C.A.S.C.A

COMMENTS TO META OVERSIGHT BOARD

Case on Scams Features AI Video of Brazilian Soccer Legend Ronaldo Seeming to Endorse Online Game

February, 2025

Comments to Meta Oversight Board

Case on Scams Features AI Video of Brazilian Soccer Legend Ronaldo Seeming to Endorse Online Game

Authors: Tanmay Durani, Inika Dular, Kavya Mittal, Viraj Singh

Research Consultant: Dr. Ivneet Walia, Associate Professor of Law and Officiating Registrar, RGNUL

CENTRE FOR ADVANCED STUDIES IN CYBER LAW AND ARTIFICIAL INTELLIGENCE [CASCA] is a research-driven centre at RGNUL dedicated to advancing scholarly research and discourse in the field of Technology Law and Regulation. As a research centre of a leading institution in India, we are committed to promoting interdisciplinary research, fostering collaboration, and driving innovation in the fields of cyber law, artificial intelligence, and other allied areas.

For more information

Visit **cascargnul.com**

Disclaimer

OVERSIGHT BOARD'S CALL FOR PUBLIC COMMENTS

(Released February 6, 2025)

An AI-manipulates video surfaced on Facebook in September 2024, depicting Brazilian soccer player Ronaldo Nazário endorsing the Plinco app, an online game. The video featured Ronaldo, with audio imitating his voice, falsely claiming that playing Plinco could yield higher earnings than typical jobs in Brazil. The post encouraged users to download the game using a link provided in the post itself, which, however, redirected users to a different game called Bubble Shooter. Despite being reported by a user as fraudulent, the content remained accessible and garnered approximately 600,000 views before Meta removed it for violating its policies on fraud and spam.

The Oversight Board had taken up this case to scrutinise the enforcement of Meta's policies concerning scams and deceptive practices involving AI-manipulated media and the impersonation of public figures. Public comments were solicited to provide insights on:

(i) The socioeconomic impact of endorsements by deepfakes imitating public figures on the public and the figure being imitated, especially in Brazil.

(ii) The effectiveness of Meta's enforcement practices for its policies against scams, specifically for content that contains fake personas and the impersonation of public figures, in Brazil and other regions.

(iii) How Meta's announcement on January 7, 2025, about ending proactive enforcement of certain policies could impact the amount of deepfake endorsements on Meta's platforms, particularly in regions where user reports are less frequent.

The comments contributed by CASCA strive to help Meta strengthen its approach to combating deepfake scams, uphold user trust, and ensure that its platform does not become a breeding ground for misinformation and financial exploitation. The original Call for Public Comments can be accessed here.

COMMENTS TO OVERSIGHT BOARD

I. Socio-Economic Impact of Deep-fakes

The AI-manipulated video exploits Brazil's economic vulnerabilities by falsely claiming that Plinco offers higher earnings than traditional professions like teaching or bus driving (average salaries: ~BRL 2,500/month). This preys on low-income populations seeking financial stability, exacerbating inequality and distrust in public institutions. Studies indicate that Meta's platforms are a breeding ground for such scams, with 79.3% of financial frauds in Brazil originating on Facebook, Instagram, or WhatsApp.[1] Research shows that Brazilian consumers tend to form deep emotional connections with celebrities, which consequently shape their consumer habits and purchasing decisions.[2] This cultural dynamism makes Brazil more susceptible to deepfake endorsement fraud, where wrongdoers exploit the trust people place in celebrities and their associated products. Furthermore, since this technology can convincingly alter a person's identity in media files, the average consumer may struggle to distinguish between authentic and fraudulent endorsements.[3]

In Brazil, the problem augments on Meta's platforms, where malicious advertisers exploit marketing tools to target audiences on lines of demography and geography. The economic fallout of deepfake endorsements is witnessed through the exploitation of government welfare programs. They offer access to both genuine and made-up government programs, with 40.5% of advertisements impersonating the federal government.[4] This creates confusion among the masses and undermines their trust in legitimate government financial inclusion programs. Thus, this can have a broader socioeconomic impact by reducing the effectiveness of Brazil's actual social mobility programs. Brazil is hyperconnected, with 118 mobile phones per 100 individuals and a fifth-largest social media market worldwide.[5] However, this connectivity also brings the

---

[1] "Meta Networks Are a Breeding Ground for Online Scams, Study Shows" (*Agência Brasil*, February 8,2025)<https://agenciabrasil.ebc.com.br/en/geral/noticia/2025-02/meta-networks-facilitate-online-scams-study-shows> accessed February 19,2025.

[2] Ricardo Boeing and Caroline Schurhaus, "The Effect of Celebrity Endorsement on Brazilian Consumer Behavior: Does It Really Matter?" (2014) 7 International Business Research <https://ccsenet.org/journal/index.php/ibr/article/view/34410> accessed February 19,2025.

[3] "What Are AI Scams and How Do You Stop Them? | Fraud Prevention" (*Sift*) <https://sift.com/blog/what-are-ai-scams-and-how-do-you-stop-them> accessed February 19, 2025.

[4] "Meta Networks Are a Breeding Ground for Online Scams, Study Shows" (*Agência Brasil*, February 8, 2025) <https://agenciabrasil.ebc.com.br/en/geral/noticia/2025-02/meta-networks-facilitate-online-scams-study-shows> accessed February 19,2025.

[5] "Topic: Social Media Usage in Brazil" (*Statista*, February 11, 2025) <https://www.statista.com/topics/6949/social-media-usage-in-brazil/#topicOverview> accessed February 19,2025.

challenge of information poverty, where access to reliable information becomes increasingly scarce.[6] Deepfake media circulating in various formats, including clips, images, and endorsements on Meta platforms, had instances of synthetic content, compromising electoral integrity. Brazil strived to tap this crisis through strict legislative action in October 2024 Elections. After the AI-manipulated video of a municipal candidate campaigning for elections surfaced, Brazil implemented rules banning unlabeled AI-generated content in electoral campaigns. [7] Thus, the combination of Brazil's celebrity-influenced consumer culture, widespread social media use, and the growing precision of deepfake technology to impersonate the federal government augments the threat to both economic security and social stability in the country.

II. Effectiveness of Meta's Enforcement Practices

Over the last decade, Meta has adopted multiple enforcement practices to protect its users from harmful content. However, with cases like the one at hand, the effectiveness of these enforcement practices seems far from reality. Meta's weak enforcement practices have often been called out, even by its own Oversight Board. The cases of digital fraud are increasing in Brazil rapidly, according to a FICO study.[8] In another Witness Media Lab study, Meta-owned Facebook was considered one of the prominent platforms for deepfake circulation in Brazil,[9] making it more important for Meta to practice caution and diligence in enforcing its community standards.

To strengthen its enforcement practices, Meta has taken several steps, like incorporating a new AI system in 2021 to label harmful content, which its traditional method might have overlooked.[10] The company relies heavily on technology, with over 90% of objectionable posts being flagged by it before any of Meta's

---

[6] Walter Matli, "The Application of Information Poverty Theory to Health Information and Misinformation in the Digital Age" [2024] Research Square (Research Square) <https://www.researchsquare.com/article/rs-4651238/latest>accessed February 19,2025.

[7] Acarvin, "Brazil's Electoral Deepfake Law Tested as AI-Generated Content Targeted Local Elections - DFRLab" (*DFRLab*, December 2, 2024) <https://dfrlab.org/2024/11/26/brazil-election-ai-deepfakes/> accessed February 19,2025.

[8] 'Identity theft for digital fraud is on the rise in Brazil, says FICO' (*FICO, 22 March 2024*) <https://www.fico.com/en/newsroom/identity-theft-digital-fraud-rise-brazil-says-fico> accessed 19 February 2025

[9] The WITNESS Media Lab, *DEEPFAKES: PREPARE NOW (Perspectives from Brazil)* (July 2019) <https://lab.witness.org/wp-content/uploads/sites/29/2019/10/WITNESS-Deepfakes-Brazil-Prepare-Now-Updated.pdf> accessed 19 February 2025

[10] 'Our New AI System to Help Tackle Harmful Content' (*Meta,* 8 December 2021) <https://about.fb.com/news/2021/12/metas-new-ai-system-tackles-harmful-content/> accessed 19 February 2025

users.[11] Its latest, and one of the biggest developments in a while, has been the adoption of a Community Notes model, making its third-party fact-checking program redundant in the US in light of its strong support for freedom of speech-[12] in compliance with Article 19 of the Universal Declaration of Human Rights.[13]

Despite these active efforts taken by Meta to strengthen its enforcement practices, the effectiveness of these practices can be questioned. With Meta ending its fact-checking program and relying on user reporting, the regions with fewer users flagging objectionable posts may see more scam posts circulating. This is even worse in the context of Brazil, where the projected revenue of the deepfake industry is projected to rise to US$ 330.1 million by 2030.[14] The company trusts its technology for most of its labelling, however, these technical tools may not be as effective as humans, and can have a poles-apart approach if designed too aggressively or too modestly.[15] This usually happens because humans are better able to identify minute detail differences in pictures or videos, however, more advanced detectors are still in their learning stage. Moreover, AI tools to detect deepfake videos often fail as they overlook the minute irregularities in the mouth region in a lip-sync video. Even in the present case, it was a user who reported the AI-generated video and not Meta's enforcement tools, signifying how the company's dependency on technology to keep fraudulent activities in check may not be sufficient. Recent cases of failures of Meta to curb the circulation of deepfake content on its apps are also seen in the instance where deepfake scam ads of businesspersons Dick Smith and Treasurer Jim Chalmers surfaced in Australia, and the company took down the flagged videos only following a government condemnation.[16] What is called the internet's biggest scam is another deepfake video of Elon Musk circulating on Facebook leading to huge losses to viewers, indicating how gullible an average viewer is to deepfake videos.[17] Similarly in the 2024 case of deepfake videos of

---

[11] 'How technology detects violations' (*Meta*, 18 October 2023) <https://transparency.meta.com/en-gb/enforcement/detecting-violations/technology-detects-violations/> accessed 19 February 2025

[12] Joel Kaplan, 'More Speech and Fewer Mistakes' (*Meta,* 7 January 2025) <https://about.fb.com/news/2025/01/meta-more-speech-fewer-mistakes/> accessed 19 February 2025

[13] Universal Declaration of Human Rights (adopted 10 December 1948 UNGA Res 217 A(III) (UDHR) art 19

[14] 'Brazil Deepfake Ai Market Size & Outlook, 2024-2030' (*Horizon*) <https://www.grandviewresearch.com/horizon/outlook/deepfake-ai-market/brazil> accessed 19 February 2025

[15] 'How enforcement technology works' (*Meta,* 12 November 2024) <https://transparency.meta.com/en-gb/enforcement/detecting-violations/how-enforcement-technology-works/> accessed 19 February 2025

[16] David Murray, 'Deep fake video scam using entrepreneur Dick Smith and Treasurer Jim Chalmers targets Australians' (*Gina Rinehart*, 27 November 2023) <https://www.ginarinehart.com.au/deep-fake-video-scam-using-entrepreneur-dick-smith-and-treasurer-jim-chalmers-targets-australians/> accessed 19 February 2025

[17] Brian New, 'Deepfakes of Elon Musk are contributing to billions of dollars in fraud losses in the U.S.' (*CBS News*, 24 November 2024) <https://www.cbsnews.com/texas/news/deepfakes-ai-fraud-elon-musk/> accessed 19 February 2025

celebrities urged people to invest in a fake opportunity, causing major losses to Australians[18] which made the company tighten its regulations upon further government condemnation.[19] The general infectivity of Meta's enforcement practices has also been underscored in its Oversight Board's multiple decisions like the *Explicit AI Images of Female Public Figures* case where the challenges of enforcing Meta's community standard, and the company's decision not to take it down "*was in error.*"[20] Conversely, the Board has also criticised Meta's "*overenforcement of Meta's Bullying and Harassment policy*" in another case of *Derogatory Image of Candidates for U.S. Elections*.[21]

Addressing human rights concerns, AI regulations of regions like the UK and GCC have regulated AI-modified content.[22] While Meta's investments[23] in technological advancements signal their commitment towards content moderation, their real-world enforcement seems questionable in light of these contradictions and inconsistencies- oscillating between under and over-enforcement. The need for its Oversight Board to intervene and overturn Meta's posts-related decisions reflects a fundamental gap between policy formulation and implementation.

III. Impact of Ending Proactive Enforcement

Meta's change in the way they enforce content moderation is a significant step that will impact both users and advertisers. Previously, machine(s) were used to automate the decision-making process when a piece of content was considered violative of the platform's community guidelines. Moreover, the content (if found against these policies) was removed before it could be available to users at large. This proactive enforcement did attract concerns about the leaning of the organisation, but it was considered a far better

---

[18] Hanan Dervisevic, 'Meta responds to deepfake celebrity scam pressure by announcing new rules for financial advertisers' (*ABC News*, 2 December 2024) <https://www.abc.net.au/news/2024-12-02/meta-forces-financial-advertisers-verify-information-scams/104672484> accessed 19 February 2025

[19] 'About authorisation requirements that may affect financial services advertisers' (*Meta*) <https://en-gb.facebook.com/business/help/719892839342050> accessed 20 February 2025

[20] Explicit AI Images of Female Public Figures (2024-007-IG-UA, 2024-008-FB-UA, Meta Oversight Board,) https://www.oversightboard.com/decision/bun-7e941o1n/

[21] https://www.oversightboard.com/decision/fb-fwixegxq/

[22] 'AI-generated deepfakes: what does the law say?' (*Rouse*, 4 September 2024) <https://rouse.com/insights/news/2024/ai-generated-deepfakes-what-does-the-law-say> accessed 20 February 2025 https://rouse.com/insights/news/2024/ai-generated-deepfakes-what-does-the-law-say

[23] 'How Meta invests in technology' (*Meta*, 19 January 2022) <https://transparency.meta.com/en-gb/enforcement/detecting-violations/investing-in-technology/> accessed 20 February 2025

approach to content moderation in comparison to the reactive approach. However, a notification dated 7th January 2025[24] from Meta changed the proactive enforcement approach to a more reactive one.

Shifting the Onus

The new approach takes inspiration from X's 'community notes' wherein the action upon a piece of content is taken after receiving feedback/flagging reports from users. The users are empowered to add context and relevant details to posts that they find contentious. The new system of content moderation shifts the onus from fact-checking organisations to individual users. A piece of content that may violate community guidelines may only be acted upon if it receives a threshold of complaints from the users. It will be incorrect to say that the new approach completely omits the use of automated systems[25]. However, the usage of automated systems would come in the picture when a post is flagged for a 'high-severity violation' (drugs, terrorism, self-harm, suicide, etc). This approach is particularly problematic because of three reasons: (a) it limits automatic content moderation to only illegal content or high-severity information. This leaves room for misinformation that might not qualify as illegal and/or of high severity but continues to mislead users until it is reported by the users themselves. (b) This approach will only leave domestic laws as a remedy to remove content that may not be suitable to be put up on the platform and does not qualify as illegal. A spokesperson from the Department of Science, Information and Technology of UK remarked Meta's obligation to follow UK's Online Safety Act regardless of the in-house content moderation practices.[26]However, many domestic technology laws only include the removal of illegal content and are lax on misinformation. (c) Lastly, it will also impact the behavior of users of the platform. It will become less of an 'option' and more of a 'responsibility' for them to keep reporting such content from time to time. Another concern is how the rapid increase of deepfakes can beat this system. For instance, a post on this platform that uses a celebrity deepfake to promote a fraudulent application will first attract a threshold of community notes from volunteers. Then, it will be analyzed at the backend before it is removed. The delay in removing such a post will only cause more harm. Throughout the time the post will be open for people

[24]Joel Kaplan 'More Speech and Fewer Mistakes' (Meta, January 2025) https://about.fb.com/news/2025/01/meta-more-speech-fewer-mistakes/ accessed 20 February 2025.

[25] Robert Booth 'Ditching Facebook factcheckers major step back public discourse' The Guardian (London, 7 January 2025) https://www.theguardian.com/technology/2025/jan/07/ditching-facebook-factcheckers-major-step-back-public-discourse accessed 20 February 2025

[26] Sheera Frenkel 'Meta's Mark Zuckerberg Meets with President-elect Trump at Mar-a-Lago' *The New York Times* (New York, 10 January 2025) https://www.nytimes.com/2025/01/10/technology/meta-mark-zuckerberg-trump.html accessed 20 February 2025.

to see, it will pose a risk, especially to people with lower literacy levels and citizens of countries with weak digital laws, to be scammed.

Lack of Uniformity for Advertisers

While Meta has stated that the provision of 'community notes' will not apply to paid advertisements, advertisers find themselves in a state of flux. This is because while this does not apply to paid ads, it applies to organic posts. This is pertinent to note because of two reasons: (i) it will lead to the disappearance of organic content since there will always be a risk of content moderation (which is now unpredictable due to its user-centric approach). (ii) It will also push for only paid ads because they are not subject to the new form of content moderation. While it will discourage smaller brands with less budgets to be able to advertise with ease, it will only increase the profits of Meta. From a larger perspective, this will also discourage all the SaaS companies that offered to 'vet' the advertisements to be run on Meta as organic sponsored posts. These companies used previous datasets and filters to give a green light to the ads, but with the increased volatility of user-based reactive enforcement, they will take a blow.[27]

Paving the Way for Advertisers

Advertisers need to be wary of the internal and external developments to the platform. There is little clarity over the monetisation eligibility status of organic content marred by community notes. Moreover, with the landscape predicted to become increasingly filled with paid ads, there will be changes in crucial indicators such as Cost Per Click (CPC) of the ads. Moreover, with the volatile regulatory landscape of Meta coupled with the variation in domestic laws, advertisers are now expected to put aside a significant amount for brand strategy. Three important suggestions can be used: (a) Ensuring branding uniformity by content creation that is appropriate across jurisdictions and the new framework, (b) establishing and vocalising brand values in order to circumvent backlash from community notes, and (c) internal suitability strategy to articulate the kinds of content your brand seeks to align with and avoid.[28]

---

[27] 'Threads of Wisdom: Experts React to Meta's Policy Changes' (Institute for Rebooting Social Media, 17 January 2025) https://rebootingsocialmedia.org/2025/01/17/threads-of-wisdom-experts-react-to-metas-policy-changes/ accessed 20 February 2025.

[28] AJ Brown, 'A Shifting Onus: What Meta's Content Moderation Changes Mean for Advertisers' (Brand Safety Institute, 9 January 2025) https://www.brandsafetyinstitute.com/blog/meta-content-moderation-changes accessed 20 February 2025.

IV. Recommendations

To effectively address the multifaceted challenges highlighted in the case of the AI-manipulated Ronaldo video and to strengthen Meta's approach to deepfake scams, the Oversight Board should consider the following recommendations for Meta:

Firstly, **Meta must re-evaluate and reinforce its commitment to proactive detection and removal of harmful deepfake content, especially those related to scams and impersonation,** and particularly in regions identified as highly vulnerable like Brazil. The shift towards a reactive enforcement model, while potentially empowering users, carries significant risks, especially when dealing with sophisticated and rapidly spreading deepfakes. Relying primarily on user reports and community notes for content like deepfake scams that exploit economic vulnerabilities and manipulate public figures places an undue burden on users and can result in delayed action, causing substantial harm in the interim. Meta should reinvest in and enhance its proactive AI-driven detection technologies specifically tailored to identify deepfakes, focusing on visual and auditory anomalies, contextual analysis, and known scam tactics. This proactive approach should be prioritized for content categories with high potential for harm, such as financial scams, and should be geographically targeted towards regions known to be disproportionately affected by such activities, ensuring that vulnerable populations receive adequate protection.

Secondly, to complement technological advancements, **Meta should significantly enhance the training and resources allocated to its human content reviewers, especially those focused on identifying nuanced forms of manipulation like deepfake scams.** While AI detection tools are crucial, they are not infallible and often struggle with the subtle nuances of deepfake technology and the cultural context within which scams operate. Human reviewers, particularly those with regional expertise and language proficiency, are essential for identifying complex cases and making informed judgments. Meta should invest in specialized training modules for reviewers that focus on deepfake identification techniques, common scam patterns prevalent in different regions (including Brazil), and the socioeconomic contexts that make certain populations more vulnerable.

Thirdly, Meta needs to develop and implement regionally tailored enforcement strategies that account for the specific socioeconomic and cultural contexts of different regions, particularly those with high vulnerability to deepfake scams like Brazil. A one-size-fits-all enforcement approach is inadequate given the diverse global landscape in which Meta operates. In regions like Brazil, where celebrity endorsements hold significant sway and economic vulnerabilities are prevalent, a more proactive and culturally sensitive

enforcement strategy is necessary. **The regional strategy must include strengthened advertiser verification processes specifically designed to prevent the impersonation of public figures and institutions, which is demonstrably prevalent in fraudulent advertising within Brazil.** This includes enhanced identity checks, cross-referencing advertiser information against databases of public figures and institutions, and potentially incorporating AI-driven analysis to detect inconsistencies or red flags in advertiser profiles and content. **This regional approach should also encompass localized public awareness campaigns in Portuguese to educate users about deepfakes and online scams, partnerships with Brazilian fact-checking organizations, and the development of reporting mechanisms that are easily accessible and culturally appropriate for Brazilian users.** Furthermore, Meta should consider allocating additional enforcement resources to regions identified as high-risk for deepfake scams, ensuring that enforcement efforts are commensurate with the scale and potential impact of these threats.

Fourthly, Meta needs to implement stricter Targeting Restrictions for gambling-related and other potentially predatory advertising content. The current demographic targeting tools, while useful for legitimate advertising, can be exploited to disproportionately target vulnerable populations with harmful content like gambling scams. **Meta should consider limiting the granularity of demographic targeting available for gambling and similar categories**, restricting the use of highly specific and potentially discriminatory targeting criteria that could be used to prey on individuals based on socioeconomic status, age, or other sensitive characteristics.

Fifthly, to foster greater transparency and accountability, Meta should enhance its communication with users regarding content moderation decisions, particularly in cases involving reported scams and deepfakes. Users who report content should receive clear and timely updates on the status of their reports and, when content is removed or remains, a clear explanation of the reasoning based on Meta's community standards. Moreover, in the context of the shift towards reactive enforcement, Meta should clearly communicate to its users and advertisers the implications of this change, especially regarding the handling of organic and paid content and the potential impact on content visibility and reach.