



PAY-TO-PLAY: META'S COMMUNITY (DOUBLE) STANDARDS ON PORNOGRAPHIC ADS

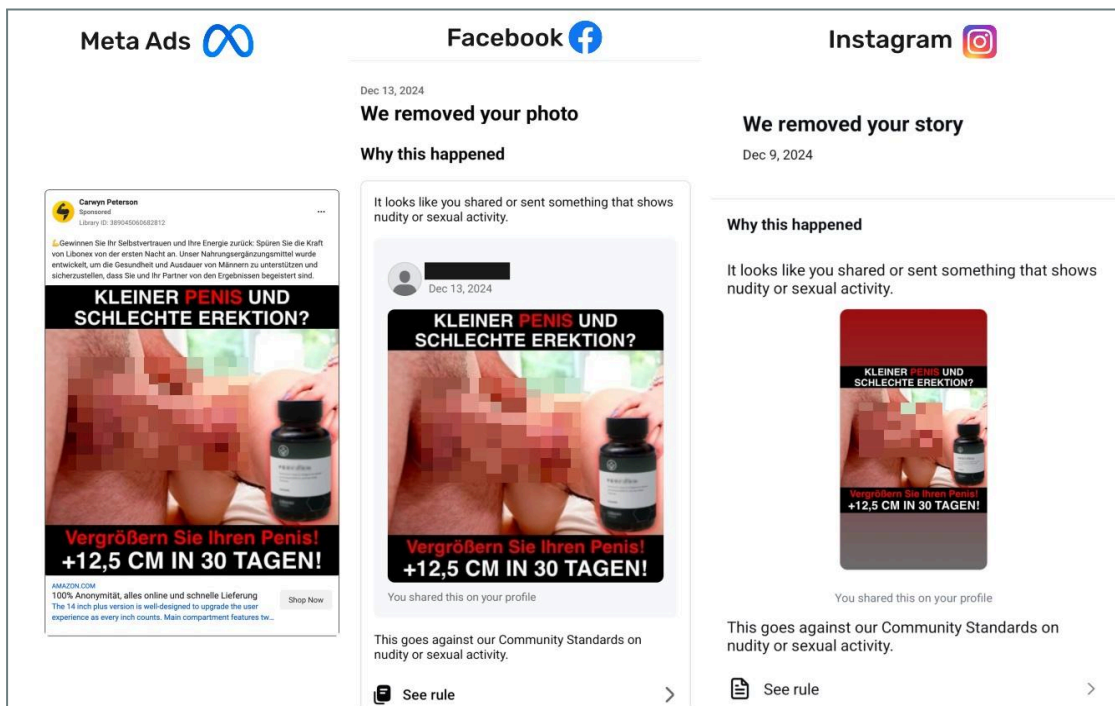
Executive Summary

Expanding on [our previous investigations](#) on coordinated malicious activity on Meta Ad Library, we identified over **3,000 pornographic advertisements** illustrated with fully explicit **adult nudity and sexual activity**, which were **reviewed and approved by Meta** despite constituting a clear violation of Meta's Community and Advertisements standards.

These **pornographic advertisements**, promoting dubious **sexual enhancement products**, generated over **8 million impressions** in the European Union last year, targeting a predominantly 44+-year-old male demographic.

To demonstrate that the lack of moderation is not due to a technical limitation, we uploaded the same visuals as standard non-sponsored posts on Instagram and Facebook. They were promptly removed for violation of Meta Community Standards.

Our findings suggest that although Meta has the technology to automatically detect pornographic content, it does not use it to enforce its community standards on advertisements as it does for non-sponsored content. This **double standard** is not a temporary bug but has persisted since at least December 2023.



We blurred explicit content, though the original ads were unredacted.

Remarks

This report contains partially blurred, sexually explicit images. We strive to balance reader sensitivity with the need to represent actual content approved by Meta in its advertising ecosystem.

Furthermore, as documented in this research, Meta can detect and remove plain pornographic content that violates its community standards. As such, we expect Meta to retroactively moderate the ads discussed in this report following its publication. In preparation for such an event, which would prevent further analysis by third parties, we curated an archive of the identified ads. Screenshots and their associated media are available upon request for regulators and other relevant actors.

Finally, this investigation was made possible by Meta's diligent implementation of a public ad repository mandated by Article 39 of the DSA. Meta's provision of a functional and relatively comprehensive ad library and API demonstrates a higher transparency standard than its peers, as [documented by the Mozilla Foundation](#). This report showcases how researcher data access serves as an effective mechanism for platform accountability through public scrutiny. We urge all platforms to fulfill their data provision requirements, with the ad repository being an enforcement priority.

Credits

Research: Dr. Paul Bouchaud

Report: Dr. Paul Bouchaud, Raziye Buse Çetin, Natalia Stanusch, Salvatore Romano, Marc Faddoul

Graphic & Brand Design: Denis Constant / Ittai Studio <https://ittai.co/>

All other content (c) AI Forensics 2024

Email : info@aiforensics.org

Social Media: [Linkedin](#) | [Bluesky](#)

This AI Forensics research was funded by core grants from [Open Society Foundations](#), [Luminate](#), and the [Limelight Foundation](#).

Table of Contents

Executive Summary	2
Remarks	3
Credits	3
Table of Contents	4
Methodology	5
Data Collection	5
Dataset Characteristics	6
Collecting the Visuals	7
Deduplicating the Visuals	7
Detecting Pornographic Content	9
Annotating the Visuals	10
Findings	11
Conclusion	18

Methodology

Data Collection

From December 8th-11th, 2024, we collected 14,363 advertisements that had run in France, Germany, Spain, Italy, and Poland over the previous year. The ads, collected through the Meta Ad Library API, were queried for containing the following expressions¹: “strong erection,” “erection cm,” “small penis,” “long penis,” “pornhub,” or “penis erection.”

These targeted keywords, translated to the languages of the respective countries analyzed, emerged from preliminary exploratory analysis and manual searches on the Ad Library, where we identified numerous advertisements featuring sexually explicit content promoting purported sexual enhancement products marketed toward men.

By this choice of keywords, we do not, nor claim to, fully characterize the extent of pornographic advertisements on Meta's advertising systems. Instead, we highlight systemic shortcomings from Meta with a small number of queries.

To broaden our data collection beyond this handful of keywords, we additionally collected all advertisements run in the countries analyzed by the advertiser pages that published these advertisements in the first place. We focused on 50 pages that published at least 10 non-moderated pornographic ads in our initial keyword search dataset. This snowball sampling yielded 1,784 additional unique advertisements.

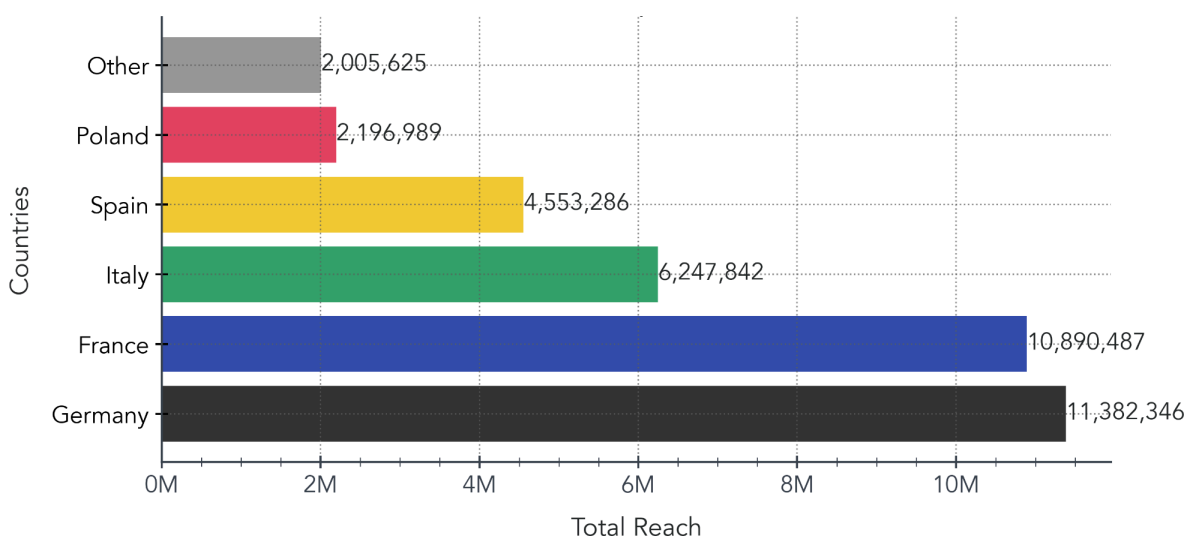


Figure 1: Total reach per country

¹ In French, we also added “Cela a aidé mon ami” as it was found during the exploratory analysis to be used in some sexual enhancement ads.

Dataset Characteristics

We collected 16,147 unique advertisements, predominantly targeting Germany and France, as displayed above. These ads predominantly reached males, accounting for 94.9% of the 37.3 million impressions.

The age distribution, displayed below, skewed toward older demographics for male users, aligning with marketing claims targeting sexual performance decline. We display examples of such clear marketing below.

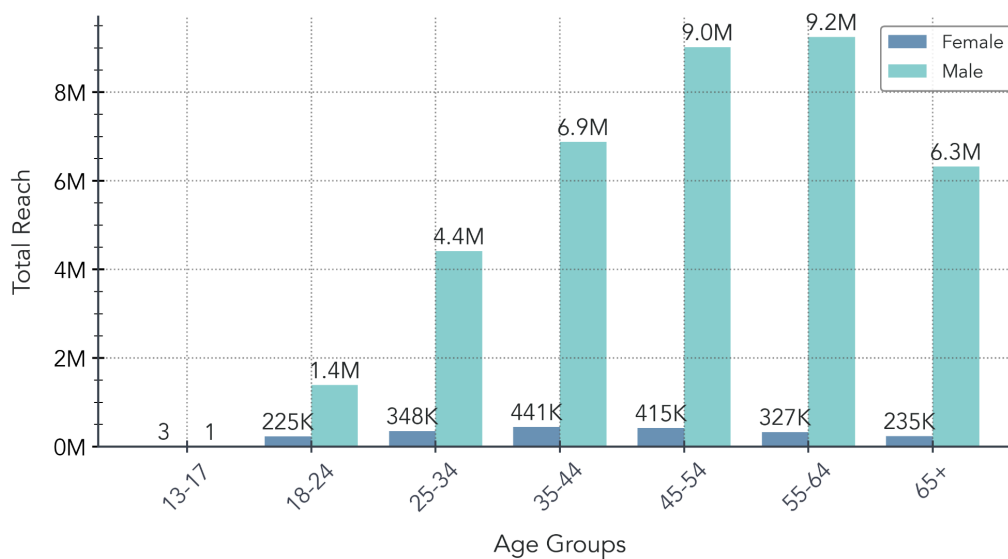


Figure 2: Age distribution of ads per gender



Figure 3: Examples of advertisement targeting older men

Collecting the Visuals

As of December 2024, Meta had removed 24.7% of the ads in our dataset for violating its “Advertising Standards.” As creative content of ads removed by Meta are replaced with a disclaimer, we cannot perform the detection of pornographic content on those.

The Meta Ad Library compared favorably with other major platforms’ ad repositories, allowing full-text search and implementing OCR capabilities. However, its API only provides text content and links, not visual content. Therefore, we programmatically accessed Meta Ad Library entries to screenshot and download ads’s visual content for the 12,160 non-redacted ads.

Upon loading their preview, and despite no indication in the API results, 3,979 ads displayed the disclaimer, “*This ad was run by an account or Page that we later disabled for not following our Advertising Standards.*” Once again, this redacted visual content and prevented further analysis.

Also, as some ads can use multiple versions of creative, we manually observed cases of advertisements illustrated with pornographic visuals in one version and blunt household items in another. For simplicity, we collected only the visuals displayed in the library preview. As such, our results represent a conservative lower bound on the actual amount of pornographic content.

Overall, we successfully downloaded 3,504 images and 3,721 videos used to illustrate the ads. The remaining 956 ads were not illustrated by single images or videos or displayed error messages.

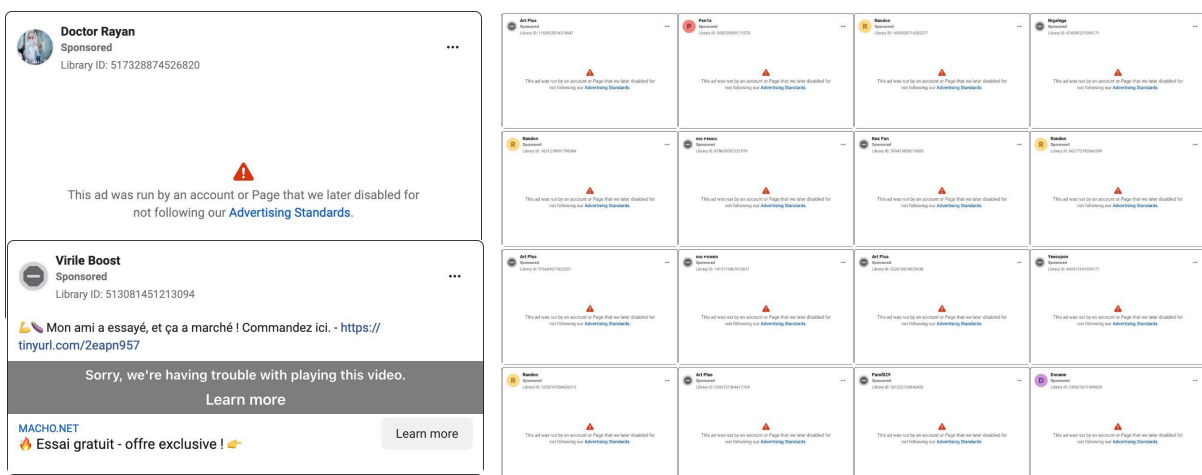


Figure 3: Examples of advertisements displaying a disclaimer or an error preventing the collection of their visual content.

Deduplicating the Visuals

To reduce computational and manual annotation burdens, we deduplicated images and videos before analyzing sexually explicit content. To this end, we leveraged Meta's open-source perceptual hashing algorithm [Pdq](#); for videos, we extracted the frame at the one-second mark (using vPdq-like methods yielded similar results). This approach identified both exact duplicates and near-identical variants, such as different compressions and slight cropping. With a Hamming distance threshold of 17, we identified 484 unique images and 286 unique videos.

The most frequently duplicated images appeared in our dataset in 116, 99, and 64 ads.



Figure 4: Most duplicated visuals

The most duplicated video appeared 240 times in our dataset:

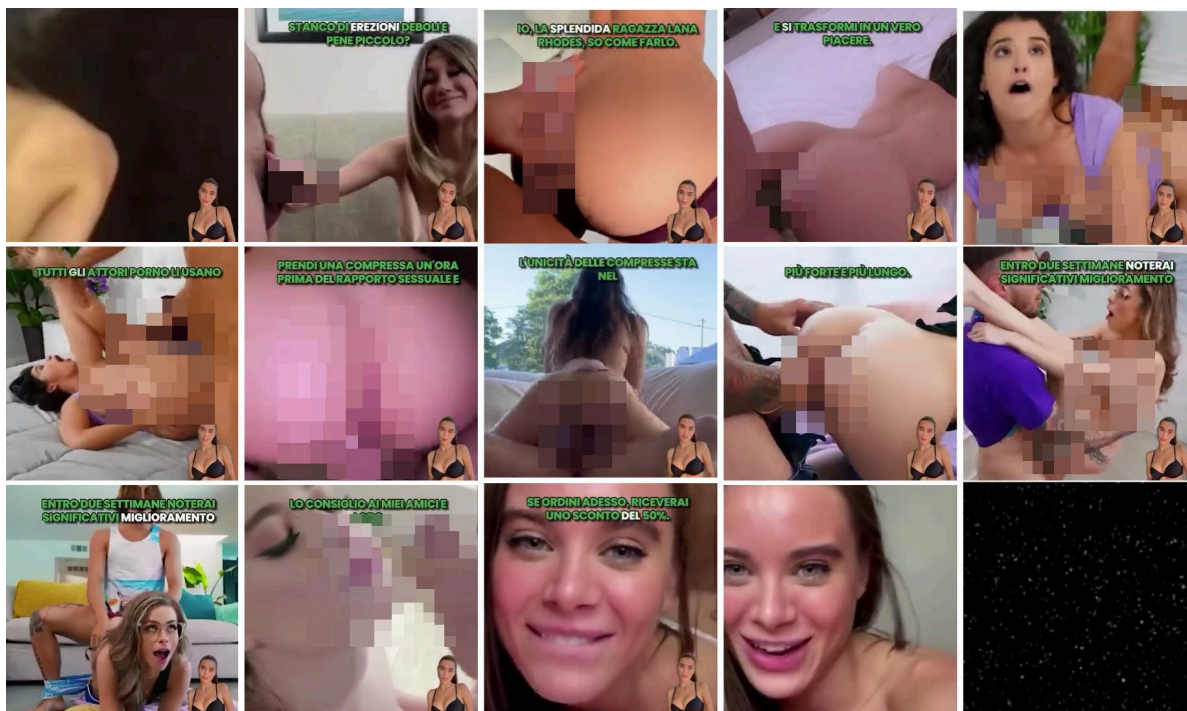


Figure 5: Frames from the most duplicated video

Detecting Pornographic Content

After deduplication, we detected sexually explicit content using an open-weight model released by Yahoo research in 2016 for pornographic content detection. We refer readers to their [blog post](#) for further details.

We selected this 8-year-old model due to its open-weight nature and to highlight that detecting plain pornographic content is a well-established machine-learning task. Thus, any potential shortcomings in content moderation stem from internal moderation policies and implementation choices rather than technological limitations.

For videos, we extracted 100 frames (one every 10th frame in the first 10 seconds) and applied the model to each. Rather than using mean results, which could be fooled by padding videos with neutral content (e.g., 4 seconds of pornographic content followed by 6 seconds of random noise), we used the 90th percentile score (alternative quantiles yielded similar results). For instance, a score of 0.7 will mean that 10% of frames have a NSFW score above 0.7.

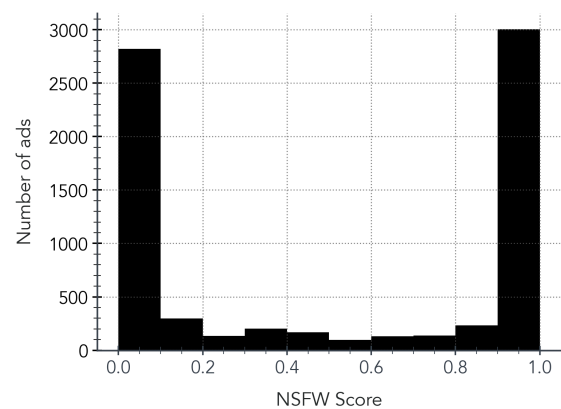


Figure 6: Distribution of visuals' NSFW Score

As displayed in Figure 6, we observe that half of the ads are scored above 0.5 by Yahoo's NSFW detector, with thousands being judged as definitely porn by the model.

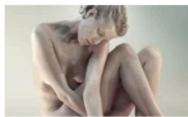


Figure 7: Examples of visuals distributed across the NSFW score range
(in the third image this is a finger)

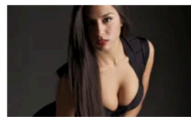
Annotating the Visuals

We performed semi-manual annotation aided by the NSFW detector, following [Meta's Adult Nudity & Sexual Activity advertising policy](#).

In particular, we report what Meta declares as forbidden:



✘ This image depicts near nudity.



✘ This image shows someone in a sexually suggestive pose.



✘ This image depicts a pose that simulates sexual activity.



✘ This image depicts near nudity.

- Depiction of nudity otherwise permitted by the [Community Standards](#) or near nudity such as nudity covered only by digital overlay
- Imagery focused on individual body parts such as groin, buttock or female breast(s), or depicting poses simulating sexual activity
- Visible female nipples in medical/health context while targeting users aged 18 or younger
- Sexual activity where otherwise permitted by the [Community Standards](#)
- Depiction of gestures that signify genitalia, masturbation, oral sex, or sexual intercourse, or sexually suggestive activities like clothed simulated sex, sexual dancing or kissing with visible tongue
- Depiction of logos, screenshots or video clips of known pornographic websites
- Sexual Audio

Figure 8: Meta's Adult Nudity & Sexual Activity advertising policy

To ensure unambiguous identification in our analysis, we applied stricter criteria. This means we focused exclusively on explicit content: full nudity, breasts, visible genitals, or sexual activity. We excluded suggestive imagery, as it is deemed more subjective and can appear in non-pornographic content.



Figure 9: Examples of visuals excluded from our final pornographic ads dataset (while falling within Meta's policies on nudity if applied literally)

Findings

Pornographic ads were shown over 8 million times in the EU over the last year.

We identified at least 3,316 unique advertisements, illustrated with 446 different pornographic visuals (382 unique images and 64 unique videos), cumulating in a total reach of over 8.2 million in the European Union (the total reach being a lower bound on number of impressions and upper bound on number of unique reached).

Far from a momentary “bug” in Meta’s moderation, the pornographic advertisements ran over the entire observation window from December 2023 through December 2024. As Meta removes advertisements precisely one year after their campaign completion, and considering our data collection in December 2024, December 2023 marks the earliest possible detection.

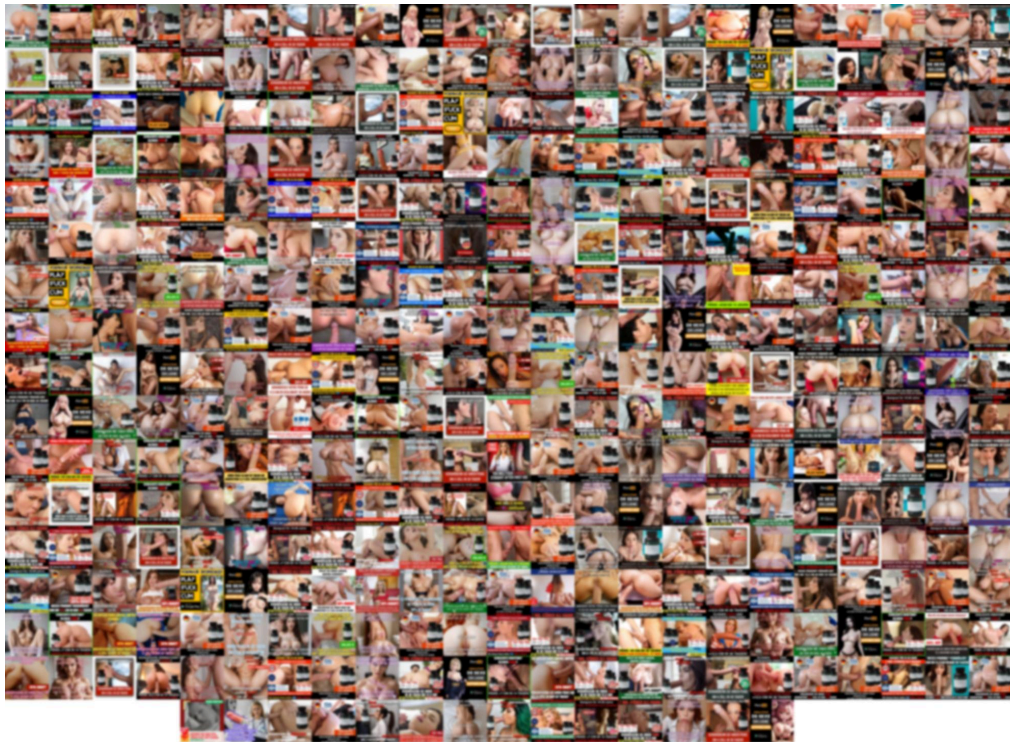


Figure 10: 382 unique images used in the ads approved by Meta

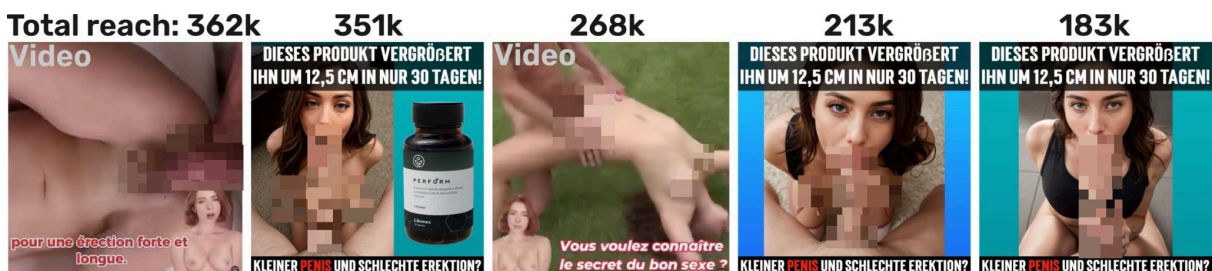


Figure 11: visuals that accumulated the highest total reach across duplicates

The same pornographic visuals get removed by Meta when uploaded from a user account.

Meta claims to review every advertisement before going live on the platform. This review process has been previously criticized for not approving content and advertisements related to sexual health and education of women, trans, or non-binary people while allowing similar content when targeting men. Similarly, Meta was accused of [taking down](#) and [banning](#) user content that is sex-positive and [policing](#) female bodies while [approving](#) explicit advertisements of “AI girlfriends” on Facebook and Instagram.

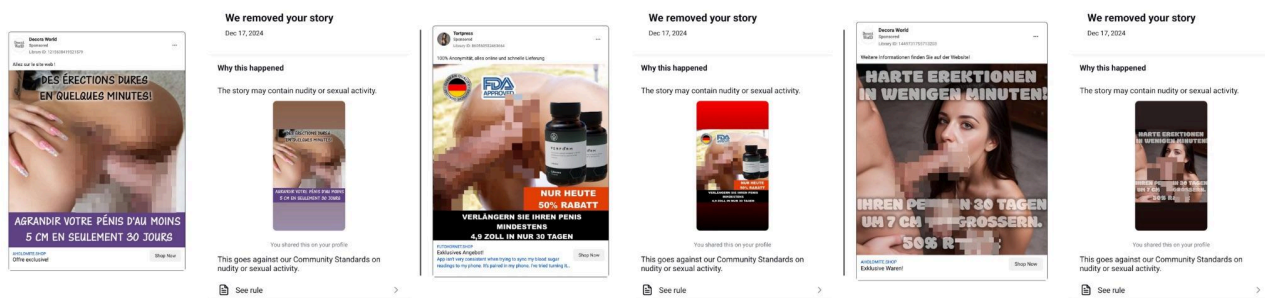
In their [Risk Assessment Reports of August 2024](#), Facebook and Instagram identify Adult Nudity and Sexual Exploitation as a systemic risk. The platforms state that it “remains a highly adversarial and profit-motivated space, particularly as threat actors develop new ways to circumvent detection and enforcement...”


Two hypotheses then emerge to explain why Meta’s ad review process fails to address such blatant violations of its Community and Advertising Standards. First, advertisers might have developed sophisticated methods to circumvent Meta’s automated systems, as Meta claims. The other possibility could be that Meta’s advertising moderation standards were lowered to such a level that they approved advertisements in clear violation of their own advertising policies.

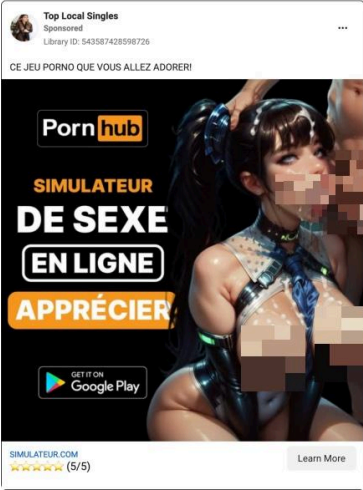
To test these hypotheses, we conducted a simple experiment:


- 1) We created new Instagram and Facebook accounts with new phone numbers. We set the accounts to private with no friends to isolate moderation decisions from user reports.
- 2) We uploaded the same visuals used to illustrate pornographic ads from our detected ads dataset.
- 3) Within minutes, Instagram and Facebook removed all visuals for violating Meta’s community standards on nudity and sexual activity.

Our experiments support the second hypothesis: **Meta’s algorithms can identify sexually explicit content and remove it when posted from a user account, yet when monetized through their advertising system, the same content is approved and actively distributed to millions of users.**



Meta Ads 



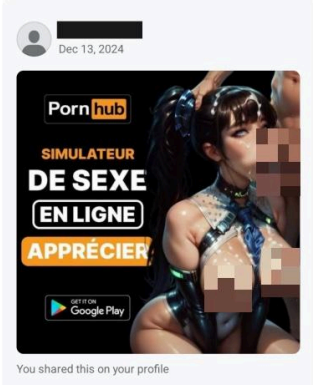
Facebook 

Dec 13, 2024

We removed your photo

Why this happened


It looks like you shared or sent something that shows nudity or sexual activity.



You shared this on your profile

This goes against our Community Standards on nudity or sexual activity.

[See rule](#)


Instagram 

We removed your story

Dec 9, 2024

Why this happened


It looks like you shared or sent something that shows nudity or sexual activity.





You shared this on your profile

This goes against our Community Standards on nudity or sexual activity.

[See rule](#)

Meta Ads 




Facebook 

Dec 13, 2024

We removed your photo

Why this happened


It looks like you shared or sent something that shows nudity or sexual activity.



You shared this on your profile

This goes against our Community Standards on nudity or sexual activity.

[See rule](#)


Instagram 

We removed your story

Dec 9, 2024

Why this happened

It looks like you shared or sent something that shows nudity or sexual activity.



You shared this on your profile

This goes against our Community Standards on nudity or sexual activity.

[See rule](#)

Figure 12: Examples of pornographic visuals approved on Meta Ads but rejected on Instagram and Facebook

AI-generated images, audio, and deepfakes are used to illustrate the ads

Use of AI-generated imagery & video: We note numerous pornographic advertisements employing AI-generated visuals.



Figure 13: Examples of AI-generated pornographic visuals approved on Meta Ads

Celebrity deepfakes: The ads included multiple deepfakes of [Vincent Cassel](#) (one example [archived here](#)), a non-pornographic French actor, promoting sexual enhancement medication through manipulated audio overlaid on pornographic content. Similarly, we found an audio deepfake of [Doctor Michel Cymes](#), who appears to be endorsing sexual enhancement drugs, staged as a fake interview, with non-pornographic imagery used.



Figure 14: Examples of celebrity deepfakes

AI-generated audio: In numerous instances, AI-generated voices of pornographic actors were used to lure and promote products, with the scripts and captions automatically translated for each targeted country.



Figure 15 : Examples of pornographic content, with AI-generated audio, translated in French, German, and Spanish

Incestuous porn ads promote hook-up websites across the EU

During the exploratory analysis, we also identified ads containing pornographic imagery with incestuous themes promoting dubious hook-up dating websites (such as mariagrisha[.]uno). This website was advertised through at least 867 ads across 4 of the 5 studied countries: Germany (381 ads), France (273 ads), Spain (197 ads), and Italy (16 ads), with no presence in Poland. These advertisements accumulated a total reach of 538k impressions.

The non-removed ads we collected revealed an identical pattern across languages: a WhatsApp-like chat featuring a fabricated conversation between a “daughter” and her “father,” where the “daughter” solicits sexual intercourse, accompanied by nude images. The same racy dialogue was (poorly) translated across different languages such as:

Daughter: Dad, I’m turning 18 today. I’m not a little girl anymore

Father: Happy birthday, send me your wet pussy

Daughter: Mom’s not home, I’m caressing my pussy alone at home

or

Daughter: Daddy, I’m ovulating and I’m going crazy

Father: Do you want daddy to come in?

Daughter: I want to put your cock in my mouth again

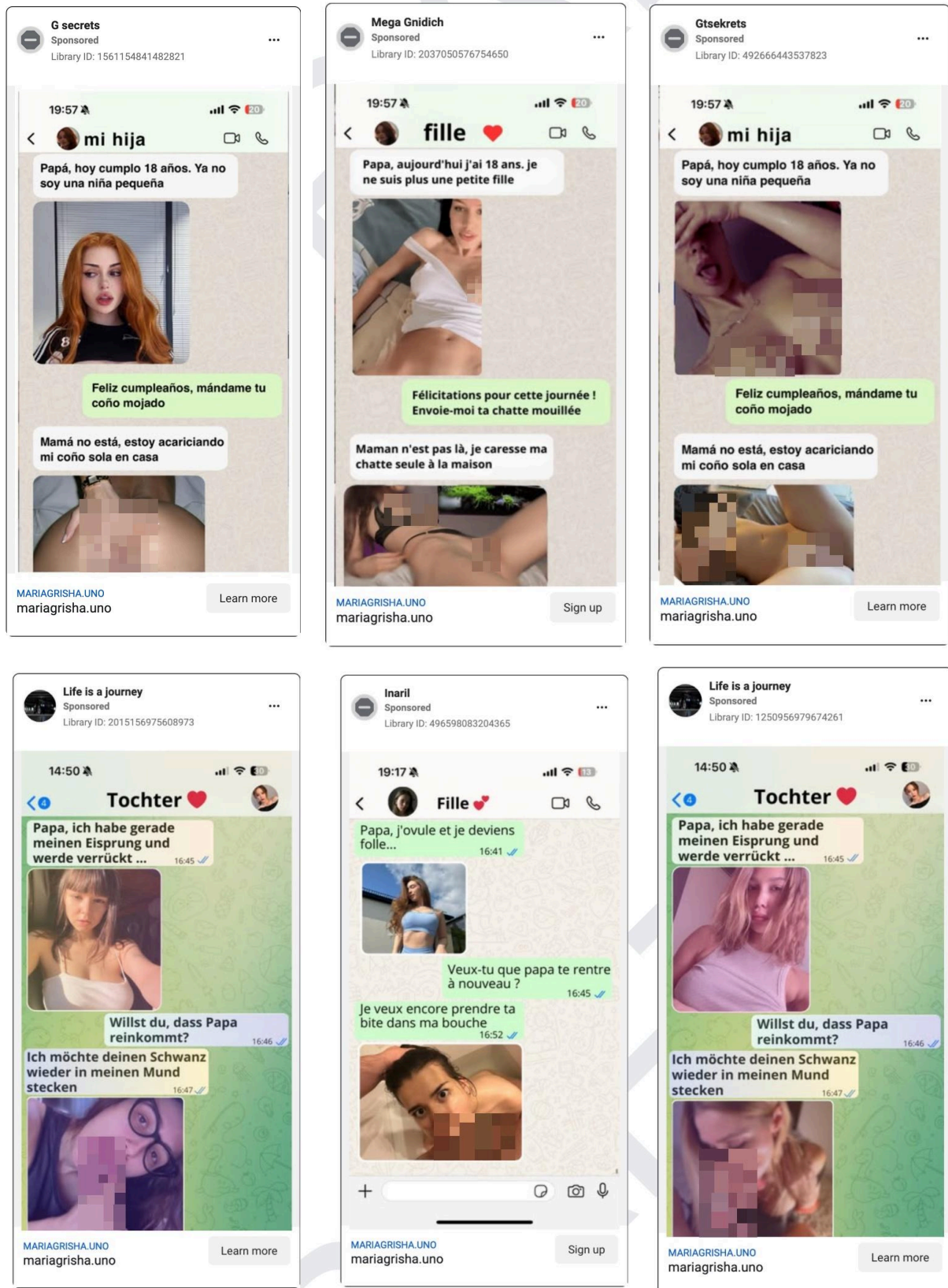


Figure 16: Examples of incestuous porn ads, translated across countries.

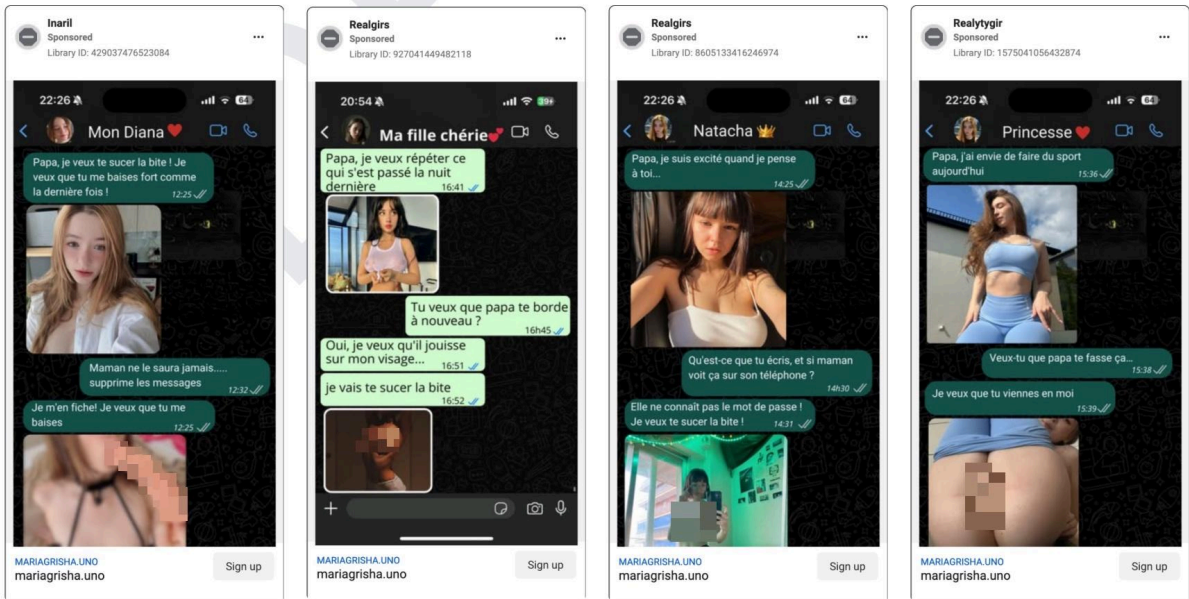


Figure 17: Other variations of incestuous porn ads

Similarly, we replicated our experiment and published the very visuals used in those incestuous porn ads as organic Instagram stories. All explicit imagery was swiftly taken down for moderation of Instagram Community Standards on nudity or sexual activity.

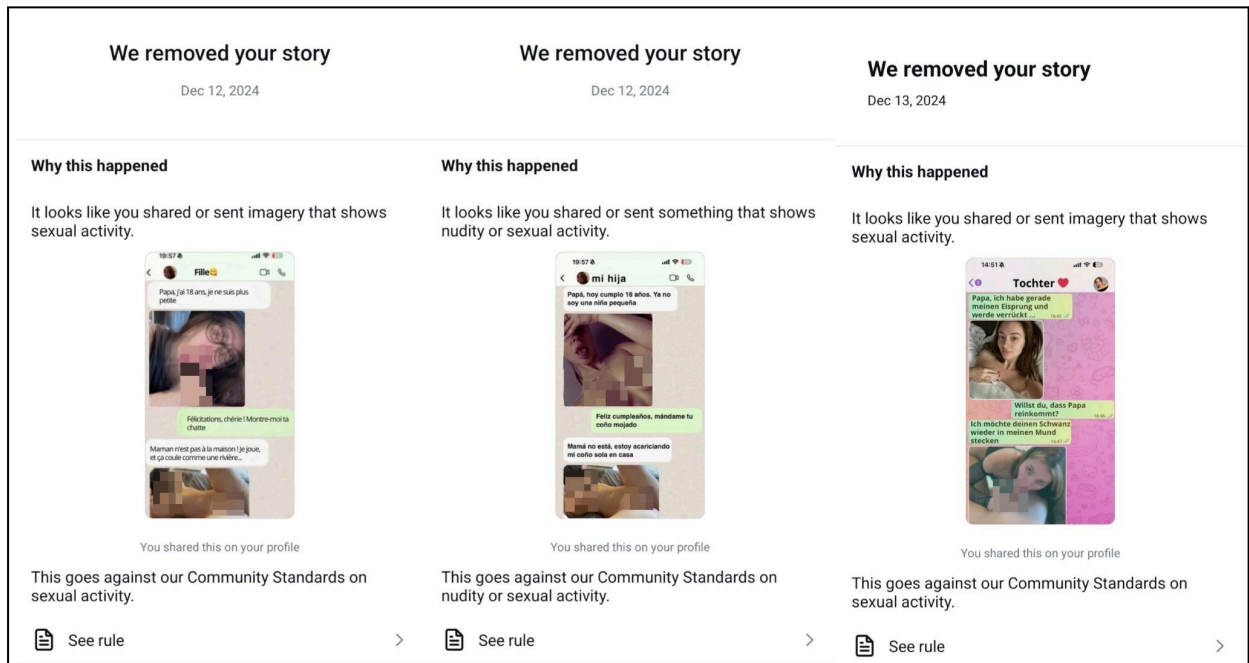


Figure 18: Explicit incestuous porn ads approved on Meta Ads but rejected on Instagram

Conclusion

Our research highlights a systemic double standard in moderation between organic and paid content. Indeed, we have shown i) the presence of plain pornographic content, depicting adult nudity and sexual activity, in thousands of advertisements reviewed, approved, and distributed by Meta, and ii) that Meta possesses the technical capability to moderate these exact same pornographic visuals when uploaded from user accounts.

Facebook and Instagram stated in their Risk Assessments conducted per Article 34 of the European Digital Services Act that:

“ Meta proactively reviews all advertisements before they are able to be published on Meta’s platforms [...] An ad is broken down into its various components, such as the title, images, and other text, to review if any portion of the ad violates Meta’s Advertising Standards. The decisions made during these reviews determine whether ads are approved and go live, or if they are disapproved and returned to the advertiser. ”

*Facebook Risk Assessment page 55,
Instagram Risk Assessment page 53*

As such, the statements made in Instagram and Facebook Risk Assessments may appear misleading, as the “*proactive review*” of advertisements to supposedly enforce the platform’s Advertising Standards largely fails to do so. It would indeed appear that **Meta has lowered their moderation standards** on their advertising ecosystem—their core source of revenue—to **such an extent** as to **approve pornographic content**.

Beyond these immediate findings regarding Meta’s double standard in content moderation practices, this investigation underscores the critical importance of civil society organizations having access to comprehensive transparency datasets. Meta’s implementation of the Ad Library demonstrates how such datasets enable independent verification of risk assessments published by Very Large Online Platforms and Search Engines. Therefore, we urge all platforms to fulfill their data provision obligations under the Digital Services Act, with advertisement repositories being an enforcement priority.
