Center for Missing and Exploited Children. We recommend that Meta, as an intermediary must report child pornography on its platform to Indian law enforcement agencies to foster a safer community experience for those between the ages of 13 and 18.Under Sections 14 and 15 of the POCSO Act, 2012 , child pornography is criminalised in India. Further, Sections 19 read with 21 of the Protection of Children from Sexual Offences Act, 2012 all persons are responsible for reporting apprehension and/or knowledge of an offence under the Act to the SJPU or the local police, failing which, the person can be held criminally liable.[1] We recommend that the strict compliance of the same must extend to Meta for any reported content on their platform upon review.This is particularly relevant as the National Human Right Commission sent a notice to the Indian Union Government, various State Governments and Union Territories flagging an increase in Child Sexual Abuse Material (CSAM) on social media platforms by 250-300%, emphasising on the psychosocial harms of such material and the urgent need for a stakeholder dialogue on the same . As Meta is an important stakeholder in ensuring that such content is regulated, we recommend a mechanism to facilitate such compliance as stated within the POCSO. The complaints process followed by Meta should be published widely and make it more visible to ensure access. A zero-trust mindset, should be adopted for deepfake content wherein the initial position of META towards such content is one of distrust . The practice within METAs Adult nudity and sexual activity policy wherein the default position is to take down such content should be implemented by utilizing auto-detection technology. Such content should be allowed to be shared only upon review of the relevant context. A label identifying deepfake content should be mandated for the ease of users to recognize such content as AI generated. The traceability would disallow the proliferation of such content without the necessary context. A chatbot based in India run by META in collaboration with Deepfake Analysis Unit an initiative by Misinformation Combat Alliance has been introduced to tackle the issue of deepfake content by verifying its authenticity. Users can send audio or videos to the chatbot to verify, whether the given media is deepfake/AI-generated. This chatbot or feature to verify a medias authenticity only exists in India, exclusively on WhatsApp. Moreover, the chances of someone checking the authenticity by going out of the way are very low. Furthermore, it does not lead to any repercussion on the content itself, it just tells one person whether a video/audio theyve reported is authentic or AI-generated. 4.Metas enforcement of its derogatory sexualized photoshop or drawings rule in the Bullying and Harassment policy, including the use of Media Matching Service Banks.The inclusion of derogatory sexualised photoshop or drawings in a category separate from adult non-consensual sexual activity to counter deepfakes is a mis-categorisation as the creation of these images and videos using a public or private figure is inherently a non-consensual use of their likeness being depicted in a sexual manner.While deepfakes of children are covered by the Child Sexual Exploitation, Abuse and Nudity policy along with other pornographic content involving children. However, adult deepfake content in the Indian context falls within the Bullying and Harassment policy instead of the Adult Sexual Exploitation policy which affects the way content violations are processed by the Meta community guidelines. Further, the categorisation of deepfakes in the Bullying and Harassment policy, which seeks to cover a broad policy rationale of ensuring curbing threats and unwanted malicious contact for users of the app does not adequately account for the distinctions between other categories enforced under this policy (such as denial of violent tragedies) and sexually explicit content such as deepfakes. The usage of deepfakes, especially in the context of private individuals where the main motivator for such content is revenge porn, impacts the targeted individual more directly and requires faster review and resolution to be taken down by Meta.Specific to the evolution of technological law, it has been noted that pre-existing categorisation may not adequately address evolving technical challenges, thereby asking policymakers to consider the basis and objective for the distinctive categorisation when created, and considering which applies best to the problem sought to be addressed . As noted specifically in the context of regulating deepfakes, the incorporation of these into pre-existing policy without re-evaluating it from a critical perspective on the object of the policy could distort the actual legal issue we seek to address here .The usage of deepfakes is more closely related in its objective to other forms of Online Gender Based Violence , which should be regulated and enforced under the policy that regulates similar non-consensual sexual activity to adequately enforce the objective sought.META is testing a new feature as a policy for Nudity Protection in DMs. It aims to protect children from unsolicited sexual images and discourage them from even sending their own naked pictures by giving them a confirmation message before the picture goes through by detecting nudity in the pictures. Since the AI Detection is being done for nudity by Meta currently, it should be extended to examine deepfakes as well. Deepfake content should also be de-indexed to remove search results and online references when such content has been removed either by default or through reporting. Such deindexing is in line with the right to be forgotten, an internationally recognized standard of data protection .5.The challenges of relying on automated systems that automatically close appeals in 48 hours if no review has taken place.The Information Technology (Intermediary Guidelines and Digital Media Ethics Code) Rules, 2021 under Rule 3(2) speaks to the grievance redressal mechanism for intermediaries. The proviso to Rule 3(2)(a)(i) states that complaints regarding obscene, pornographic, paedophilic, invasive of anothers privacy including bodily privacy, insulting or harassing based on gender material should be acted upon as expeditiously as possible and should be resolved within seventy-two hours of such reporting. Further, Rule 3(2)(b) states that the intermediary shall, within twenty-four hours from the receipt of a complaint made by an individual or any person on his behalf under this sub-rule, in relation to any content which is prima facie in the nature of any material which exposes the private area of such individual, shows such individual in full or partial nudity or shows or depicts such individual in any sexual act or conduct, or is in the nature of impersonation in an electronic form, including artificially morphed images of such individual, take all reasonable and practicable measures to remove or disable access to such content which is hosted, stored, published or transmitted by it. Such automatic closure of the reports is in contravention of the IT Rules, and Meta policy must be amended to be in compliance with Rule 3 of the IT Rules. The amended policy should ensure Rule 3 also demonstrates the need for pornographic material (real or deepfake) to be regulated under a common policy, which reviewers may take up on priority given the gravity of the same, and the compliances given under Indian Law.

Link to Attachment

PC-27073

2024-007-IG-UA, 2024-008-FB-UA	PC-27074	Central & South Asia
Case number	Public comment number	Region
Withheld	Withheld	English
Commenter's first name	Commenter's last name	Commenter's preferred language
Withheld		No
Organization		Response on behalf of organization

Full Comment

Clarity is required in the terms of use and community policies. Does "default removal" mean automatic removal or upon reporting? (Adult Sexual Content Policy)Deepfakes should be automatically removed. It disproportionately affects woman more. It is a danger to the reputation and dignity of women and exploits public figures. Intermediaries have a role because you are directly profiting off of such viral content and thus contributing to the exploitation unless safeguards are there. The review system is a problem of META and not the user. The effect should not be upon the user for a problem with the review system. The determination of age by the AI to determine if children are depicted in such deepfake imagery is not clear. The intermediary should preemptively remove such data to avoid proliferation to other sites.

Link to Attachment

No Attachment

2024-007-IG-UA, 2024-008-FB-UA	PC-27075	Europe
Case number	Public comment number	Region
Sarah	Andrew	English
Commenter's first name	Commenter's last name	Commenter's preferred language
Avaaz Foundation		Yes
Organization		Response on behalf of organization

Full Comment

"Mom, they say theres a naked photo of me going aroundthat they did it with an artificial intelligence app. Im scared. Some girls have also received it. Quote from a child victim of fake sexual images distributed on WhatsApp and TelegramThese remarks are submitted by Avaaz, a global civic organization representing around 70 million people including two million from Spain. We are dedicated to combating the harmful impacts that the misuse of artificial intelligence can inflict on our communities. Our advocacy efforts have contributed to the inclusion of human rights protections in the European AI Act, during which we have investigated various harmful uses of AI, including the creation and spread of AI-generated pornography through social media platforms. The Board asked for input on the nature and gravity of harms posed by deepfake pornography and contextual information about its use globally,We wish to highlight a distressing case from Almendralejo, Spain, to the Oversight Board. In September 2023, reports surfaced that teenage girls were distressed, scared and suffering from anxiety after discovering that manipulated images of themeither topless or in even more compromising stateswere being shared by their male classmates via WhatsApp and Telegram. Over 20 girls, aged between 11 and 17, stepped forward as victims. These images were crafted using photographs of the girls, fully clothed and often taken from their personal social media accounts, then altered by an app that simulated them without clothes.It is important to note that the suspects involved were also minors, eleven boys aged between 12 and 14, now within the permissible user age for WhatsApp, which has recently been lowered to 13. While the Oversight Board is concerned with the impact on public figures, it is crucial to address that Metas platforms are being exploited to harm vulnerable young individuals as well with no media profile or support. As the opening quote shows, the distress is real. The devastating impact on these young girls' lives is profoundly exacerbated because the manipulated images are circulating, even at smaller scales, amongst people they know, people they see daily at school, leading to severe emotional distress and isolation.Currently, the creation of such AI-generated images is not criminalized in Spain, and pending regulations under the EU AI Act will not take effect until 2026. In this period, when damaging content can be produced and distributed so readily, platforms like WhatsApp bear a significant responsibility to limit such content to the fullest extent possible, balancing the rights to privacy and free speech. However, when the content involves sexualized images of children, the imperative to act decisively and responsibly is even more critical.This is a call for the Board to show ethical leadership in protecting our youth and to step up and widen the scope of its investigation. We urge the Oversight Board to consider these issues seriously and, whilst respecting free speech, lay down guidance for WhatsApp to transform how it approaches the encrypted dissemination of child fake pornographic images at any scale, whatever the public profile of the victim. Avaaz FoundationApril 2024

Link to Attachment

PC-27075

2024-007-IG-UA, 2024-008-FB-UA	PC-27076	Central & South Asia
Case number	Public comment number	Region
Withheld	Withheld	English
Commenter's first name	Commenter's last name	Commenter's preferred language
Withheld		No
Organization		Response on behalf of organization

Full Comment

The harm caused by deepfake morphed images has particularly affected women and young girls. Given the Indian context of a deeply patriarchal society that actively dissuades and prevents women from participating fully in society, including in the digital space, morphed images are used as weapons to terrorise and subdue women. There is a specific pattern of misogyny, victim blaming and shaming and bullying, which results in severe mental trauma, as well as other additional repercussions, including loss of access to phones and the internet, loss of mobility (especially for young girls), and further effects such as bringing shame to the family/community Ive worked closely with women across India, both rural and urban, training and advocating for increased access to technology and internet, including using social media platforms to share news and information. I have witnessed firsthand the challenges women and girls face in accessing phones and the internet. Further, I have seen the various ways they have been bullied to silence their voices. Specifically, I have supported female activists being trolled with morphed images, which have damaged their credibility and reputations, and caused significant harm to their mental health. I am myself a woman, and use multiple social media platforms, and consider it to be especially worrying when in one of the specific cases described, the first complaint was automatically closed because it was not reviewed within 48 hours. Meta must improve and enhance faster actioning on reports, and include a zero tolerance approach to digitally manipulated nude and pornographic content. Meta has the capacity and needs to commit to challenging and preventing misuse and abuse of their platforms. Policies and enforcement processes MUST be stringent, with action to identify and prevent digitally manipulated nude and pornographic content, including reuploads.

Link to Attachment

No Attachment

2024-007-IG-UA, 2024-008-FB-UA	PC-27077	Europe
Case number	Public comment number	Region
Withheld	Withheld	English
Commenter's first name	Commenter's last name	Commenter's preferred language
Withheld		No
Organization		Response on behalf of organization

Full Comment

Explicit AI Images of Female Public FiguresNon-consensual deepfake as a form of gender-based violenceTo properly address the problem of non-consensual deepfake, it is important to understand it as a form of gender-based violence or, preferably in this context, of cyberviolence against women. According to a September 2019 report, synthetic media and deepfakes are on rise. Since late 2017, the phenomenon has developed rapidly, in terms of both technological sophistication and societal consequences. The findings reveal the presence of 14.678 deepfake videos in the cyberspace, as for September 2019, meaning that an increase of 100% has taken place since the previous measurements taken in December 2018, reaching a total of 7.9641 . While the political and intellectual community has concerns for politics and cyber security, yet, as the report confirms, the proliferation of this piece of technology has mostly affected women. Hence, very large online platforms such as Facebook and Instagram should put in place policies and practices that hear and voice their perspectives and experience. To do so, very large online platforms should bear in mind the impact of the creation and the distribution of intimate images without consent, which has multiple, severe, continuous and heinous consequences on offended subjects, both in the private and in the public sphere. As reported by a Cyber

Civil Rights Initiative survey with total 1606 respondents, 361 offended parties, 82% suffered significant impairment in social, occupational, or other important areas of functioning and 54% had difficulty focusing on work or at school. Additionally, 55% and 57% fear for their current professional reputation and their professional advancement, 13% had difficulty getting a job or getting into school, 8% quit their job or school and 6% were fired or kicked out of school. Moreover, one out of three declared that this jeopardized their relationships with family or friendships, while 13% lost a significant other and 40% fear the loss of a current or future partner. Further victimisation is another potential consequence of this kind of gender-based abuse. According to the same study, 49% experienced online harassment or stalking by users who have seen the material, while 30% experienced harassment or stalking in person or over the phone. Findings of a study based on in-depth quality interviews and inductive analysis conducted between February 2014 and January 2015 with 15 female self-identifying as offended subjects show the devastating impact on emotional and mental health, as well as similarities with sexual assault. Nearly all respondents experienced a general loss of trust in others, diagnosis of PTSD, anxiety and depression disruptive of everyday life and sleep patterns, lower self-esteem and confidence, a sense of loss of control on past, present and future. Besides, negative coping mechanisms were present too, including avoidance, denial, excessive drinking of alcohol and obsessing over ones victimization, while common positive coping mechanisms were seeing a counsellor or therapist, speaking out and helping others, relying on support system such as family or friendships, and focusing on moving on.Non-consensual deepfake within the socio-cultural indian context Very large online platforms such as Facebook and Instagram should also place the phenomenon of non-consensual deepfake within the particular geographical and cultural context where it takes place. The cases concerned took place in India. Violence against women, and especially rape culture, constitutes a sore spot for the country. The first ever documented case of non-consensual dissemination of intimate image took place in India in 2001. when a 16 years old Delhi schoolboy created a website and posted intimate images of schoolgirls and teachers without their consent, as well as details of their sexual preferences . Being the first ever report of teen revenge through cyberspace, the case captured an overall attention of the media, the public, the legal scholars and the police, leading to a new trend later imitated by many adults to perform gender-based violence. Since social and cultural changes occurred in India, under the influence of western trends as well as the entry in the web 2.0 era, the consensual capturing of sexual performances with the partner by the partner himself or by automatic devices is commonplace between the younger generations.Besides, categories such as revenge porn, Porn India, India teen porn, Indian Desi Girl and

South Indian Mallu etc. are popular in pornographic websites, even on YouTube platform45. According to legal scholars, the consensual capturing of sexual performances with the partner is a growing tendency among the Indian teenagers, leading to an expectable rise of the non-consensual phenomenon, also regarding adults, as the key words searching confirm. Deepfake as a weapong used against female public figures Not only non-consensual deepfake has serious and harmful individual effects on the targeted offended subjects, which may range from damages to reputation to psychological and mental health issues, from the reduction of educational and professional opportunities to social isolation, stigmatisation, discrimination and even physical danger or suicide, it has collective consequences, too. All women are potential targets of deepfake, as highlighted by numerous scholars and activists in the field. This is due to two main reasons. Firstly, the large amount of personal multimedia data available online, most of the time willingly uploaded, i.e. photographs, videos, voice registrations. Secondly, the availability of scraping tools combined with the incorporation of deepfake technology into popular and commercialised desktop and smartphone applications. The solely awareness of such constitutes undoubtedly a threat to womens engagement in society.The case of Rana Ayyub is exemplary in this regard . Ms Ayyub is an independent journalist and writer whose work has included investigations into alleged crimes committed by public and government officials. According to information received by the experts, the issue intensified after a malicious Tweet on 20 April falsely quoted Ms Ayyub as supporting child rapists and saying that Muslims were no longer safe in India.After the Tweet was published, Ms Ayyub received a barrage of hate-filled messages, which included calls for her to be gang-raped and murdered, and made numerous references to her Muslim faith. Her phone number and home address were posted on a social network, and the threats against her are continuing even though she has clarified that the Tweet was false. The experts expressed further alarm that a faked pornographic video purporting to show Ms Ayyub was also recently circulated online, triggering new threats.While police began an investigation into the threats 10 days after Ms Ayyub filed a complaint, she has reportedly not yet received any police protection.Very large online platforms such as Facebook and Instagram should therefore put in place swift procedures to ensure the protection of female public figures targeted by deepfake campaigns. Practical guidance for Facebook and Instagram to counter non-consensual deepfake Very large online platforms such as Facebook and Instagram shall: engage in tackling misogynistic conduct, speech and content in their terms of service, including any conduct, speech and/or content of users based on hatred of, contempt for or prejudice against women and girlsengage in tackling non-consensual intimate imagery distribution in any form in their terms of service, including photographs, videos, streaming videos, manipulated media, deepfake photographs and videos portraying sexually explicit or, more broadly, intimate images of a person that were not originally intended for public distribution or disseminated without the depicted persons consent.take into account the social context of the users during the content policy stipulation process, in particular by embedding social, cultural, geographical and gender-based considerations in their polices and products tackling non-consensual intimate imagery distribution.content policy drafting, internal content moderator staff training and supportive programs development shall draw on gender expertise, including professionals, organizations and grass-root groups dedicated to prevent and contrast gender-based violenceengage in providing for effective human assistance, monitoring and control to affected users in removal procedures, instead of primarily relying on automated processes, particularly when the context of the platform requires so.provide affected users for system-wide removal of signalled content at source, by means of AI-based detective mechanisms and, eventually, fingerprinting systems, to proactively detect and remove disseminated non-consensual intimate content.provide for open, clear and detailed information on their internal governance and staff structure, in particular by indicating the number, localisation, specialisation and instruction of content moderators who respond to reported non-consensual intimate content.

Link to Attachment

PC-27077

2024-007-IG-UA, 2024-008-FB-UA	PC-27078	Central & South Asia
Case number	Public comment number	Region
Soumya	Nair	English
Commenter's first name	Commenter's last name	Commenter's preferred language
DID NOT PROVIDE		No
Organization		Response on behalf of organization

Full Comment

DID NOT PROVIDE

Link to Attachment

PC-27078

2024-007-IG-UA, 2024-008-FB-UA

Case number

PC-27079

Public comment number

Asia Pacific & Oceania

Region

Anushrut

Commenter's first name

Sharma

Commenter's last name

English

Commenter's preferred language

DID NOT PROVIDE

Organization

No

Response on behalf of organization

Full Comment

SUGGESTIONS TO THE OVERSIGHT BOARD1.It is suggested that META puts the labelling policy, that it intends to bring forth, into use on an urgent basis. 2.It drafts a separate clause for content created via AI, sexually suggested or otherwise and reviews it before the content is ever live at a platform. 3.Asks the user to self-verify or testify that the content that they are uploading. It may not lead to 100% verification but would at least warn and discourage people from putting AI-generated sexually explicit content. 4.Proper AI-based detection software to detect AI-generated content especially on Public Accounts with immense reach. The threshold can be based on the number of followers of the account. 5.Helplines for all platforms to report AI-generated content.

Link to Attachment

No Attachment

2024-007-IG-UA, 2024-008-FB-UA	PC-27080	Asia Pacific & Oceania
Case number	Public comment number	Region
Harshita	Hari	English
Commenter's first name	Commenter's last name	Commenter's preferred language
DID NOT PROVIDE		No
Organization		Response on behalf of organization

Full Comment

Deepfakes have been heavily influential in the current social media discourse. Meta as an entity needs to ensure that there is a pre-detect mechanism which can help prevent the use of non-consensual use of deepfakes. As deepfakes are mostly used against female influencers, it is important to have a mechanism where deepfakes are identified before uploading or go through review mechanisms right after uploading to understand the nature of the media itself.

Link to Attachment

[PC-27080]

2024-007-IG-UA, 2024-008-FB-UA	PC-27081	Asia Pacific & Oceania
Case number	Public comment number	Region
Reshma	Naykodi	English
Commenter's first name	Commenter's last name	Commenter's preferred language
DID NOT PROVIDE		No
Organization		Response on behalf of organization

Full Comment

Deepfake pornography poses a grave threat, especially for women and public figures who are easily targeted. These manipulated videos inflict irreparable damage by spreading false and degrading content without consent, resulting in reputational harm, cyberbullying, and emotional trauma. Women, particularly public figures, face malicious intent, exposing them to heightened risks of harassment and defamation.Public figures are currently the prime targets, but soon, ordinary women will be vulnerable targets in the name of love, rage, and more, leaving them helpless. Urgent action is crucial to combat this insidious abuse and safeguard the dignity and rights of every individual.

Link to Attachment

No Attachment

2024-007-IG-UA, 2024-008-FB-UA

Case number

PC-27083

Public comment number

Central & South Asia

Region

Withheld

Commenter's first name

Withheld

Commenter's last name

English

Commenter's preferred language

Withheld

Organization

No

Response on behalf of organization

Full Comment

Public opinion shared by Dr.Debarati Halder, LL.B, LL.M (International & Constitutional Law), LL.M (Criminal & Security Law)., Ph.D (Law) (NLSIU)The nature and gravity of harms posed by deepfake pornography including how those harms affect women, especially women who are public figures.Online victimization by creating porn contents on the basis of real images and deepfake porn content based victimization on the basis of real images may seem synonymous, but they have different legal meanings and the levels of harm may differ. A porn content may be created with upskirt photography that may be defined as voyeur image and this is recognized by criminal laws in many jurisdictions. But Deepfake porn is a different issue which may include doctoring of image without authorization and this may also include unauthorized access to device, content modification and unauthorized use of advanced digital technologies for an intentional causing of harm to the victims who may not be aware how she is being victimized. A deep look into this issue from cyber victimological understanding may suggest that the gravity of harm may be wider than any other form of victimization: this may not only traumatize the victim, but this may also cause privacy infringement, monetary loss and reputation damage for the victim. It is a common myth that women who are public figures may be not suffer victimization by deep fake porns like other private individuals who may have been victimized by fake avatars including by deep fakes. Women who are public figures may see their images being distributed on different platforms for different purposes which may include expression of political opinions and ideologies, face of the company, achievers and brand ambassadors . deep fake porns targeting such women may have deep impact on their privacy as well as their familys privacy. It may also cause threat to their reputation as well as to their physical integrity. While some women who are public figures, may afford to defend their rights in the courts for removing such deepfakes , for punishment of the perpetrators and for compensation for reputation damage, many may not afford the same even if they may have acquired a stand in the society by their hard work especially in countries like India. They may face trolls, public humility and even job loss due to the wide apprehension that their deepfakes may affect the reputation of the organizations they are working for. Even though we may see that due large scale awareness building by the government and NGOs, large part of the society in India as well as in many parts of the world, may have become aware about various methods of online victimization of women (including deepfakes), many may not accept that the victim was completely innocent. In my role as a cyber-rime victim counseller and an expert researcher in the field of cyber law and cyber victimology, I have experienced how deeply deepfake porns may affect the victim and the society at large. Women are silenced and their rights to speech and expression through information and digital technology are restricted by families when they get to see that empowered women are also falling victims of such patterns of victimizations. Some women (even if they are empowered) may opt out irrational coping mechanisms that may further push them to re-victimization. The gravity of harm may widen due to rejection of grievance complaints on deepfake victimization. This may further impact on intermediarys decisions on the offensive natures of the deep fakes. this in turn may encourage the perpetrators to be more vigorous for committing more harm to women at large. Since intermediaries like Meta may not want to abide by the laws in India, the entire judicial enquiry system may take longer duration for restitution of justice. Sometimes the entire system may fail the victims due to nonchalant attitude and this may motivate self harms like suicide, counter attacks by the victims which may in turn accommodate the perpetrators to play the role of victims and dismiss the objects of restorative justice.Metas enforcement of its derogatory sexualized photoshop or drawings rule in the Bullying and Harassment policy, including the use of Media Matching Service Banks.Usage of AI has become a norm for everyone for creating harassing and offensive contents specially to victimize women and girls. In my opinion this phenomenon is increasing because the perpetrators an access the AI tools easily from the intermediaries. Meta through Facebook, Instagram and WhatsApp provide many AI tools for photoshop. Perpetrators can also access AI tools from different platforms and websites to download images of women and then change the images to represent them in sexually explicit manners. Here we need to see two liabilities: (i) liability of the perpetrators that may be addressed by existing laws including that of Information Technology Act, 2000(amended in 2008), Indian Penal Code (now amended as Bharatiya Nyay Sanhita, 2023), Indecent Representation of women Prohibition Act etc, and (ii) liabilities of the intermediaries for not developing strategies as per Indian socio-legal standard and providing AI tools for photoshopping without providing proper guidelines for fair usage of the same. Intermediaries may also be held for not controlling the photo downloading options. It is understood that intermediaries may provide mechanisms to the users/subscribers for controlling third party access to their contents and images. But this may be violated by proxy stalkers and followers. Further, for the public figures and celebrities, there may be no control mechanism and their images can be downloaded from any platform and social media platforms like Facebook and Instagram are the most preferred platforms for re-sharing the morphed /photoshopped images. While Meta needs to understand the above liabilities, Meta as a company also need to understand the consequences from cybervictimological perspectives. In my research on cyber victimology I have shown the impact of image based harassment on women irrespective of socio-economic class. The remedy lies not only in the hands of courts, but also in the hands of the companies like Meta. My suggestions for improving the strategies of Meta in such image based harassment including AI tool based image based harassment is as follows: Consider the meaning of nudity not from US perspectives but from Indian psychological-socio-legal perspectives. Some women may post their images in swimsuits intentionally knowing the consequences of the same. But they are a minority group who may afford to cope with trolls, manage to followup with police complaints for unethical personal image distribution and may afford to spend money for hiring good lawyers. Others cannot. For them, morphed, deepfakes and nude pictures may turn dangerous as they may be targeted by their own society and they may face victim blaming in the police stations and courts. Hence Meta must expand the scope of strategies for addressing bullying, harassment and nudity. Each of these heads must have different strategies. Meta needs to understand that the concept of bullying must not be mixed with harassment and nudity from Indian socio-legal perspectives. The challenges of relying on automated systems that automatically close appeals in 48 hours if no review has taken place.Metas endeavor to block the circulation of offensive image by using different AI tools is commendable. But the reports on offensive images must be attended within 24 hours and it should not be through automated systems.. The Information Technology (Intermediary Guidelines and Digital Media Ethics Code) Rules, 2021 very clearly mention intermediary liability for handling the reports of harassment within a stipulated time frame. Meta needs to follow these Rules more strictly. Once a complaint is rejected by Meta due to automated system of identifying the content as offensive and consequently closure of the case due to nil review, victims may become extremely frustrated, depressed, traumatized and withdrawn. Not every victim may know the result of the complaint within the stipulated time of 48 hours in India. This may escalate chances of re-victimization. In such cases the intermediary may become liable for possible self-harm of the frustrated victim because such images may continue to be shared, downloaded and archived by people who may not know the victim personally, but may create larger harm by circulating these images at different times for different purposes. Reference: Halder, D. (2021). Cyber Victimology: DecodingCyber Crime Victimization. Boca Raton, FL, USA: Routledge, Taylor andFrancis Group. ISBN: 9781498784894 Halder D., & Jaishankar, K (2016.) Cyber crimes against women in India.New Delhi: SAGE Publications. ISBN: 9789385985775Halder Debarati., (2016) Celebrities and Cyber Crimes: AnAnalysis of the Victimization of Female Film Stars on the Internet. Temida- The journal on victimization, human rights and gender. 19(3-4), 355-372.ISSN: 14506637 (UGC Listed Journal).Halder D. (2015). "Cyber Stalking Victimisation of Women: Evaluating theEffectiveness of Current Laws in India from Restorative Justice andTherapeutic, in Jurisprudential Perspectives," Temida - The journal onvictimization, human rights and gender, pp.103-130. ISSN: 1450-6637Halder D., & Jaishankar, K. (2014). Online Victimization of AndamanJarawa Tribal Women: An Analysis of the Human Safari YouTubeVideos (2012) and its Effects. British Journal of Criminology, 54(4), 673-688. (Impact factor 1.556). ISSN: 00070955

Link to Attachment

No Attachment

2024-007-IG-UA, 2024-008-FB-UA	PC-27085	Europe
Case number	Public comment number	Region
Lasara	Kariyawasam	English
Commenter's first name	Commenter's last name	Commenter's preferred language
University of Oxford		Yes
Organization		Response on behalf of organization

Full Comment

The impact of AI-generated sexual content on girls and women in South AsiaWhilst deepfakes of Indian celebrities has become a hot global topic, other South Asian countries are not far behind. Creating deepfakes of public figures and disseminating these seem to have had no consequences for culprits thus far in some South Asian countries. For instance, the top first and third search results for AI generated nudes in Sri Lanka are two links: Nudity Sri Lankan Sinhala Women AI Art Generator, and Desifakes.com respectively. The first site is a platform for users to create their own AI-generated pictures, and the second site contains multiple AI generated pornographic content of women who are public figures in Sri Lanka. Whilst these online platforms are publicly accessible with no age restrictions, no news-site or authority has brought this issue to public attention. In fact, it is unclear whether the women who are victimised by these websites are aware of the existence of these websites. A South Asian research study that analysed deepfake-related news across several media outlets in Bangladesh, India, and Pakistan found that more than 50% of news in Pakistani newspapers were related to the dangers of deepfakes (Sunvy et al., 2023). Earlier this year, a deepfake pornographic video of a Bangladesh female celebrity circulated on the internet and after a reviewing of the video, an online fact checking research platform confirmed that it was a fake (Hossaion, 2024).Whilst the impact of victimisation can be intense, this could be particularly extreme for people from certain cultural backgrounds. For instance, the conservative and patriarchal values of South Asian countries such as India and Pakistan make the gravity of deepfakes a grave concern for girls and women (Altaf & Javed, 2024). Moreover, shame is embedded in the South Asian culture and is valued as a virtue particularly for women (Abeyasekera et al., 2019). In fact, in countries such as Sri Lanka, where shame denotes purity, sexual indecency regardless of whether it is real or rumoured could harm the reputation of a woman and her familys honour, which could even jeopardise her ability to get married (Abeysekera & Marecek, 2019). This emphasis on honour as an indicator of a woman or a girls value is embedded in similar cultures such as the Indian and Pakistani culture (Altaf & Javed, 2024). Just being accused of engaging in sexual impropriety or having shamed oneself or their family has been linked to suicide or attempted suicide among young girls in Sri Lanka (Abeysekera et al., 2019; Abeysekera & Marecek, 2019). Several years ago, a 21-year-old Indian woman committed suicide after her pictures were morphed into pornographic content and shared and tagged on Facebook. Her suicide note revealed that this drastic measure had to be taken because she could not fight the stigma any longer and that her own parents and police did not believe that she was innocent (Madhav, 2016; Sudhir, 2016). On the other hand, being victimised of honour killings for having shamed ones family is an increased risk for victims such as Pakistani girls and women irrespective of whether the images are AI-altered.These cultural determinants and traditional boundaries of certain cultures particularly in South Asia signify the severity of the experience for victims who have been and might be victimised of AI-generated pornographic content. The lack of social support, victim-blaming culture, stigma on mental health and sexual behaviour would all add to the distress already caused by being victimised of this heinous crime. In addition, victims might not speak English although they might be using social media platforms such as Facebook and Instagram which means that they might not know how to report this content to the relevant social media platforms. According to Altaf and Javed (2024), being victimised or the fear of being victimised might not only impact on the mental wellbeing of South Asian girls and women, but it could also potentially curtail womens willingness to seek and participate in education and employment due to the heightened risk of reputational damage. Therefore, it is of utmost importance that policymakers and social media platforms take into account the extra cultural pressures for certain users and the damaging effects of AI-generated content. Any AI-generated pornographic content should be immediately blocked on social media platforms before several of its users end up taking their own lives to a crime they did not commit in the first place.Suggestions for META: When reporting an inappropriate picture, video, or post: people should be able to choose the reason. If it is nudity: the report should be addressed with immediate effect (as of now: Facebook allows people to choose nudity as a reason for asking to report a content. However, the response to this does not occur immediately, and some content does not get blocked or taken down. Instead, any content that is reported as nude should at least be temporarily taken down, until a real person or a committee decide whether or not the content is inappropriate). For any child users using Messenger (if at least one account belongs to a child): the term nude should be flagged with a caution message (for instance, if someone on the other end requests a nude from a child, the child should get a pop up message with the dangers of sending a nude/caution that this could be a case of sextortion, along with links to support resources). Any nude sent or viewed from a childs account to be blurred by default Information on helplines to pop up when a message or post is shared on suicide/self-harm (e.g., Facebook Messenger or Instagram inbox) or when the term is searched on social mediaDr. Lasara KariyawasamUniversity of Oxford

Link to Attachment

No Attachment

2024-007-IG-UA, 2024-008-FB-UA	PC-27086	Latin America & Caribbean
Case number	Public comment number	Region
André	Fernandes	English
Commenter's first name	Commenter's last name	Commenter's preferred language
Law and Technology Research Institute of Recife (IP.rec)		Yes
Organization		Response on behalf of organization

Full Comment

The general context of deepnudes as tools for gender-based violence on the InternetIn a simplified definition, deepfakes, whether pornographic or not, are falsifications resulting from image and video editing programs that possess (or make use of) an artificial intelligence algorithm at their core. Such algorithms are built based on deep learning and require significant computational power and a large amount of data to serve as a learning base, allowing for better performance (measured by a set of metrics). These models are also very hard to explain, meaning it is difficult to determine the elements justifying a decision made by the algorithm to produce its final output.Moreover, the creation of such falsified products that place victims in a situational context of pornography, typically targeting minorities such as women, transgender individuals, and children, is referred to as deepfake pornography - for the purpose of distinguishing this phenomenon, which requires non-consent, we understand, exemplified by the most famous case, to call these deepfakes "deepnudes".The creation of deepfakes has been growing exponentially worldwide: in 2023, 95,820 videos were found, a 550% increase compared to 2019. While this type of media has garnered attention for its potential political risks, the data paints a different picture - 98% of deepfake videos in 2023 are pornographic; and among these, 99% of the targets are women, according to a study by Home Security Heroes.If the non-consensual sharing of intimate images was already a growing concern, often linked to a pornographic worldview, what deepfake pornography brings to the scenario of digital gender violence and gender violence facilitated by technology is that sexual content doesn't even need to be present in the original image for people to share it - it can be artificially generated.In this way, any woman with her images available online would be vulnerable to experiencing harm from the creation of false content that could easily be mistaken for reality. This phenomenon not only dehumanizes the victims but also serves as a conduit for misogyny, often aiming to diminish or question the legitimacy and capability of women to occupy public spaces and positions of power.This type of digital abuse escalates within the realm of public figures, particularly targeting women disproportionately, manipulating their images to create sexually explicit content against their will. These attacks not only inflict immediate damage to the victims' reputation and mental well-being but also perpetuate a hostile online environment that may discourage female participation in public and professional spheres. The study "Protecting Public Figures Online" (Cover, et al, 2024) illustrates a significant gap in digital platform policies, often treating public figures as "fair game" for abuse, overlooking the nuances among different types of public figures and the institutional support each may receive.It is worth highlighting the inherent political dimension in the use of deepnudes tools, particularly when targeted at women in positions of influence. Perpetrators often aim to undermine the power of these women by resorting to attacks meant to ridicule and discredit, targeting their sexual dignity and honor. These attacks not only reflect resistance to the increasing female power in traditionally male-dominated spheres but also seek to reinforce outdated and harmful gender stereotypes, using sexualization as a weapon.Regardless of whether the images are real or not, the consequences for the victims can be just as devastating as revenge pornography, whose impact is well-documented in the literature. Victims of digital abuse often experience high rates of mental health issues, such as anxiety, depression, self-harm, and suicide, according to research conducted by Asher Flynn, a professor at Monash University. The effects can be felt both physically and mentally, impacting their employment, family, and social life, as well as having an inhibiting effect on women's freedom of expression.A revealing and concerning statistic on this matter is that in the United States, 73% of male users of deepfake pornography don't feel guilty about it (2023, Home Security Heroes). The main reasons cited include: knowing it is not actually the person; believing it doesn't harm anyone; viewing it as simply a more realistic version of sexual imagination; and considering it not much different from traditional pornography.The fact that users do not understand how actions that violate the notion of consent, which according to literature, have violent effects on victims, illustrates the significant gap in understanding on the subject and the need for an educational process on the various types of gender violence.The possible role of moderation in sexual and purportedly sexual contentThe way Meta handles these cases sheds light on the complex nuances and challenges of content moderation on social platforms. It is clear that implementing the Media Matching Service Bank as part of the automated enforcement system is an effort to enhance accuracy and efficiency in detecting recurrent violations. However, relying too heavily on this technology raises significant questions, especially when considering the intricate dynamics of gender, race, and power that influence the creation and circulation of digital imagery.Firstly, systems relying on automated decisions often lack the ability to interpret context and nuances, which can lead to misguided and unfair decisions. This practice speaks to a context of operational efficiency optimization, where models are driven to achieve ostensibly objective performance markers, but disregard concerns about the fairness and justice of moderation procedures.By relying on algorithms to identify and remove inappropriate content, platforms incur (a) a Sophie's Choice and (b) an analytical error. Sophie's Choice involves the quantitative aspect: how to filter, moderate, manage a large quantity of human-related content? The analytical error lies in giving a sort of blind trust, not based on epistemology, that the internal process of models, upon receiving inputs and generating outputs, constitutes an exercise in seeking objectivity or, even worse, diversity.In this realm, there is a risk of unjustly excluding legitimate expressions of freedom of speech, such as art, satire, or meaningful discussions. Safiya Umoja Noble's insightful analysis in her book "Algorithms of Oppression" provides a crucial starting point for exploring the complexities of reliance on automated systems and media matching algorithms, highlighting how digital algorithms, when not properly regulated and controlled, can perpetuate and even amplify biases and prejudices present in society.With this in mind, the Media Matching Service Bank, which is used to identify and automatically remove content previously deemed violative, may introduce and reinforce significant biases. Operating under the premise that past data can adequately predict and identify future violations, achieving failure is the likely outcome, especially when considering that this bank may be fueled by racially biased phrenological data, reflecting entrenched prejudices and historical inequalities inherent in the training and validation databases themselves.If a significant number of images flagged for policy violations belong to certain ethnicities or social groups, the system may learn to associate these groups with policy violations, even if such associations are unfair and harmful.These biases can be exacerbated over time, creating a cycle of algorithmic discrimination that perpetuates and reinforces intergenerational inequalities. An illustrative case is cited by Noble, who, when discussing the representation of racialized women, especially young black women, demonstrates how these women are often associated with negative and harmful stereotypes, such as pornography and objectionable behaviors. This distorted and harmful association contributes to the perpetuation of prejudices and stigmas against black women, reinforcing social inequalities and injustices.Recognizing the limitations and dangers of relying too heavily on automated systems and media correspondence banks in content moderation, it becomes clear that there's a need for a more integrated and careful approach. This should involve significant human participation in reviewing and judging individual cases. Not only will this help avoid unfair decisions, but it will also lead to a deeper understanding of the complexities involved. Thus, by removing content using an automated system based on biased data, there's a risk of censoring valid expressions of freedom (sexual, in the context of the cases), especially in scenarios where nudity or other forms of bodily expression are artistically or culturally significant.The line between protection against harassment and undue censorship is delicate and complex, requiring an approach that respects both individual dignity and freedom of expression. The need for human review is therefore not only a matter of accuracy but also of justice. Human reviewers, equipped with cultural context, sensitivity, and the ability to interpret nuances, are essential to ensure that content policies are applied in a way that honors both the platform's intent and users' rights. On the other hand, it is indispensable to ensure the dignity of the work of these reviewers, who often face challenging working conditions and disturbing material. It is crucial that their working conditions reflect the seriousness and difficulty of their tasks, ensuring adequate support and measures for their well-being.Nevertheless, the issue of automatically closing reports after 48 hours without review highlights significant limitations in the governance of these automated systems. By constraining the available time for review and analysis of reports, the platform runs the risk of making hasty or inappropriate decisions, which could lead to the unjust removal of legitimate content or the persistence of harmful content online.This approach not only disregards the need for time and reflection to ensure fair and balanced decisions but also exposes users to potential harm, as inappropriate content can remain active on the network for prolonged periods. Thus, ensuring transparency, explainability, and auditability in online content moderation processes becomes essential. These are fundamental in ensuring users' trust, their ability to interpret, judge, and hold platforms accountable, while also protecting individual rights.By making moderation policies, practices, and algorithms accessible and understandable, platforms give users a sense of control over their online environment, which enhances safety and satisfaction in their experience.It's worth mentioning that in a broader sense of security, as well elucidated by Bruce Schneier in the text "The Psychology of Security," concrete security measures should also be approached from the perspective of perceived security or the appearance of security. This fosters a virtuous cycle of best practices and education in technology management.Transparency enables users to grasp how their interactions are handled and what the expectations are regarding appropriate behavior, fostering a more inclusive and democratic environment. Furthermore, explainability ensures that decisions are comprehensible and justifiable, which is crucial for maintaining user trust and preventing arbitrariness or injustice.Finally, auditability enables independent and rigorous oversight of processes, helping to identify and correct biases, flaws, or abuses, thereby contributing to continuous improvement and enhancement of these systems. Together, these principles form the basis for effective and ethical content moderation, balancing user protection with the preservation of freedom of expression and respect for the diversity of opinions and perspectives.The adoption of these measures should also involve increasing the scope of multi-sector participation, by creating environments within the business model that encompass diverse inputs, across regional, sectoral representation, and legitimacy divides. Such measures, which integrate a slice of environmental complexity into business management, enhance the capacity for ongoing dialogues and granularity in the cognition and decision-making of these cases. This approach thus addresses the pretensions of universalist solutions that overlook the contextual needs of users in different societies, while also enabling consensus-building on actions, akin to algorithmic tuning based on the human element as the central filter.

Link to Attachment

PC-27086

2024-007-IG-UA, 2024-008-FB-UA	PC-27087	United States & Canada
Case number	Public comment number	Region
Erica	Olsen	English
Commenter's first name	Commenter's last name	Commenter's preferred language
National Network to End Domestic Violence		Yes
Organization		Response on behalf of organization

Full Comment

The proliferation of deepfake pornography poses multifaceted harms, encompassing not only privacy violations and reputational damage but also profound implications for societal trust, consent, and gender equality. Women are disproportionately targeted by this form of technology-facilitated abuse. And, women public figures and Black women are often targeted the most. The weaponization of their likenesses intensifies already substantial risks women face in their personal and professional lives. Image-based abuses, including deepfake pornography, have been described as torture for the soul with devastating impacts on social connections, sense of self, and willingness to trust others.Until Meta and the online social media industry develop more robust tools for the automated detection of deepfake pornography, community reporting will continue to be the first line of defense against abusive behavior on social platforms. As with any report of abuse or harm to an authority, the willingness of a person to make a report is directly related to their belief that their report will help to address the harm they are reporting. When a persons willingness to report is eroded by being ignored, both that person and the platform they are engaging in are less safe. To encourage reporting and have safer platforms, Meta should end the practice of automatically closing reports within 48 hours. Whatever benefit Meta experiences from this practice, these cases demonstrate that the cost of unassessed closed reports is safety. To users of Metas platforms, this practice sends a message that Meta will only take safety concerns seriously if they arent too busy. It is a fundamentally dismissive and belittling practice that discourages reporting. Instead of automatic closures, Meta should direct a portion of its considerable machine learning resources to help human moderators to appropriately triage incoming reports such that reports that are likely to be actioned on are responded to within 48 hours, and reports that are less likely to be actioned can be responded to in a greater amount of time. If, due to triaging, a report is not responded to with a meaningful amount of time, users should be given the option of resubmitting their report with an elevated status instead of having to resubmit the same issue as a new report, or appealing to the Oversight Board.

Link to Attachment

No Attachment

2024-007-IG-UA, 2024-008-FB-UA	PC-27088	Europe
Case number	Public comment number	Region
Olga	Jurasz	English
Commenter's first name	Commenter's last name	Commenter's preferred language
Centre for Protecting Women Online (The Open University, UK)		Yes
Organization		Response on behalf of organization

Full Comment

DID NOT PROVIDE

Link to Attachment

PC-27088

2024-007-IG-UA, 2024-008-FB-UA	PC-27092	Latin America & Caribbean
Case number	Public comment number	Region
Rubiela	Gaspar	English
Commenter's first name	Commenter's last name	Commenter's preferred language
Hiperderecho		Yes
Organization		Response on behalf of organization

Full Comment

Comments sent in the attached file

Link to Attachment

PC-27092

2024-007-IG-UA, 2024-008-FB-UA	PC-27093	United States & Canada
Case number	Public comment number	Region
Dhanaraj	Thakur	English
Commenter's first name	Commenter's last name	Commenter's preferred language
Center for Democracy & Technology		Yes
Organization		Response on behalf of organization

Full Comment

CDT Response to the Oversight Board's call for public comments: "Explicit AI Images of Female Public Figures"Authors: Dhanaraj Thakur and Asha AllenApril 2024IntroductionThe Center for Democracy & Technology (CDT) submits these comments in response to the Oversight Board's request for public comments on "Explicit AI Images of Female Public Figures." CDTs work includes assessing the impacts of online abuse on digital platforms and advocating for solutions that protect free expression, privacy and security, and other fundamental rights. Deepfakes are synthetic manipulations of identities and expressions in the form of video, images, or audio which make it appear as if someone says or does something they never did. As one well known study noted, the vast majority of deepfake videos are pornographic, and almost all of those are targeted at women. Other researchers have for some time highlighted concerns that women journalists and politicians are often targeted by deepfakes. The problem of false and sexualized information about women is not new. For example, researchers and journalists have identified cheapfakes or shallowfakes as the manipulation of media in less sophisticated ways compared to deepfakes, including crudely editing, mislabeling, or misrepresenting the original context of an image or video. In fact, researchers note that as far back as the 19th century there were documented examples of women in the U.S. who were warned that photographers could combine a photo of their face with that of another woman's body in a sexualized way. A key difference stemming from the advent of AI today is the increasing ease with which machine learning tools are becoming accessible and affordable through a network of websites and apps that allow users to produce and share deepfakes very quickly and regularly.The phenomena of deepfakes, cheapfakes, and their antecedents are not random observations but part of a systemic problem - patriarchy. Patriarchy exists where positions of power in political, social, and economic structures and organizations in a country are dominated by men. When control of these systems of power are perceived to be under threat, as is the case with the increase in the number of women running for political office in the U.S., we observe a disproportionate amount of online harassment and abuse targeted at those women. Such harassment and abuse based on one's gender expression (or online gender-based violence - GBV) can take a range of forms, including non-consensual image/video sharing, and more specifically creating and sharing fake images/video without consent.However, deepfakes are not only an expression of violence, as they also are created to spread false information about persons or groups based on their gender identity (i.e., gendered disinformation). Such disinformation campaigns often include deepfakes as an attempt to undermine a woman's ability to participate in representative politics by harming them, their staff, and their political candidacy in potentially severe ways. The use of deepfakes targeted at women in politics in particular is a form of online GBV and gendered disinformation that is meant to challenge, control, and attack their presence in spaces of public authority. The Impacts of Deepfakes on Women in Public Life. Deepfakes can impact politically engaged women, including candidates, journalists, advocates, and civic leaders, in a variety of ways, not only on the individual level, but on women as a group. For those experiencing these videos and images firsthand, they can prove to be persistent forms of distraction: by trying to regularly refute such attacks and falsehoods, women candidates will have less time to focus on substantive issues, and the wider discussion about them will follow that pattern as well. Further, such experiences can cause personal harm, such as distress, as well as a chilling effect on political or other speech. More broadly, the severity of some deepfake videos as part of a larger campaign of online harassment and disinformation targeted at a woman politician can make other women who are interested in politics more likely to reconsider their ambitions, which in turn harms efforts to build and sustain inclusive democracies. Similarly, the prospect of harassment and other kinds of abuse that can follow from deepfakes can actively discourage women and gender nonconforming individuals in political roles from expressing themselves online in a way that might draw public attention and scrutiny. Women who are the subject of these campaigns can also face significant long-term effects as, given their severe nature, some of these attacks can yield physical and psychological damage that requires longer recovery times, with implications for their political careers. These harms are equally experienced by women journalists who, as essential civic space actors, are often confronted with similar campaigns aimed to discredit their journalistic efforts and which, in some cases, lead to threats against their physical safety.Intersectionality and other harms Gender only represents one type of identity and is perhaps only a starting point for trying to understand the various impacts of deepfakes. In reality, individuals traverse multiple identities all the time, and disinformation can also operate across race, gender, and other aspects of identity simultaneously. Recognizing the reality of intersectional identities challenges researchers and policymakers to understand both how a person may have to contend with multiple sources of oppression at the same time, and the unique impact from this multifaceted oppression. Among other problems, this means that the unique experiences and needs of people who are, for example, neither white nor male can go unexamined and unaddressed in research and policymaking. Limiting our analysis and measures to address deepfakes to the population at large may in turn undermine our ability to effectively counter the harm that such disinformation campaigns and online GBV have on democracy and attempts to advance gender equality. When we think about deepfakes we should therefore recognize that they will be used to exploit existing forms of discrimination not only based on gender, but also a range of other identities such as disability status, LGBTQIA+ communities, age, religious background and immigration status. Although there is limited research using intersectionality to examine deepfakes, we do have some related evidence from other forms of online GBV and gendered disinformation targeted at women politicians that may be instructive. From a study of posts on Twitter/X that targeted a representative sample of all candidates that ran for Congress in the 2020 U.S. election we found that:Women of color candidates were twice as likely as other candidates to be targeted with or the subject of mis- and disinformation, which often included cheapfakes or manipulated images and photos.Women of color candidates were the most likely to be the target of particular forms of online abuse, including sexist abuse (as compared to white women), racist abuse (as compared to men of color), and violent abuse (four times more than white candidates and two times more than men of color.)Women of color candidates were also most likely to be targeted with or the subject of posts that combined mis- and disinformation and abuse. When we interviewed women of color that ran in those elections, they reported feeling diminished, questioning their worth, and other negative effects. In other words, they perceived the purpose was for them to drop out of politics and to accept the oppression they faced.The use of automated solutions to address these harmsWhile many online platforms use various forms of automated technologies to detect and analyze user generated content, some actors are developing new techniques to evade such systems. Deepfakes emerged as one such circumvention effort. There may be some legitimate use cases of the technologies underlying deepfakes, in fields like movie production, game design, or improving quality of real-time video streams. That said, deepfake detection is presently a major industry priority and challenge. Using AI based tools to detect deepfakes can introduce additional challenges, however, depending on how they are incorporated within existing trust and safety systems. For example, such detection systems could introduce bias depending on the models and data used; they may lack explainability, which could be important particularly during an appeals process with end-users; they often cannot assess context as well as humans; and it's difficult to assess their performance because of a lack of useful metrics, particularly those that can be easily understood by non-experts. Another potential challenge could be the design of detection systems that focus on deepfakes in general. However, as noted earlier, most deepfakes are pornographic and target women specifically, which should have implications for how such systems are designed in the first place. These limitations point to the importance of complementary mechanisms such as user reporting of deepfakes. In fact, in the cases outlined by the Oversight Board in this call for public comments, one includes user-reporting. Given the focus on female public figures and that (as noted earlier) women of color political candidates in the U.S. are more likely to be targeted with online abuse and disinformation, it is also important to understand their perspectives on reporting mechanisms. Most candidates we spoke to in our study described using social media platform reporting mechanisms at least once. Of these reports, only one respondent successfully petitioned the platform to remove content that was false or abusive. According to these women of color candidates, the platform least responsive to user reporting was Facebook. In another study which included women journalists, users felt that "reporting mechanisms on social media platforms are often profoundly confusing, time-consuming, frustrating, and disappointing."A Harmonised Approach to tackling deepfakes in the context of Online GBVThe two cases being reviewed by the Oversight Board are of particular relevance to legislative advancements in the European Union. The EU co-legislative bodies have just adopted the final text of the Directive to combat violence against women and domestic violence, which aligns with the international standards already established by the Council of

Europe Istanbul Convention and its first General Recommendation. Once transposed into national law, this means that across the EU, the production, manipulation, or altering of an image, video, or similar content which make it falsely appear as though a person is engaged in sexually explicit activities, without that persons consent, and subsequently making that content publicly accessible, will be a criminal offense punishable up to at least one year of imprisonment. The decision of the Oversight Board in these cases therefore may need to take into due regard these obligations and the necessity of ensuring coherence at a global level. General Recommendations for Meta to address Deepfakes targeted at Women in Public LifeMeta should clearly articulate policies that prohibit content such as deepfakes that harasses or abuses someone on the basis of gender or race. These policies, and the moderation processes that enforce them, should adopt an intersectional approach that considers the unique ways in which abuse can manifest against women with multiple identities.With regard to women politicians, Meta should ideally provide transparency reports around election mis/disinformation before, during, and after an election. These reports could provide a holistic view into content moderation and integrity operations by the service during the period around a specific election, and should include a focus on online GBV, gendered disinformation, and deepfakes that target political candidates, broken down by demographics. Meta should make data available to independent researchers that enables them to study the nature and impact of deepfakes, gendered mis- and disinformation, and online GBV on political candidates. This includes annual risk assessments performed in the context of Article 34 of the DSA, which expressly requires mitigation of risks related to the spread of disinformation and GBV.Meta should take additional steps to protect and prevent abuse, particularly explicit AI images and other sexualized deepfake abuse targeting women political candidates, journalists, and other public figures. They should:Offer tools that allow users to report content that violates the companies policies against abuse or mis- and disinformation including additional tooling (e.g., granular levels of control) for verified accounts including women public figures, to quickly escalate abuse reports to specially trained moderators.Ensure that content moderation systems, including human moderators and algorithmic systems, are attuned to the needs of and the threats faced by women public figures, and women public figures whose identities maybe particularly targeted in a given society (e.g., women of color in the U.S. and women of caste oppressed and religious minority communities in India).

Link to Attachment

PC-27093

2024-007-IG-UA, 2024-008-FB-UA	PC-27094	United States & Canada
Case number	Public comment number	Region
Ilse	Knecht	English
Commenter's first name	Commenter's last name	Commenter's preferred language
Joyful Heart Foundation		Yes
Organization		Response on behalf of organization

Full Comment

April 30, 2024Joyful Heart Foundations Comments to Meta on Deepfake Abuse The Joyful Heart Foundation is honored to provide comments for Meta on the terrible scourge of Image Based Abuse (IBA) and deepfake technology. We are grateful for the work that Meta has done to address this issue, especially related to the exploitation of children. More must be done to project kids and adults who are increasingly targets of online sexual assault. In the U.S., 1 in 12 adults reported being victims of image-based abuse. Real rates could be much higher; 1 in 3 people reported being victims in the UK and Australia. This exponentially growing problem threatens women and girls health and safety and often derails their lives forever. The FBI recently warned that young boys are often becoming victims of sextortion. More must be done to stop the creation, publication, selling, threats to disseminate, and dissemination of intimate images of people who suffer greatly as a result of this exploitation. When addressing image based abuse, the issue of real or synthetic images naturally surfaces. The problem of nonconsensual deepfake pornography has made countless headlines in the last few years, building to a crescendo with the Taylor Swift incident. In less than 24 hours, the altered sexual images of her were viewed 45 million times, showing how quickly these photos can spread once they make it to the internet. The creation of deepfakes used to be limited to someone with enhanced skillssomewhat of a computer whizto make a believable version, but new technologies, increasingly available to anyone, makes it easier than ever to create a photo or video that looks very real. Prevalence It's clear that this global problem is spiraling out of control. Various numbers exist but all recent reports agree that 96-98% of all deepfakes are pornographic in nature, and the vast majority were created without the subject's consent. According to a 2023 report by Home Security Heroes, 98% of deepfake visuals are pornographic in nature, and 99% target women. According to Sensity, 96% of deepfakes are sexually explicit and feature women who didnt consent to the creation of the content. Many of these altered images include underage girls. The pace at which this industry is growing is alarming. The advocacy group, My Image My Choice, reports that there are currently (as of January 2024) 276,149 deepfake images online with a total number of 4,219,974,115 views, a 1,780% increase compared with 2019. Professor Danielle Citron, at the University of Virginia School of Law, asserts there are more than 9,500 sites devoted to this type of offense. Tens of millions of viewers visit these websites which are all easily discoverable via Google search. These videos are advertised on various platforms, one called Discord, which is used by 100 million people (https://discord.com/. This is not an issue that is hidden in the shadows; it's blatantly being posted all over the internet. It's not a surprise that nonconsensual pornographers target young people more than older women. People between the ages of 20 and 29 years were twice as likely as those aged over 40 to be victims of image-based sexual abuse (Panorama Global Report, 2024). Native Alaskans, Indigenous North Americans or African Americans were found to suffer more from the threat to distribute intimate material compared to White participants. In the context of interpersonal violence, it's become increasingly common for an abuser or sex trafficker to share or threaten to share an intimate image or video to exert power and control over the victim (Maddocks, 2018). In the case of domestic violence, the general public and the media call this revenge porn, of course; the intent of the perpetrator is to get back at someone, to harm that specific person and manipulate or punish them as part of their abusive pattern of behavior. While Joyful Heart Foundation recognizes this comment is supposed to be focused on people in the public eye who have been exploited, everyone can be a target of a deepfake developer. Middle and high schooler girls in California, New Jersey, and New York, and many other states have recently been the subject of deepfakes. Too many ordinary women and girls have received that text or call from a friend, you need to look at this or Im sorry to tell you this and find their faces pasted on a body that is not theirs, doing things that they have not done, nor would ever do. It's important to note that many of these videos depict women and girls in degrading and sadistic situations, such as being tied up and urinated on or of course sexually assaulted. Importantly, there have been an incredible amount of women in the public arena who have been targeted by deepfake creators: ASMR YouTubers, politicians or other public officials, celebrities, pop stars, journalists. Any woman in the public realm is potential prey for these bad actors. Victim Impact Survivors of Image based abuse (IBA) suffer real life anguish & deep personal violation. The rallying cry of those working on this issue is the image may be fake, but the harm is real. It's absolutely crucial to not underestimate the impact image based abuse has on victims. Whether images are original, AI generated, or digitally altered, if they are shared nonconsensually, the action is injurious to the targeted person. Research finds that victims of IBA suffer from many psychological, mental and physical effects, including PTSD, depression, and anxiety. Many who have endured this cruelty say they feel violated, vulnerable, and a loss of safety and control over their lives, feelings that survivors of sexual assault often cite. Some survivors have committed suicide after discovering deepfake videos were made with their likeness.Survivors experience fear and anxiety because in many deepfake cases they dont know who created or disseminated the image. Some liken the experience to being stalked, making them look over their shoulder all the time. Taylor, the 22-year-old engineering student who is the subject of the SXSW Special Jury Award-winning feature documentary, Another Body, says she experienced extreme OCD and anxiety, causing her to reevaluate her social circle. Many survivors cite social isolation and feeling alone as a common after effect. Social stigma, public judgment, shame, and humiliation are all heaped on survivors of IBA.IBA survivors often face severe financial repercussions as they change or lose jobs after being targeted in this way. For some, careers they have dreamed about are washed away due to the images living online in perpetuity. Withdrawing from their public life, whether its online or in-person, is common. Those with lived experience have explained that when this happened to them it was life shattering, a wash of pain, and hell on Earth. Sadly, for most survivors, getting the content removed is a long and arduous path. Many have described the seemingly never-ending process of identifying images, requesting take down, follow up, and on and on as often futile, frustrating, and exasperating.   Legislative and other recourseAt this moment in time, there are few laws on the books that address deepfake abuse at the state level, and now prohibition exists at the federal level. It's clear the U.S. needs a comprehensive legislative approach to image based abuse. But this will take time, and survivors dont have time to wait. What can be done right now is in the hands of companies like Meta, who can immediately create rules around what is allowed on their platforms, what is not, and how and when it will be taken down. Meta can create practices that are survivor centered and provide clear paths to survivors to help them get images taken down expeditiously. Joyful Heart Foundation exists to help survivors heal and reclaim joy in their lives and to end injustice that is perpetrated against a person because of their gender. Everyone deserves to live a life free from degradation, dehumanization, and abuse. Everyone deserves to have control over their body and self image, whether in person and online. Nonconsensual intimate images in general serve no public good interest and Meta should create a system whereby the images are immediately removed, and if need be, reinstated after a full investigation is conducted. While we are not experts in technology and will therefore refrain from suggesting technological answers to these issues, we are confident that the responses do exist. If humans can create artificial intelligence, they can surely create systems that identify it and protect against it. Removing images that are posted against a persons will and that can unravel a persons life should be a top priority of Meta, and removing it should be done with expediency. We stand ready to assist in any way we can and We are grateful for the opportunity to provide information about the impact of the posting of nonconsensual deepfake sexual images on survivors. Joyful Heart Foundation looks forward to Metas next moves to protect women and girls from harassment and sexual violence online. With gratitude,Ilse Knecht Policy and Advocacy Director

Link to Attachment

PC-27094

2024-007-IG-UA, 2024-008-FB-UA	PC-27095	United States & Canada
Case number	Public comment number	Region
Sam	Gregory	English
Commenter's first name	Commenter's last name	Commenter's preferred language
WITNESS		Yes
Organization		Response on behalf of organization

Full Comment

DID NOT PROVIDE

Link to Attachment

PC-27095

2024-007-IG-UA, 2024-008-FB-UA

Case number

PC-27096

Public comment number

Latin America & Caribbean

Region

Ana

Commenter's first name

Ardila

Commenter's last name

English

Commenter's preferred language

Fundación Karisma

Organization

Yes

Response on behalf of organization

Full Comment

The nature and gravity of harms posed by deepfake pornography including how those harms affect women, especially women who are public figures. Discrimination and violence against women has become entrenched in the online space, with devastating consequences. Contents with deepfake pornography are part of the behaviors that lead to digital violence suffered by women, which can lead them to leave these spaces for good or remain in them at the expense of their peace of mind. Thus, in addition to putting their lives, health and integrity at risk, these aggressions threaten their freedom of expression and information, their privacy and their personal data. Sometimes, even that of their closest people and affect the public space and debate where their voices are silenced.It is important to highlight that deepfake pornography content, for the most part, involves the use of images or video of a woman to make and viralize sexual content without her consent. This has a disproportionate impact on women's lives, from being subjected to sexual harassment and aggression, to the expected roles, social isolation, increased anxiety, depression, and even suicide. These consequences often extend over time and carry over into their offline lives. It should be noted that people that experience this specific type of violence are usually ones that defy traditional gender roles, such as public women, activists, journalists, female politicians, sex workers or people with diverse sexual orientation or gender identities.Strategies for how Meta can address deepfake pornography on its platforms, including the policies and enforcement processes that may be most effective. Taking into account the above and recognizing that Meta is not responsible for third-party publications that violate its community policies, we consider that a good practice to address deepfake pornography on its platforms is the inclusion of policies to assist victims of publications with this type of content, as a clear form of digital gender-based violence.We insist that, although Meta is not called to assume responsibility for non-compliance with its content policies by third parties, it is important to recognize that it lends the possibility that people, regardless of community policies, publish any type of content, which, although it can be removed as a sanction for non-compliance with policies, has already caused damage, almost irreparable for a person (usually a woman). In this sense, we name it from good practices because, starting from the question of how to address gender violence on their platforms, the need to address it in a comprehensive manner and not exclusively from the moderation and curatorship arises. The above is based precisely on the effects of publications such as those that gave rise to the cases under study. Affectations that persist in women victims of this violence, even when the publications have been removed from the platforms. The viral capacity of social media demand different measures by different actors. This is why, based on the theory of restorative justice with a gender approach, we find it advisable for Meta to explore alternatives of attention for this type of cases that focus on the victims, especially when it comes to women and gender-dissident people, and not exclusively on the punishment of the publisher or on the elimination of the content that generates the affectation. Restorative justice was born in the face of the inability of criminal justice to respond to the needs of victims and society in general in terms of compensation for the harm caused. Under the retributive justice model, the punishment of those responsible for violence, by itself, does not contribute to improving the lives of the victims or to promoting social change. In this sense, when the focus is placed solely on punishment (or in this case the elimination of the content), the victim is displaced, giving her an almost invisible role, even when the after-effects of the harm remain with her.However, restorative approaches and practices undoubtedly go beyond the criminal justice system and even the justice systems, configuring ways to address the subjectivities of the victims and their specific needs to recompose what the victimizing event has affected.In this framework, implementing reparation measures for victims of deepfake pornography from a gender-based restorative justice approach highlights its importance, taking into account that, as Rita Laura Segato expresses, "inequalities and the patriarchal system are a constant in women's lives, and the breeding ground for situations of disadvantage and inequality".Women's lives before the victimizing events are already marked by multiple forms of violence and inequalities; therefore, the gender approach in reparation actions allows understanding and addressing inequality and structural violence as a factor that puts women at risk of experiencing digital violence with greater intensity. In addition to the above, the gender approach must be accompanied by an intersectional approach because, although the cases studied involve women who are public figures, not all cases happen to these women and, instead, many of them may be happening to women who do not have the necessary resources to make visible and even address the damage caused by digital gender-based violence. According to Judith Butler, there is a perception of women's lives as "unweepable" or less valuable, especially when they come from racialized or structurally impoverished contexts. In a digital context where sexism and classism insert stereotypes according to which only certain lives are valuable and protected (heterosexual, white, middle or upper class, non-migrant, cisgender), reparation policies for victims of digital violence with a gender and intersectional approach bet on the effective redress of the damage thought from the context of each person, in which multiple identities and forms of oppression converge.In conclusion, when considering the impacts of digital violence of publications with deepfake pornography, in addition to wondering how to address the complaints from the moderation and curatorship of content, it is worthwhile and advisable to ask how, from Meta as a digital public space, these situations can be addressed from alternatives of care and support for women victims of this violence, seeking to restore their integrity and dignity.

Link to Attachment

PC-27096

2024-007-IG-UA, 2024-008-FB-UA	PC-27097	Asia Pacific & Oceania
Case number	Public comment number	Region
Withheld	Withheld	English
Commenter's first name	Commenter's last name	Commenter's preferred language
Withheld		No
Organization		Response on behalf of organization

Full Comment

Deepfake pornography targeted at female public figures causes severe harm at both the individual and societal level. At the individual level, it can cause emotional distress and reputational damage. At the societal level, it contributes to a misogynistic culture which punishes women who dare to assume powerful roles and silences other women with similar aspirations.To address deepfake pornography, Meta should commission a study to evaluate the efficacy of different content moderation techniques. Here, it is crucial to find the right mix of automation and human review. If there is too much automation, it may compromise free speech. And if there is too much human review, the removal process may be too slow. These are extremely important decisions and they must be grounded in actual data available with Meta about the total volume of complaints, the proportion of frivolous versus genuine complaints, the accuracy and speed of automated systems versus human reviewers, etc.An exception should be carved out for deepfake pornography from Metas policy that automatically closes appeals in 48 hours. Women who have already been victimized by deepfake pornography should not be subjected to further harm merely because of Metas inability to promptly review their complaint.

Link to Attachment

PC-27097

2024-007-IG-UA, 2024-008-FB-UA	PC-27098	United States & Canada
Case number	Public comment number	Region
Lucy	Qin	English
Commenter's first name	Commenter's last name	Commenter's preferred language
Massive Data Institute at the McCourt School of Public Policy, Georgetown University		Yes
Organization		Response on behalf of organization

Full Comment

Image-based sexual abuse (IBSA) is a spectrum of violence that includes the non-consensual creation of synthetic intimate content using AI. While this is frequently referred to as deepfake pornography, in line with IBSA terminology [1] we refer to it as AI-generated non-consensual intimate imagery (AIG-NCII). IBSA also includes other forms of abuse such as the non-consensual distribution of consensually shared intimate content. As a whole, IBSA is unfortunately not rare: an estimated one-third of adults have been subjected to IBSA [1] (similar to other types of sexual violence [2]). IBSA causes significant harms as victim-survivors often experience severe health consequences, such as post-traumatic stress disorder, anxiety and depression [3][4][5]. Additionally, IBSA may threaten their physical safety, reputation, or job security [6].

The abuse is more prevalent in marginalized groups such as young women [7][8][9], LGBTQ+ people [4][10], migrants [10], and survivors of intimate partner violence [11]. Victim-blaming attitudes are common in IBSA and victim-blaming has been associated with poorer outcomes for victim-survivors, acting as a barrier for reporting and help-seeking, worsening victim-survivors already impacted mental health [3]. Due to societal stigma and insufficient education around IBSA, individuals may also be unaware of available resources in the event that they do experience harm [12]. IBSA and AIG-NCII are growing global issues [13]. Policy on IBSA is sparse in most countries [14][15]; in the US specifically, legal scholars have called for legislation to sufficiently address its harms [16][17][18] and President Bidens October 2023 Executive Order on Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence [19] calls for a report identifying the existing standards, tools, methods, and practices, as well as the potential development of further science-backed standards and techniques, forpreventing generative AI from producing child sexual abuse material or producing non-consensual intimate imagery of real individuals (to include intimate digital depictions of the body or body parts of an identifiable individual). However, policymaking efforts toward combatting AIG-NCII have not kept pace with technological advancements. Given these gaps, platforms must bear the responsibility of protecting its users and proactively develop strategies that reduce harms produced by AIG-NCII. Coinciding with gaps in policymaking, legal recourse as an avenue for pursuing justice may be prohibitively expensive and, particularly in the case of AI-generated content, may fall outside of existing legal frameworks. Legal recourse for non-synthetic NDII is limited in its effectiveness as it may also be prohibitively expensive, lack jurisdiction (e.g., for content shared outside of the U.S., content shared anonymously), and/or not pursued due to fears of additional harassment. In ongoing research in which we interviewed organizations that support victim-survivors, the practitioners we spoke to emphasize the additional barrier of finding a lawyer who is trained to assist with issues around IBSA. This challenge of finding support extended to therapists and law enforcement, which practitioners mentioned as potentially ill-equipped and ill-trained to assist in cases of IBSA due to the recent proliferation of harms and IBSA being a relatively novel issue. Given limited options for support, platforms can directly assist victim-survivors by removing harmful content as quickly as possible. The existing processes for content removal of all forms of non-consensually distributed intimate content (NDII) regardless of whether AI or human-created from platforms are ineffective. In ongoing research, we interviewed 20+ victim-survivors of NDII. They reported frustrations with content takedown on various social media platforms. They cited long response times, a lack of response, and the need to make multiple reports from different accounts. This is consistent with the experiences of non-profit organizations who support victim-survivors by reporting content on their behalf. In their 2022 annual report, the Revenge Porn Helpline (a UK-based organization) stated that making takedown requests is a manual process that requires persistence and time as most requests need additional follow-up due to platforms being unresponsive and/or uncooperative [20]. Interviews with NGO staff members that support victim-survivors in South Asia also described similar frustrations and lengthy delays in response times [21]. In our research, we conducted interviews with organizations that support victim-survivors. Our findings are consistent with these narratives. We interviewed individuals from 6 organizations that support victim-survivors of various forms of intimate image abuse, including those who have had AI-generated content shared of them. They reported that the most pressing and urgent form of support victim-survivors needed was content removal. As a victim-survivor support organization shared with us, The immediate overriding need is simply [to] get the images down. Its very rare that someone in crisis is asking for a lawyer, or police, or even a therapist or anything like that.Lack of response to content removal requests such as the two cases describe have consequences. While waiting for content to be removed, victim-survivors and/or those supporting them are repeatedly exposed to abusive content while checking to see if it has been removed. Each time, this can directly re-traumatize victim-survivors. For staff that support victim-survivors, this has contributed to what the Revenge Porn Helpline noted as, vicarious trauma. In their 2022 annual report, they stated that, viewing and reporting content is incredibly stressful. It can take a large toll on the team, mentally and emotionally...the imagery can be disturbing and the comments can be vile. This kind of exposure can impact everyone differently and can form the basis for compassion burnout [20].The 48-hour window of response is far too lengthy as it allows the content to be further stored and shared, potentially leaving the platform. In addition, the automatic closing of cases beyond 48-hours is unacceptable in that repeatedly filing content removal requests taxes victim-survivors and may require them to revisit the abusive content. In the meantime, while reports are ignored, the non-consensual replication and sharing of intimate content greatly increases harm caused to victim-survivors. As long as this content is present on Metas platforms, it can be stored by anyone viewing the content. This creates greater the potential for it to re-emerge, creating long-term stress about whether the content may be re-shared. This also drastically increases the workload for the victim-survivor when pursuing content removal, especially if the content is shared on other platforms. Content initially shared on Meta may end up on similar social media platforms like X or TikTok, but even more concerning is when it is shared to other platforms or online channels that are solely dedicated toward non-consensually sharing intimate content.

These groups often create repositories of NDII and are used to engage in further abuse such as doxing and harassment [22][23]. Currently, Metas Bullying and Harassment Policy does not explicitly address IBSA or AIG-NCII. This needs to be an explicit part of the policy with dedicated options for reporting intimate image abuse that receives expedited review within the same business day. For example, on Instagram, reporting all forms of intimate image abuse currently falls under Nudity or sexual activity or Bullying or harassment. Within these categories, there is no explicit option for reporting non-consensually shared intimate content, whether AI-generated or not. While all forms of harassment need to be dealt with swiftly, non-consensual shared intimate content must be treated with even greater urgency. All forms of IBSA should be triaged with dedicated staff as quickly as possible through a designated reporting subcategory.Alongside improving mechanisms for reporting AIG-NCII and other forms of IBSA, Meta has a responsibility toward proactively identifying advertisements for apps that are used to create AIG-NCII. As per reporting from 404 Media, a reporter noted five different apps that advertised on Metas platforms the ability to nudify [24]. Though Meta removed these ads after being pointed toward them, Meta must invest resources in proactively identifying similar ads and continuing to promptly remove them. Meta should also commit more resources to their Trusted Flagger program so that organizations that support victim-survivors may quickly remove non-consensually shared intimate content on the behalf of victim-survivors. In our research, smaller organizations mentioned facing barriers to being included in Trusted Flagger programs across platforms. One organization shared with us that they were unable to get direct access to a Trusted Flagger Program because of their size. However, they were handling a large scale of content removal request from victim-survivors themselves and had to access a Trusted Flagger program through a larger organization in their country that did not directly support victim-survivors. Meta should prioritize and quickly review requests to join the Trusted Flagger program and actively work toward expanding this network. IBSA and in particular, AIG-NCII, will only continue and increase in scale as generative AI tools become more available. Strategies, policies, and interventions must be developed now to protect users and create safe environments for them to share images and express themselves online. Though it will require cultural change to prevent IBSA perpetration as a whole, Meta has the power to reduce the amount of harm experienced by victim-survivors in the aftermath. By dedicating resources toward improving reporting processes for IBSA and creating a more trauma-informed experience, Meta has the opportunity to lead the industry in protecting users.[1]A. Powell, A. J. Scott, A. Flynn, and S. McCook, A multi-country study of image-based sexual abuse: Extent, relational nature and correlates of victimisation experiences, J.

Sex. Aggress., vol. 0, no. 0, pp. 116, 2022, doi: 10.1080/13552600.2022.2119292.[2]Centers for Disease Control and Prevention, Preventing Sexual Violence\textbarViolence Prevention\textbarInjury Center\textbarCDC. Jun. 22, 2022. Accessed: Dec. 15, 2021. [Online]. Available: https://www.cdc.gov/violenceprevention/sexualviolence/fastfact.html[3]S. Bates, Revenge Porn and Mental Health: A Qualitative Analysis of the Mental Health Effects of Revenge Porn on Female Survivors, Fem. Criminol., vol. 12, no. 1, pp. 2242, Jan. 2017.[4]A. A. Eaton and C. McGlynn, The Psychology of Nonconsensual Porn: Understanding and Addressing a Growing Form of Sexual Violence, Policy Insights Behav. Brain Sci., vol. 7, no. 2, pp. 190197, Oct. 2020, doi: 10.1177/2372732220941534.[5]C. McGlynn et al., Its Torture for the Soul: The Harms of Image-Based Sexual Abuse, Soc. Leg. Stud., vol. 30, no. 4, pp. 541562, 2021, doi: 10.1177/0964663920947791.[6]T. Kirchengast and T. Crofts, The legal and policy contexts of revenge porn criminalisation: The need for multiple approaches, Oxf. Univ. Commonw. Law J., vol. 19, no. 1, pp. 129, Jan. 2019, doi: 10.1080/14729342.2019.1580518.[7]Amanda Lenhart, Michelle Ybarra, and Myeshia Price-Feeney, Nonconsensual Image Sharing: One in 25 Americans Has Been a Victim of Revenge Porn, Data & Society Research Institute, Dec. 2016.[8]A. A. Eaton, S. Noori, A. Bonomi, D. P. Stephens, and T. L. Gillum, Nonconsensual Porn as a Form of Intimate Partner Violence: Using the Power and Control Wheel to Understand Nonconsensual Porn Perpetration in Intimate Relationships, Trauma Violence Abuse, vol. 22, no. 5, pp. 11401154, Dec. 2021, doi: 10.1177/1524838020906533.[9]T. Mckinlay and T. Lavis, Why did she send it in the first place? Victim blame in the context of revenge porn, Psychiatry Psychol. Law, vol. 27, no. 3, pp. 386396, May 2020, doi: 10.1080/13218719.2020.1734977.[10]L. Vitis, Private, Hidden and Obscured: Image-Based Sexual Abuse in Singapore, Asian J. Criminol., vol. 15, no. 1, pp. 2543, Mar. 2020, doi: 10.1007/s11417-019-09293-0.[11]C. McGlynn, E. Rackley, and R. Houghton, Beyond Revenge Porn: The Continuum of Image-Based Sexual Abuse, Fem. Leg. Stud., vol. 25, no. 1, pp. 2546, Apr. 2017, doi: 10.1007/s10691-017-9343-2.[12]Lucy Qin, Vaughn Hamilton, Sharon Wang, Yigit Aydinalp, Marin Scarlett, and Elissa M. Redmiles, Did They Consent to That?: Safer Digital Intimacy via Proactive Protection Against Image-Based Sexual Abuse. Mar. 07, 2024. [Online]. Available: https://arxiv.org/abs/2403.04659[13]A. Flynn, A. Powell, A. J. Scott, and E. Cama, Deepfakes and Digitally Altered Imagery Abuse: A Cross-Country Exploration of an Emerging form of Image-Based Sexual Abuse, Br. J. Criminol., vol. 62, no. 6, pp. 13411358, Dec. 2021, doi: 10.1093/bjc/azab111.[14]K. Williams, Exploring Legal Approaches to Regulating Nonconsensual Deepfake Pornography. TechPolicy.Press, 2023.[15]R. Williams, Text-to-image AI models can be tricked into generating disturbing images, MIT Technol. Rev., 2023, [Online]. Available: https://www.technologyreview.com/2023/11/17/1083593/text-to-image-ai-models-can-betricked-into-generating-disturbing-images/[16]D. K. Citron and R. Chesney, Deep Fakes: A Looming Challenge for Privacy, Democracy, and National Security, Calif. Law Rev., vol. 107, p. 1753, 2019.[17]R. A. Delfino, Pornographic Deepfakes: The Case for Federal Criminalization of Revenge Porns Next Tragic Act, Fordham Law Rev., vol. 88, p. 887, 2019.[18]A. P. Gieseke, The New Weapon of Choice: Laws Current Inability to Properly Address Deepfake Pornography, Vanderbilt Law Rev., vol. 73, p. 1479, 2020.[19]Executive Order No. 14110 - Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence, no. 14110. U.S. Government Printing Office, 2023.[20]Z. Ward, Revenge Porn Helpline Report 2022, Revenge Porn Helpline, 2022. [Online]. Available: https://swgfl.org.uk/research/revenge-porn-helpline-2022-report/[21]N. Sambasivan et al., They Dont Leave Us Alone Anywhere We Go: Gender and Digital Abuse in South Asia, in Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, Glasgow Scotland Uk: ACM, May 2019, pp. 114. doi: 10.1145/3290605.3300232.[22]N. Henry and A. Flynn, Image-Based Sexual Abuse: Online Distribution Channels and Illicit Communities of Support, Violence Women, vol. 25, no. 16, pp. 19321955, Dec. 2019, doi: 10.1177/1077801219863881.[23]S. Semenzin and L. Bainotti, The Use of Telegram for Non-Consensual Dissemination of Intimate Images: Gendered Affordances and the Construction of Masculinities, Soc. Media Soc., vol. 6, no. 4, p. 205630512098445, Oct. 2020, doi: 10.1177/2056305120984453.[24]Emanuel Maiberg, Instagrams Nudify Ads, 404 Media, Apr. 22, 2024. [Online]. Available: https://www.404media.co/email/d2bebba9-5808-44fc-8352-d93d1791a5ff/?ref=daily-stories-newsletter

Link to Attachment

PC-27098

2024-007-IG-UA, 2024-008-FB-UA	PC-27099	United States & Canada
Case number	Public comment number	Region
Andrea	Powell	English
Commenter's first name	Commenter's last name	Commenter's preferred language
The Reclaim Coalition		Yes
Organization		Response on behalf of organization

Full Comment

April 30, 2024 Oversight Board 1601 Willow Road, Menlo Park, CA 94025 Re: Public Comment for Explicit AI Images of Female Public Figures Dear Meta Oversight Board, Thank you for giving The Reclaim Coalition the opportunity to engage and provide commentary on the case studies and Metas policies therein related to the synthetic nonconsensual creation and distribution of deepfake abuse of public figures. The Reclaim Coalition to End Online Image-based Sexual Violence, formally launched in June 2023, is a global, survivor-centered movement working collectively to end online image-based sexual violence in all its forms, powered by Panorama Global. We envision a world where everyone may be freely and safely onlinewithout the threat of image-based sexual violence. The Reclaim Coalition serves as an inclusive platform for coordination, collaboration, and convening amongst leading advocates, tech experts, civil society groups, and government regulatory bodies. This global network of experts working across 20+ countries is joining forces for the first time to end online image-based sexual violence and ensure all survivors have access to justice and healing. The Harms to Victims Caused by Nonconsensual Exploitation AI Images: Nonconsensual Explicit AI Images, often referred to as deepfakes or nonconsensual deepfake abuse cause profound and often life altering harm. As reported by The Reclaim Coalition in a recent New York Times op ed, most survivors with whom we have engaged have contemplated suicide as a result of their very public intimate image abuse. The fact that the content is created by AI does not diminish the harms that follow and, in some cases, amplify them.   Even though the online content is AI-generated, the actual harm experienced by victims is very real.Survivors have little recourse for healing or justice, and there are few pathways to hold perpetrators accountable and stop them from doing this to more people.Survivor leaders within The Reclaim Coalitions lived experience expert community consistently describe how they have experienced thoughts of suicide, financial effects due to trying to pay out-of-pocket for image removal services, lost relationships, dropping out of school due to shame and trauma, post-traumatic stress and paranoia, body dysmorphia, and inabilities to trust and create new relationships. In short, what is called a deepfake has real, lasting damage that can ruin lives. These harms are not related to the public status, age, gender orientation or profession of the victim. From a fear and harm perspective, it does not matter if you are Taylor Swift, Selena Gomez or a 18-year-old girl in Australia who just got into the college of her dreams and her classmates decided to make deepfakes of her and her female classmates. Or, for that matter a 13-year-old girl in California. I had images made of my face edited onto another womans body that was then shared on twitter and posted online. It was extremely violating and as though I was not even a person, just a body to be manipulated for others pleasure and my shame. - Megan Sims, Advocate and Survivor, The Reclaim Coalition The scale of this type of abuse, which disproportionately impacts women and girls, is truly astonishing. Not only has deepfake abuse grown faster in 2023 than all prior years combined, but research found that women represent 99% of those targeted by deepfake pornography." This type of deepfake contentnonconsensual and explicitmakes up 98% of all deepfake videos online. According to research from our partner, My Image My Choice, the top 40 websites created to house deepfake abuse content now host a cumulative 270,000 videos with more than 4 billion views. This represents a 3000% increase from prior years, which also means that an alarming volume of women are now being depicted in deepfakes without their consent. The harms are profound and far-reaching. Beyond the personal trauma of online sexual violence, this type of abuse has a silencing effect that leads to people stepping back from vital arenas like politics, journalism, and public discourse. Nearly 9 in 10 women restrict their online activity due to, or in anticipation of, online harassment and online sexual violence.   Digital content can spread across multiple platforms and countries making it difficult to remove or track and protective laws in one country would not be enough to protect all victims across the world.

National responses and mechanisms have to be supported by strong interconnected, international responses. The technology sector also needs to play a role by not promoting this content on their platforms and removing it, especially when it has been reported as offending content. Image-Removal of AI generated nonconsensual sexual violence content must be done with effective regulation: While removal is necessary and vital for individual victims, at a macro-level it remains a superficial response that fails to tackle the root causes of this problem. All removal does is seek to take offending content down at any one time in any one place. But more importantly, removal will only be effective if it is done simultaneously with strong enforcement. Removal without enforcement does little to meaningfully regulate this scourge. Both must occur. Specific Ways to Detect and Remove Nonconsensual Explicit AI Images: The challenges are multifaceted. Online IBSV and deepfake abuse are new iterations of gender-based violence. At its core, this is a societal issue, and we must address the root causes of misogyny. Thats a huge undertaking that will take awareness, education, and policy shifts. Without legislation in place, tech companies that enable this abuseand often profit off itare not incentivized to prevent and take down abusive content. The lack of legislation means that perpetrators can create and distribute abusive deepfake content with zero sense of accountability. This has now become a monetized and profitable industry. There are striking similarities to online sex trafficking. There are creators of nonconsensual deepfake content making over $20k a month and hiring assistants to manage the demand for their nonconsensual content. We need legislation, technology company accountability, and robust services for survivors, including hotlines, mental health support, and image removal services. There is a clear role within this paradigm for Meta both in Facebook and Instagram to play a leading role. We need an entire paradigm shift with respect to how law and public policy treat peoples digital identities, images, videos, audio, and texts. Technology companies developing and deploying this technology should not be permitted to expropriate any material from, about, or of, people without their express and informed consent. This should be the gold standard. -Noelle Martin, (Australian Human Rights Attorney, Lived Experience Expert, Researcher at the UWA Tech & Policy Lab Metadata Tracking: Adult survivors of synthetic NCII need access to effective and swift identification of their abuse content as well as removal of existing and future content. These robust systems should be able to quickly and with accuracy identify photographic and video-based content. A key element to succeeding in this goal will be ensuring that there is a comprehensive metadata system that tracks the origins and the nature of the synthetic or AI-generated content.   These systems should support platforms in their enforcement of harm and content moderation per their overall user agreements. This includes pathways to work with not only survivors but law enforcement who may require the data to identify perpetrators in possible criminal investigations. Part of this should include mandatory policies to use both hashing and AI technology to verify the authenticity of possible intimate or misinformation digital content before it is shared online, thus reducing the increase of URLs with the abuse content and limiting the burden and cost to online platforms. Furthermore, when this explicit sexual material is uploaded, there must be identification permissions that the platforms or social networks must have so as not to harm third parties. Labeling Requirements: There must be regulations requiring all AI-generated content to be labeled. This could be visible, such as a watermark. It could also be invisible, such as immovable fragments of code that computer programs can flag if/when needed. Consent Requirements There must be regulations requiring tech companies to ask for consent from users before creating deepfakes using their images. This is important for all content, not just explicit content. Whenever an algorithm is given content, it becomes better at recreating that content. For example, Taylor Swift's fans uploaded a bunch of non-explicit photos of her to deepfake sites in order to create fake album covers. Those sites were then far better at creating realistic-looking pornography of her because they had been trained using so many of her photos in multiple iterations. User Rights: Technology companies need to have easily accessible, clear and intuitive reporting pipelines so victims or end users can report instances of CSAM and/or image-based abused. These pipelines should be standardized across platforms to support removal. Advertising: It should be illegal for companies to advertise their deepfake/AI services with the mention or implication of pornography. Content Moderation and Trust and Safety: Last year, large tech platforms were laying off their content moderators staff who review image removal requests by survivors and their advocates. All technology platforms could learn from survivors that when they fail to remove this content, they are not only likely violating their own user agreements but they are further harming survivors some as young as 11 years old. There is a large and growing network of Trust and Safety professional networks that have dedicated leadership who, together with lived experience experts such as those contributing to this report, can advocate and advise for the best practices of content moderation. Platform Responsibility: Tech companies should be accountable for hosting and distributing nonconsensual deepfake porn. This could involve regulations requiring platforms to implement effective content moderation policies and tools to detect and remove deepfake porn. Technology companies need to be held accountable for responsiveness to and timely action of end user and victim reports of the distribution of image-based abuse via AI- generation, CSAM or non-consensual sharing of intimate images. Technology platforms need to be held account for the blocking of future uploads of or de-indexing of known and previously reported CSAM and image-based abuse imagery (real or AI-generated). Conclusion: Digital content can spread across multiple platforms and countries making it difficult to remove or track and protective laws in one country would not be enough to protect all victims across the world. National responses and mechanisms have to be supported by strong interconnected, international responses. The technology sector also needs to play a role by not promoting this content on their platforms and removing it, especially when it has been reported as offending content. When we listen and invest in recommendations by survivors of lived experience with AI generated sexual violence harms, we can prevent more victims in the future.

Link to Attachment

PC-27099

2024-007-IG-UA, 2024-008-FB-UA	PC-27100	United States & Canada
Case number	Public comment number	Region
Omny	Miranda Martone	English
Commenter's first name	Commenter's last name	Commenter's preferred language
Sexual Violence Prevention Association (SVPA)		Yes
Organization		Response on behalf of organization

Full Comment

DID NOT PROVIDE

Link to Attachment

PC-27100