AI-Manipulated Video Promoting Gambling

2025-018-FB-UA

Summary

The Board has overturned Meta's decision to leave up a Facebook post showing an AI-manipulated video of a person who appears to be Brazilian soccer legend Ronaldo Nazário endorsing an online game. Taking the post down is consistent with Meta's Community Standards on fraud and spam. Meta should also have rejected the content for advertisement, as its rules prohibit using the image of a famous person to bait people into engaging with an ad.

Based on public reporting, the Board notes Meta is likely allowing significant amounts of scam content on its platforms to avoid potentially overenforcing a small subset of genuine celebrity endorsements. At-scale reviewers are not empowered to enforce this prohibition on content that establishes a fake persona or pretends to be a famous person in order to scam or defraud. Meta should enforce this prohibition at-scale by providing reviewers with often easily identifiable indicators that distinguish AI content.

About the Case

In September 2024, a user shared a post involving an AI-manipulated video of a person who appears to be retired Brazilian soccer player Ronaldo Nazário. In the video, he encourages people to download an app to play the popular online game Plinko (or Plinco).

In the video, the audio imitating Ronaldo Nazário is not in sync with his lip movements. The video also shows unrealistic AI-generated images of a schoolteacher, a bus driver and a grocery store worker, and the average salary for their jobs in Brazil. The audio claims that players on Plinko can earn more money from the game than the jobs mentioned. The video encourages users to click a download link to the app, although

this leads to a different game called Bubble Shooter. The post was viewed over 600,000 times.

A user reported the content to Meta as a fraud or scam, but the report was not prioritized. The company did not remove the content. The user appealed this decision to Meta, but this appeal was not prioritized for human review either, so the content remained on Facebook. Finally, the user appealed Meta's decision to the Board, saying the post appeared to be sponsored. If a post is boosted, an ad is created on the post.

The ad was disabled for violating the company's Unacceptable Business Practices Advertising Standard, although the original organic post remained on the platform. After the Board identified this case for review, Meta removed the original post for violating the Fraud, Scams and Deceptive Practices policy. Meta later confirmed the post also violated its Spam policy.

Deepfakes and deepfake endorsements are increasing globally, including those involving public figures promoting fraudulent political campaigns and financial scams. Reports highlight that many of the financial scams in Brazil that originate on Facebook, Instagram and WhatsApp involve AI-manipulated content.

Key Findings

Removing the content is consistent with Meta's human rights responsibilities. Misleading manipulated endorsements pose significant risks to the depicted person's rights to privacy and reputation. They also impact the public, by potentially facilitating fraud.

The Board is concerned that at-scale content reviewers are unable to remove posts that establish a fake persona or pretend to be a famous person "in an attempt to scam or defraud," even if the content contains clear indicators that it violates Meta's policies. Such content can only be removed by Meta's specialized teams, making underenforcement of its Fraud, Scams and Deceptive Practices policy more likely.

Meta is likely allowing significant amounts of scam content on its platforms to avoid potentially overenforcing a small subset of genuine celebrity endorsements. This is particularly concerning when genuine celebrity endorsements will likely have other protections against overenforcement, either through formal systems such as cross-check or points of contact at Meta. The Board therefore recommends Meta change its approach and enforce this policy line at-scale.

The manipulated or fake nature of the video is apparent. The Board finds the post violates Meta's prohibition on fake personas or pretending to be a famous person to scam or defraud, under the Fraud, Scams and Deceptive Practices Community Standard. It also violates Meta's prohibition on sharing deceptive or misleading links under its Spam Community Standard, as it promotes Plinko but links to a different game. Therefore, the Board finds that the post should have been removed when reported. Even before the post's removal, Meta should have applied an "AI info" label, under its Manipulated Media policy. Meta should also have rejected the content for advertisement, as its Unacceptable Business Practices Advertising Standard prohibits using the image of a famous person and misleading tactics to bait people into engaging with an ad.

Meta has a responsibility to "mitigate adverse human rights impacts" of monetized content that could scam or defraud – in line with the United Nations Guiding Principles on Business and Human Rights. When paid to boost content, Meta should ensure these posts do not violate its policies.

The Oversight Board's Decision

The Oversight Board overturns Meta's decision to leave up the post on Facebook.

The Board also recommends that Meta:

- Enforce at scale its Fraud, Scams and Deceptive Practices policy prohibition on content that "attempts to establish a fake persona or to pretend to be a famous

person in an attempt to scam or defraud" by providing reviewers with indicators to identify this content. This could include, for example, the presence of media manipulation watermarks and metadata, or clear factors such as video-audio mismatch.

* Case summaries provide an overview of cases and do not have precedential value.

Full Case Decision

1. Case Description and Background

In September 2024, a user posted an AI-manipulated video of a person who appears to be retired Brazilian soccer legend Ronaldo Nazário encouraging others to download an app to play Plinko. A popular online game, Plinko (or Plinco) involves dropping a ball down a peg-filled board, with players winning different prizes based on where the ball lands.

The video begins with Ronaldo Nazário speaking to the camera. While it appears realistic at first glance, the audio imitating the soccer star is not in sync with his lip movements. The video then shows unrealistic AI-generated images of a schoolteacher, a bus driver and a grocery store worker, as well as the average salary for these jobs in Brazil. The audio imitating Ronaldo Nazário's voice claims that Plinko is simple to play, and that average players can earn more money from the game than from these jobs. Finally, the video encourages users to click a link to download the app, although this leads to a different game called Bubble Shooter. The post was viewed over 600,000 times and reported over 50 times by different users.

A user reported the content to Meta as a fraud or scam, but the report was not prioritized for human review and the company did not remove the content. The user then appealed this decision to Meta. This appeal was not prioritized for human review either, so the content remained on the platform. The user finally appealed Meta's decision to the Board.

The user who appealed Meta's decision to the Board said the content appeared as a sponsored post. Meta allows users to pay to "boost" posts to increase their visibility and reach wider audiences. When a post is boosted, an ad is created based on the post. Separately, Meta told the Board that the ad was disabled for violating the company's policy. This means the post was no longer boosted for additional visibility, but the original organic post remained on the platform. After the Board identified this case for review, Meta removed the original post, finding it violated the Fraud, Scams and Deceptive Practices policy. The company also applied a standard strike on the profile of the user who created the post. Meta later confirmed that the post also violated its Spam policy.

The Board noted the following context in reaching its decision on this case:

The problem of deepfakes and deepfake endorsements is increasing around the world. Deepfakes are a prominent social and political issue in Brazil, often involving influential public figures. For example, AI-manipulated disinformation, including fake candidate endorsements, has featured in recent electoral campaigns. In 2024, the Brazilian Ministry of Sports raised concerns about social media content promoting online gambling and promising easy money without warning of the risks involved. Reports also highlight that many of the financial scams in Brazil that originate from Facebook, Instagram and WhatsApp involve AI-manipulated content (see public comment PC-31027, Centre for Advanced Studies in Cyber Law and Artificial Intelligence).

2. User Submissions

The user who reported the content called the video a lie and a scam for using Ronaldo Nazário's image to induce people to download and play a game. The user stated that Meta puts warnings and labels on other posts, but did not place any warning in this case or delete the post. They stated the content appeared as a sponsored post and is visibly false.

3. Meta's Content Policies and Submissions

I. Meta's Content Policies

Fraud, Scams and Deceptive Practices Community Standard

The Fraud, Scams and Deceptive Practices policy rationale states that Meta "aim[s] to protect users and businesses from being deceived out of their money, property or personal information" by removing content that "purposefully employs deceptive means – such as willful misrepresentation, stolen information and exaggerated claims – to either scam or defraud users and businesses, or to drive engagement."

Under the section of the policy that Meta requires "additional information and/or context to enforce," it states the company "may remove content" that "attempts to establish a fake persona or [that] pretend[s] to be a famous person in an attempt to scam or defraud." This means only specialized review teams at Meta can enforce this rule and it cannot be enforced by at-scale reviewers.

Spam Community Standard

The Spam policy rationale explains that Meta does not allow "content that is designed to deceive, mislead or overwhelm users in order to artificially increase viewership." The rules prohibit content containing misleading links, defined as "content containing a link that promises one type of content but delivers something substantially different."

Misinformation Community Standard

The Misinformation Community Standard states that Meta removes content only "where it is likely to directly contribute to the risk of imminent physical harm" or "directly contribute to interference with the functioning of political processes." The company also "require[s] people to disclose, using its AI disclosure tool, whenever they

post organic content with photorealistic video or realistic-sounding audio that was digitally created or altered."

If manipulated media does not otherwise violate the Community Standards, Meta may "place an informative label on the face of content – or reject content submitted as an advertisement – when the content is a photorealistic image or video, or realistic sounding audio, that was digitally created or altered and creates a particularly high risk of materially deceiving the public on a matter of public importance."

Unacceptable Business Practices Advertising Standard

The Standard explains that advertisements "must not promote products, services, schemes or offers using identified deceptive or misleading practices." In its guidelines for advertisements, Meta prohibits the "use [of] the image of a famous person and misleading tactics in order to bait people into engaging with an ad."

II. Meta's Submissions

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 its Fraud, Scams and Deceptive Practices Community Standard. Meta stated that the post violated its prohibition on content that "attempts to establish a fake persona or to pretend to be a famous person in an attempt to scam or defraud." Meta explained that by making it appear through AI as if Ronaldo Nazário is using or promoting an online game, the video attempted to scam people into using a product they might not otherwise download without his endorsement.

In response to the Board's questions on the enforcement of content with fake personas, Meta explained that it enforces the policy only on escalation to ensure the person depicted in the content did not actually endorse the product. Meta stated this requires a highly context-dependent analysis and specialized expertise. Meta stated: At-scale reviewers' interpretation of what constitutes a 'fake persona' could vary across regions and introduce inconsistencies in enforcement."

Meta also found the content violated its Spam Community Standard that prohibits using deceptive links. In this case, the post included a link that led to a different game than Plinko. For Meta, the use of a link to a different game suggested the content was designed to scam or defraud people or to deceptively drive engagement.

Meta informed the Board that removing this content also protected the rights and reputation of other people. It concluded that "authenticity risks outweighed the value of voice and there were no less intrusive means available for limiting this content other than removal."

The Board asked Meta 10 questions, including on its enforcement practices, how it labels manipulated media and how it approaches celeb-bait scams that use depictions of celebrities to drive engagement with organic and paid content. Meta responded to all questions.

In response to the Board's questions on labeling AI-manipulated content, Meta said it did not label the post as AI-generated content prior to its removal. According to Meta, the video "did not include industry standard indicators that it was AI-generated, nor did the user self-disclose."

4. Public Comments

The Board received four public comments that met the terms for submission – one comment each from Latin America and the Caribbean, the Middle East and North Africa,

the United States and Canada, and Central and South Asia. To read public comments submitted with consent to publish, click here.

Submissions covered the following themes: the socio-economic impact of deepfakes, the effectiveness of Meta's enforcement practices, and the impact of Meta's decision to end proactive enforcement for some categories of content.

5. Oversight Board Analysis

The Board analyzed Meta's decision in this case against Meta's content policies, values and human rights responsibilities. The Board also assessed the implications of this case for Meta's broader approach to content governance.

The Board selected this case to examine for the first time the challenges in enforcing Meta's Fraud, Scams and Deceptive Practices and Spam policies. The volume of manipulated media used for scams is expected to increase, particularly with generative-AI advances. This case falls within the Board's Automated Enforcement of Policies and Curation of Content strategic priority.

5.1 Compliance With Meta's Content Policies

I. Content Rules

The Board finds that the post violates Meta's prohibition on "establish[ing] a fake persona or pretend[ing] to be a famous person in an attempt to scam or defraud" under the Fraud, Scams and Deceptive Practices Community Standard. The video makes it appear as though Ronaldo Nazário is encouraging others to download a gambling app from which they can easily earn money. The manipulated or fake nature of the video is apparent, as the audio imitating the soccer star is visibly not in sync with his lip movements. The Board did not find any public reporting suggesting Ronaldo Nazário endorses this game.

The Board also finds that the post violates Meta's prohibition on sharing deceptive or misleading links under its Spam Community Standard, as the content "contain[s] a link that promises one type of content but delivers something substantially different." The video promotes Plinko and encourages users to download the game, but the link included leads to a different game.

II. Enforcement Action

Meta had multiple opportunities to review and remove this content before it reached the Board. Despite more than 600,000 views and over 50 user reports, Meta did not prioritize this post for review when it was first reported or when it was later appealed.

Prior to its removal, it should have been labeled under Meta's Misinformation Community Standard to indicate it contained manipulated media. In the Altered Video of President Biden decision, the Board recommended that Meta label manipulated content to prevent users from being misled about its authenticity. Meta has implemented this recommendation. In this case, the video was digitally altered to mislead users into believing that Ronaldo Nazário, a well-known public figure, was endorsing a gambling app. Under its new approach to AI-generated content, Meta should have applied an "AI info" label to indicate that the content was digitally created or altered.

Meta labels misleading manipulated content by relying on metadata and watermarks. The company told the Board it has "labeled a large volume of content following this approach," but it does not publicly disclose any statistics around the approach's efficacy. For content that does not contain such markers, of which this post is an example, there may be other indicators that it is AI-generated. In this post, the audio and video do not align, and the video includes low-quality AI-generated images. Experts consulted by the Board and public authorities highlight video-audio mismatch as a key indicator that content is AI-manipulated. They cite other indicators including

unnatural facial movements, inconsistencies in lighting and shadows, and the lack of motion continuity and coherence.

The Board is further concerned that at-scale content reviewers are unable to remove posts that establish a fake persona or pretend to be a famous person "in an attempt to scam or defraud," as prohibited by the policy. This is because such content can only be removed by Meta's specialized teams. This approach makes it more likely for Meta's Fraud, Scams and Deceptive Practices policy to be underenforced. While some posts may require specific expertise to understand they contain fake personas, others, including this one, could be enforced at-scale.

Meta should also have rejected the content for advertisement, as its Unacceptable Business Practices Advertising Standard prohibits using the "image of a famous person and misleading tactics in order to bait people into engaging with an ad." Meta has further publicly stated that it evaluates ads with facial recognition technology for celebrity deepfakes. Meta informed the Board that its ad review system, which primarily relies on automation, is "designed to review all ads before they go live." Despite these mechanisms, it appears that Meta nonetheless accepted this content for advertisement. Meta later informed the Board that the ad was disabled because it violated its Unacceptable Business Practices Advertising Standard, but the organic content remained active on the platform. When Meta found the content violated an Advertising Standard and disabled the ad, this did not trigger any additional review for possible Community Standards violations for the underlying organic post, despite clear policy overlap. To address this gap, the company could initiate an organic policy review when content is found to violate an advertising policy.

5.2 Compliance With Meta's Human Rights Responsibilities

The Board finds that removing the content from Facebook is consistent with Meta's human rights responsibilities.

Article 19(2) of the International Covenant on Civil and Political Rights (ICCPR) provides broad protection for expression. This includes "freedom to seek, receive and impart information and ideas of all kinds." The Human Rights Committee lists specific forms of expression included under Article 19 and notes that the right to freedom of expression may include commercial advertising (General Comment No. 34, para. 11).

When restrictions on expression are imposed by a state, they must meet the requirements of legality, legitimate aim, and necessity and proportionality (Article 19, para. 3, ICCPR). These requirements are often referred to as the "three-part test." The Human Rights Committee has noted the applicability of this test in the context of commercial advertising (General Comment No. 34, para. 33). The Board uses this framework to interpret Meta's human rights responsibilities in line with the UN Guiding Principles on Business and Human Rights, which Meta itself has committed to in its Corporate Human Rights Policy. The Board does this both in relation to the individual content decision under review and what this says about Meta's broader approach to content governance. As the UN Special Rapporteur on freedom of expression has stated, although "companies do not have the obligations of Governments, their impact is of a sort that requires them to assess the same kind of questions about protecting their users' right to freedom of expression," (A/74/486, para. 41).

The Board has often noted the importance of protecting political and social discourse (General Comment No. 34, para. 38). Those considerations do not apply here. The Board finds this post to be commercial speech that may be limited where the three-part test is met.

I. Legality (Clarity and Accessibility of the Rules)

The principle of legality requires rules limiting expression to be accessible and clear, formulated with sufficient precision to enable an individual to regulate their conduct accordingly (General Comment No. 34, para. 25). People using Meta's platforms should be able to access and understand the rules and content reviewers should have clear guidance regarding their enforcement.

The Board finds that the rules within Meta's Fraud, Scams and Deceptive Practices policy are sufficiently clear and accessible. It is clear to the Board that the prohibition on "establish[ing] a fake persona or pretend[ing] to be a famous person in an attempt to scam or defraud" encompasses posts that use the likeness of a public figure to fraudulently endorse a product or app.

II. Legitimate Aim

Any state restriction on freedom of expression should also pursue one or more of the legitimate aims listed in the ICCPR, which includes protecting the rights or reputations of others. Meta's prohibition on posts that establish fake personas to scam or defraud others meets two aims. First, it seeks to protect people from scams and fraud (Article 17, Universal Declaration of Human Rights). Second, it protects the rights and reputation of the persons depicted, as this content impacts their right to privacy and their ability to decide how images of themselves are created and released (Article 17, ICPPR; see also Explicit AI Images of Female Public Figures decision).

III. Necessity and Proportionality

Under ICCPR Article 19(3), necessity and proportionality requires that restrictions on expression "must be appropriate to achieve their protective function; they must be the least intrusive instrument amongst those which might achieve their protective function; they must be proportionate to the interest to be protected." In addition, under the United Nations Guiding Principles on Business and Human Rights (UNGPs), a company's actions in response to potential harm should be informed by the extent of its involvement in creating any adverse human rights impact (UNGP 19(b)).

The Board finds that Meta's eventual decision to remove the content from Facebook was necessary and proportionate. In this case, removal is the least intrusive measure to protect the public from scams, particularly those with limited digital literacy, and to protect against the misuse of Ronaldo Nazário's image. The impact on his privacy and

reputation is immediate and there could be financial harm to the public, so Meta should remove such content.

Enforcement

Meta has a responsibility to "mitigate adverse human rights impacts" of monetized content that could scam or defraud – in line with the UNGPs. If Meta is paid to increase the reach of content through its boosting program, the company should take particular care to ensure these posts do not violate its policies.

These endorsements, especially when created by generative-AI tools, may be difficult for viewers to detect. The availability of advanced generative-AI tools to create videos has increased dramatically in recent years and is likely to rise. As mentioned above, reports highlight that some financial scams in Brazil originate from Facebook, Instagram and WhatsApp, including those using deepfakes.

Deepfake Plinko advertisements and organic content do not appear to be a novel issue. As part of its research into this case, the Board searched Meta's Ad Library for "Plinko app" content. At the time of its search, the Board found over 3,900 active advertisements for such content, about 3,500 of which included videos. Of these videos, several featured similar AI-generated endorsements, including deepfakes of Portuguese soccer player Cristiano Ronaldo. In terms of organic content, the Board also found deepfakes featuring Meta's CEO Mark Zuckerberg endorsing Plinko.

The Board is concerned that at-scale content reviewers are unable to remove posts that establish a fake persona or pretend to be a famous person "in an attempt to scam or defraud," even if the content contains clear indicators that it violates Meta's policies. Meta's approach of only enforcing this policy after seeking additional context on escalation favors the underenforcement of violating content. Based on public reporting, including databases of reported incidents and journalistic reports, the Board notes that to avoid potentially overenforcing a small subset of genuine celebrity endorsements, the company is likely allowing significant amounts of scam content on

its platforms. This is particularly concerning when genuine celebrity endorsement content will likely have other protections against overenforcement, either through formal systems such as cross-check or points of contact at Meta. The Board therefore recommends Meta change its approach and enforce this policy line at-scale.

6. The Oversight Board's Decision

The Oversight Board overturns Meta's original decision to leave up the post on Facebook.

7. Recommendations

Enforcement

1. To better combat misleading manipulated celebrity endorsements, Meta should enforce at scale its Fraud, Scams and Deceptive Practices policy prohibition on content that "attempts to establish a fake persona or to pretend to be a famous person in an attempt to scam or defraud" by providing reviewers with indicators to identify this content. This could include, for example, the presence of media manipulation watermarks and metadata, or clear factors such as video-audio mismatch.

The Board will consider this recommendation implemented when both the public-facing and private internal guidelines are updated to reflect this change.

***Procedural Note:**

- The Oversight Board's decisions are made by panels of five Members and approved by a majority vote of the full Board. Board decisions do not necessarily represent the views of all Members.

- Under its Charter, the Oversight Board may review appeals from users whose content Meta removed, appeals from users who reported content that Meta left

up, and decisions that Meta refers to it (Charter Article 2, Section 1). The Board has binding authority to uphold or overturn Meta's content decisions (Charter Article 3, Section 5; Charter Article 4). The Board may issue non-binding recommendations that Meta is required to respond to (Charter Article 3, Section 4; Article 4). Where Meta commits to act on recommendations, the Board monitors their implementation.

- For this case decision, independent research was commissioned on behalf of the Board. The Board was assisted by Duco Advisors, an advisory firm focusing on the intersection of geopolitics, trust and safety, and technology.