Taiwan Job Scam Warning

2026-003-FB-UA

Summary

The Oversight Board calls for Meta to do more to stop fraudulent online labor recruitment. In addition to removing fraudulent recruitment content from its platforms that leads to offline harm, Meta should introduce an informative notice when users engage with content that may violate its policies, but where Meta's automated systems do not have enough confidence to remove it. This would provide additional protection to users from scam content that spreads across platforms.

In analyzing a case on the removal of content from a Taiwanese police department warning about job scams, the Board has overturned Meta's original decision to take down the content.

About the Case

In October 2024, a Taiwanese police department reshared a post on its Facebook page. The post contains an image of animated pigs and a bird in a police uniform holding a sign. Overlay text in Chinese describes the signals of job scams and warns job seekers. The caption includes a similar list of job scam keywords, advice on how to prevent being scammed and information on an anti-scam hotline.

In July 2025, Meta's automated systems identified the content as potentially violating the Human Exploitation Community Standard, then removed it. An administrator of the police department's Facebook page appealed to Meta. A human reviewer upheld the original decision. The administrator then appealed to the Board, stating that the post aimed to prevent fraud and was part of a governmental initiative to educate the public and raise awareness on safe employment practices.

When the Board brought the case to Meta's attention, Meta's experts reviewed the post under the Human Exploitation and Fraud, Scams, and Deceptive Practices policies and concluded it was shared to raise awareness and educate. The company restored the post.

Online labor scams by transnational crime syndicates tricking people into being trafficked or stealing people's money are a significant problem on social media. Social media posts are reportedly the fastest-growing source of scams in Taiwan, with most online scam losses stemming from Facebook ads. The Board found that many posts with signs of job scams request follow up on messaging platforms off Facebook.

Key Findings

In addition to removing fraudulent recruitment content from its platforms that lead to offline harm, Meta should explore ways to improve its technology to better distinguish non-violating anti-scam content.

There may also be a range of content that has some signals of fraudulent recruitment, but more tenuous links to harm. To protect expression while still protecting against the potential of serious offline harm, Meta should explore less intrusive means targeting these specific patterns.

For example, Meta's Messenger chats employ advanced scam detection that allows users to send recent chat messages for AI scam review when "a new contact sends a potentially scammy message." If a potential scam is detected, users receive a warning pop-up that outlines information on common scams and suggests actions including blocking or reporting the suspicious account.

To disrupt the spread of fraudulent labor recruitment across platforms and to provide additional protections to users, Meta should introduce a similar informative notice for its platform users. This notice would not apply to posts that violate Meta's Human Exploitation or Fraud, Scams, and Deceptive Practices policies, which should still be

removed. However, there is a significant gray area in enforcing this policy, given evolving evasion efforts in a highly dynamic space.

The Board finds that while it may have been challenging for Meta's classifier to assess this post, it is clearly anti-scam content. It does not violate either the Human Exploitation or Fraud, Scams, and Deceptive Practices policy. The Board finds that removing the content from Facebook was not consistent with Meta's human rights responsibilities.

The Oversight Board's Decision

The Board overturns Meta's original decision to take down the content.

The Board also recommends that Meta:

- Should introduce an informative notice to disrupt the spread of fraudulent labor recruitment across platforms. It will be applied when users are engaging with (react, comment, share or click on an external link) content that is flagged by Meta's technology as involving signals of job fraud and recruitment into labor exploitation, but left on the platform due to low or medium levels of confidence for removal.

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

Full Case Decision

1. Case Description and Background

In October 2024, a police department in Taiwan reshared a post on its Facebook page. The reshared post, which is in Chinese, contains an image of animated pigs and a bird in a police uniform holding a sign. Overlay text on the image describes the signals that

indicate content is a job scam and warns job seekers about common job scam keywords and tactics, such as offering high salaries, not requiring any work experience, or providing easy moneymaking opportunities. The image caption includes megaphone and red triangle emojis, lists keywords and provides advice on how to prevent being scammed. The caption ends with information about an anti-scam hotline.

In July 2025, Meta's automated systems identified the content as potentially violating the Human Exploitation Community Standard, then removed it.

An administrator of the police department's Facebook page appealed to Meta, and a human reviewer upheld the original decision to remove the post. The page administrator then appealed to the Board, stating that the post aimed to prevent fraud and was part of an official governmental initiative to educate the public and raise awareness on safe employment practices.

When the Board brought the case to Meta's attention, Meta's subject matter experts reviewed the post under both the Human Exploitation and Fraud, Scams, and Deceptive Practices policies, and concluded that it was shared to raise awareness and educate users on common scam tactics and labor exploitation. As a result, Meta reversed its original decision and restored the post.

The Board notes the following context in reaching its decision:

Human trafficking is a serious human rights problem, and social media can be used to recruit workers who are forced to perpetrate online scams on a global scale. In 2023, Interpol issued a global warning on human trafficking-fueled fraud, noting "the scheme, where victims are trafficked to work in online scam centers, has shifted from regional crime trend to global threat."

Scammers often share high-salary "overseas" roles through Facebook, Instagram and Tinder ads and recruiter outreach. Targets are lured to travel overseas, where their passports are seized, and they are coerced into running online fraud from fortified compounds or other forms of migrant labor. Hundreds of Taiwanese have also fallen victim to these operations coordinated by transnational crime syndicates operating

across Southeast Asia, particularly in Cambodia, Myanmar, Laos and Thailand. It is estimated that hundreds of thousands of victims are trafficked into such compounds, including about 100,000 in Cambodia and 120,000 in Myanmar. According to the US Institute of Peace, as of the end of 2023, these syndicates have annually stolen about US$64 billion worldwide.

Social media posts are the fastest-growing source of scams in Taiwan, according to the Safer Internet Lab. Facebook and the LINE messaging app stand out as the leading platforms where individuals encounter online scams. Nearly 70% of online scam losses stem from Facebook ads in the form of investment scams and product endorsements by fake celebrities. Additionally, 98% of fake ads reported to the police in Taiwan were found to have originated from Meta platforms.

In July 2024, Taiwan passed several key anti-fraud measures, including the Fraud Crime Hazard Prevention Act (FCHPA) and amendments to the Code of Criminal Procedure, the Communication Security and Surveillance Act, and the Money Laundering Control Act. In November 2024, a new version 2.0 of the anti-fraud guidelines (2025-2026) was passed, primarily focused on financial sector fraud prevention, crypto industry regulations and scam awareness. Under the FCHPA, online platforms have several anti-fraud responsibilities, including a requirement to remove fraudulent advertisements within 24 hours of receiving notice.

Meta has been fined several times under FCHPA. In May 2025, Meta's Facebook platform was fined NT$1 million (US$33,258) for failing to disclose information regarding the advertisers and funding sources behind certain advertisements on Facebook. In July 2025, the Ministry imposed a NT$15 million (US$512,864) fine on Meta for failing to disclose key information about advertisers on Facebook. In August 2025, Meta was fined a third time: NT$2.5 million (US$81,914) for its inadequate response to an order to remove false advertisement on Facebook. Taiwan's Ministry of Digital Affairs noted that "while Meta had already complied with the order ... and removed 95% of the reported ads in a timely manner, 5% failed to meet the 24-hour deadline."

In December 2025, Meta reported that in the preceding 15 months reports about scam ads had declined by more than 50% and that so far in 2025 Meta had removed more

than 134 million scam ads. Meta stated that in the first half of 2025 its teams detected and disrupted nearly 12 million accounts — across Facebook, Instagram and WhatsApp — associated with the most adversarial and malicious scammers: criminal scam centers. Earlier, a Reuters investigation reported that Meta internally projected that in 2024 about 10% of its total ad revenue came from ads linked to scams or banned goods.

2. User Submissions

The user who appealed Meta's decision to the Board explained that the post aimed to prevent fraud and was part of an official governmental initiative to raise awareness on safe employment practices.

3. Meta's Content Policies and Submissions

I. Meta's Content Policies

Human Exploitation Community Standards

According to the Human Exploitation policy rationale, Meta "remove[s] content that facilitates or coordinates the exploitation of humans, including human trafficking." The Community Standard prohibits: "Content that recruits people for, facilitates or exploits people through [several] forms of human trafficking [including] labor exploitation."

The policy includes exceptions to these rules and states that Meta "allow[s] content that is otherwise covered by this policy when posted in condemnation, educational, awareness raising or news reporting contexts."

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." The policy rationale notes that this includes content that "seeks to coordinate or promote those activities using [Meta's] services." The rationale also notes that the policy allows people to "raise awareness and educate others as well as condemn these activities."

The policy prohibits job fraud and scams, defined as content that "offers jobs with an unclear or vague job description and get-rich-quick opportunities promising money with little time investment or effort," or "offers jobs containing no job information, simply referencing job vacancies." The policy also does "not prohibit content that condemns, raises awareness of or educates others about fraud and scams, without either revealing sensitive information or promoting fraud or scams."

II. Meta's Submissions

As a result of the Board selecting this case, Meta reversed its original decision to remove the post, concluding that the content did not violate either the Human Exploitation or the Fraud, Scams, and Deceptive Practices policies. The company determined that the post was shared to raise awareness and educate users on common scam tactics and labor exploitation.

According to Meta, the post's awareness-raising and educational purposes are illustrated through the anti-scam tips and the reference to a fraud hotline. For Meta, the pig imagery is a reference to pig-butchering schemes that involve financial deception or forced labor. The company highlights that "the visual elements, such as the inclusion of the speaker and red triangle emojis, draw attention to the warning signs and reinforce the informative nature of the post." Finally, Meta considers the fact that a local Taiwanese police department shared the post to show that "the content is intended to inform its primarily Taiwanese audience of job-related scams and exploitation."

In response to the Board's questions, Meta emphasized that while both Human Exploitation and Fraud, Scams, and Deceptive Practices policies are designed to prioritize minimizing physical, financial and privacy harms to users, the company

recognizes the importance of allowing awareness-raising or condemning content. Both human moderators and classifiers review content under these policies in all markets. Content reviewers are instructed to review all captions, posts, videos or images to determine whether the content violates the policy or is being shared in an allowable context. Meta noted that in this case, after the classifier first removed the content, the reviewer on appeal misunderstood the post's educational purpose because they missed important keywords and phrases in the caption, the anti-scam hotline and the fact that the post originated from a local police department. Meta added that given that bad actors continually adapt their tactics to evade detection, the company regularly reviews its enforcement practices to identify evolving patterns of abuse and minimize the removal of non-violating content.

Meta stated that as the content highlights the same signals that are commonly associated with scams or exploitation, classifiers may struggle to distinguish violative content from genuine educational content. Classifiers are trained to detect violation indicators and the prevalence of educational or condemnation content is relatively rare. Consequently, classifiers may misclassify posts that aim to inform or raise awareness.

In response to the Board's questions, Meta explained that following the company's announcement on January 7, 2025, large language models (LLMs) are now more widely integrated as an additional review layer. Meta informed the Board that the classifier that detected the violation did not rely on LLMs or other generative AI technology. Nor did an LLM provide a second opinion for the enforcement decision, although the company is currently "using LLMs to improve the training data quality and the precision for the classifier in this case."

In response to the Board's questions, Meta stated that it "has specific channels to receive and evaluate reported fraud/scam content from regulators." Specifically in Taiwan, Meta receives takedown requests issued under the Anti-Scams Code that focuses on fraudulent ads from such entities as the Ministry of Digital Affairs (MODA), the enforcement authority under the Anti-Scams Code, law enforcement authorities

including the Taiwan Criminal Investigation Bureau and Meta's local legal representative, as required under the Code.

The Board asked questions on anti-scam policy enforcement considerations, including the policy exceptions, Meta's approach to receiving and evaluating fraud or scam content takedown requests from regulators in Taiwan, and the use of LLMs or other generative AI technologies in policy enforcement. Meta responded to all the questions.

4. Public Comments

The Oversight Board received two public comments that met the terms for submission. One was submitted from Asia Pacific and Oceania, and one from United Stated and Canada. The submissions focused on the widespread nature of fraudulent advertisements on Facebook in Taiwan, criticized the ineffectiveness of the company's enforcement and noted the impact of incorrect enforcement on job postings. To read public comments submitted with consent to publish, click here.

5. Oversight Board Analysis

The Board selected this case to assess Meta's moderation practices in enforcing its Human Exploitation and Fraud, Scams and Deceptive Practices policies, particularly in the context of social media's role in facilitating offline harms related to online job scams. This case falls within the Board's Automated Enforcement of Policies and Curation of Content and Government's Use of Meta's Platforms, two of the Board's seven strategic priorities.

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.

5.1 Compliance With Meta's Content Policies

Content Rules

The Board finds that the post does not violate either the Human Exploitation or the Fraud, Scams, and Deceptive Practices policies. It does not contain an offer of employment, vague or otherwise, and does not recruit for or facilitate labor exploitation.

While it may have been challenging for the classifier to assess this post, as it contains keyword phrases that are commonly used in online job scams, it is clearly anti-scam content. Content reviewers who are instructed to assess all parts of the posts, should have concluded that the post was non-violating, based on the following contextual cues:

- Information on an anti-fraud hotline;
- The warnings and anti-scam tips used in the overlay text and caption;
- Megaphone and red triangle emojis to invite reader's attention to the listed common job fraud keywords and phrases; and
- The use of images of pigs (seemingly referring to pig butchering scams) and the imagery of a bird in police uniform (seemingly referring to police warning of online job scams).

An accurate assessment of the above should have led the reviewer to determine without difficulty that the post was non-violating.

5.2 Compliance With Meta's Human Rights Responsibilities

The Board finds that removing the content from Facebook was not consistent with Meta's human rights responsibilities.

Freedom of Expression (Article 19 ICCPR)

Article 19 of the International Covenant on Civil and Political Rights (ICCPR) provides for broad protection of expression, including "freedom to seek, receive and impart information and ideas of all kinds."

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 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).

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). Additionally, these rules "may not confer unfettered discretion for the restriction of freedom of expression on those charged with [their] execution" and must "provide sufficient guidance to those charged with their execution to enable them to ascertain what sorts of expression are properly restricted and what sorts are not" (ibid). The UN Special Rapporteur on freedom of expression has stated that when applied to private actors' governance of online speech, rules should be clear and specific (A/HRC/38/35, para. 46). 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 on job fraud and scams and labor exploitation are sufficiently clear as applied to this case, and that the content does not violate them.

II. Legitimate Aim

Any restriction on freedom of expression should also pursue one or more of the legitimate aims listed in the ICCPR, which includes protecting the rights of others.

As a policy matter, the Human Exploitation and Fraud, Scams, and Deceptive Practices policies aim to prevent serious offline harms, in this case linked to labor trafficking. Online job scams and recruitment in labor exploitation have been proliferating in East and Southeast Asia, and globally. Victims of fraudulent online labor exploitation face serious threats to their right to life, liberty and security of the person.

In this case, the Board finds that Meta's prohibition on recruitment into labor exploitation under the Human Exploitation policy aims to protect users' right to life, liberty and security of the person; and right to be free of torture, cruel, inhuman and degrading treatment or punishment, arbitrary detention and forced labor (Articles 6-9, ICCPR).

Meta's prohibition on fraudulent job offers also pursues a legitimate aim by seeking to protect people from scams and fraud and their impact on public order (Article 19(3) ICCPR; General Comment 34, para 31; General Comment 37, para 44 as well as Article 17, Universal Declaration of Human Rights, see AI-Manipulated Video Promoting Gambling decision).

III. Necessity and Proportionality

Under ICCPR Article 19(3), necessity and proportionality require 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" (General Comment No. 34, para. 34).

In the face of proliferation of human trafficking as a global human rights challenge, social media often serves as an entry point for deceptive recruitment. Communications

then often migrate to other spaces, including messaging platforms and ultimately severe harm may materialize offline.

To meet its responsibilities, Meta should remove fraudulent recruitment content from its platforms that lead to offline harm. At the same time, mistakenly removing anti-scam content limits the ability to address those same harms. Here, an accurate and comprehensive assessment of the cues in the post should have resulted in it being left on the platform.

Meta noted challenges in automated enforcement for anti-scam content, as it often uses similar language to scam content. The company also suggested it may have limited training data for its classifiers, as non-violating content of this nature is relatively rare. Meta should explore how to improve its technology to better make that distinction. The company mentioned it is using LLMs to improve training data quality for the classifier used here, and if that improves accuracy both to protect anti-scam content and remove scam content, it should continue these efforts.

The Board is also concerned about false negatives, and acknowledges the highly dynamic nature of fraud prevention, as the company faces constantly evolving tactics to circumvent enforcement.

As part of its research in this case, the Board searched Meta's Content Library for posts from Taiwan that contain signals of fraudulent job posts, such as "easy money" or "any academic qualifications" (as translated from traditional Chinese). At the time of the search, the vast majority of the content with these terms was posted by individual accounts posting to public groups focused on job-searching. The research found that many of the posts appeared to be coordinated, where identical content was posted across multiple groups, usually using the same images, and shared on the same day (sometimes within minutes of each other). More than half of the posts mentioned that those interested in the job listing should contact the poster off platform, through WhatsApp, LINE or other messaging platforms. This off-platform migration poses additional enforcement challenges to protect against harm.

Additionally, anti-fraud measures in Taiwan focused on paid content could lead to increased fraud in organic content. As regulation requires removal of scams in ad content within 24 hours of notification, Meta may focus more on removals of this type of content in Taiwan. As bad actors may find it more difficult to place scam ads, they may migrate to using organic content for recruitment.

Within this context, to both protect anti-scam expression and protect against the serious harms associated with violating scam content, especially those connected to fraudulent recruitment, Meta should explore a broader range of tools beyond the mere removal of content. For example, some potentially violating content may contain signals of fraudulent recruitment but not with high enough confidence levels for classifiers to determine that it violates the policy, and would remain on the platform with the potential to cause harm. On the other hand, lowering the confidence thresholds to allow the classifiers to remove more potentially violating content could result in the removal of content that raises awareness or educates about scams. Therefore, to protect expression while still protecting against the potential of serious offline harm, Meta should explore the use of less intrusive means than removals, such as user notifications and anti-scam educational practices, when targeting such content.

Meta has already taken several measures in this respect. For example, it currently allows users to enable advanced scam detection in Messenger chats. According to Meta, "when it is enabled," and "a new contact sends a potentially scammy message," the users receive a warning pop up, giving them an option to send recent chat messages for AI scam review. If a potential scam is detected, the users receive another warning pop-up that outlines information on common scams and suggests actions including blocking or reporting the suspicious account. Similarly, Meta provides warnings on WhatsApp when a user attempts to share the screen with an unknown contact during a video call.

Meta should consider adopting such best practices for more users. For example, it could consider measures to make the anti-scam detection tool more prominent, such as

sending a notification to remind users of this option. It could also expand the use of notices outside of Messenger. These could be triggered when users engage with (react, comment, share or click an external link) content that has been flagged by technology as including some signals of recruitment for labor exploitation, but left on the platform due to low or medium levels of confidence for removal. Seeing informational notices when engaging with posts that meet the low or medium confidence thresholds could alert the users to exercise greater caution.

A notice could indicate, for example, that the user is leaving Meta's platforms, such as Facebook, Instagram, Threads or Messenger, describe examples of common recruitment scams, including that off-platform migration is a common pattern for fraudulent labor recruitment, and suggested actions such as blocking or reporting suspicious content. Such notices should be available in all languages supported by the platforms.

6. The Oversight Board's Decision

The Oversight Board overturns Meta's original decision to take down the content.

7. Recommendations

Enforcement

1. To disrupt the spread of fraudulent labor recruitment under Human Exploitation or Fraud, Scams, and Deceptive Practices policies across platforms and to provide additional protections to users, Meta should introduce an informative notice. It could be triggered when users are engaging with (react, comment, share or click on an external link) the content that is flagged by its technology as involving signals of job fraud and recruitment into labor exploitation but left on the platform due to low or medium levels of confidence for removal.

The Board will consider this recommendation implemented when Meta confirms that the new informative notices are provided to users in all languages supported by the platform.

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.