EU Digital Services Act (EU DSA) Annual Transparency Report - 2024

Reporting Period: January 1, 2024 - December 31, 2024
Publication Date: 2025-03-11
Previous Report: N/A

In this Transparency Report, we provide a comprehensive overview of our content moderation efforts from January 1 to December 31, 2024.

We detail the volume and types of content we removed or restricted, how we handled user reports (and how quickly), the balance between automated filters and human moderation, the outcome of user appeals, and regional patterns in moderation. We also highlight emerging trends and outline how we plan to further enhance safety on our platform.

Overview

Wildlife Studios is committed to fostering a safe and trustworthy online environment. To achieve this, we employ a combination of proactive and reactive measures to identify and address harmful content.

Our content moderation framework is guided by the following principles:

- Transparency: We strive to provide clear and accessible information about our policies and practices.
- Accountability: We are committed to holding ourselves accountable for our decisions.
- Fairness: We aim to apply our policies consistently and fairly.
- Privacy: We respect the privacy of our players.

Our enforcement efforts increased steadily throughout the year, reflecting both a growing user base and enhanced detection capabilities. Notably, moderation activity peaked in late 2024 after we deployed new anti-cheat and anti-spam tools, which led to a higher volume of automated actions in Q3 and Q4.

Our Approach to Content Moderation

Wildlife Studios utilizes automated systems and human review to proactively identify and remove content that violates our terms of service. These measures include:

- Automated detection: We use machine learning algorithms to identify potentially harmful content, such as spam, hate speech, and child sexual abuse material.
- Human review: Our content moderation team reviews content flagged by automated systems and user reports.

We also respond to user reports of harmful content. Our process for handling user reports includes:

- Review: We review each report to determine whether the content violates our terms of service.
- Action: If we determine that the content violates our terms of service, we take appropriate action, such as removing the content, suspending the account, or reporting the content to law enforcement.

This report presents data on the following aspects of our content moderation efforts:

- User reports: The number of user reports received and the types of content reported.
- Own initiative moderation: The number of items moderated proactively by Wildlife Studios.
- Orders: The number of orders received from government agencies and courts to remove content.
- Internal complaints: The number of internal complaints received about content moderation decisions.

We undertook more than 16 million content moderation actions in 2024, caused either by user reports or via our automated content moderation tools. These actions led to enforcement of our Community Guidelines through content removals, player sanctions, and other protective measures. This figure includes all interventions – from deleting toxic chat messages before they reach other players to banning accounts engaged in severe misconduct.

User Reports

In 2024, player submissions of reports about potential misconduct were a key input to our moderation process. Players can easily flag chat messages, profiles, or other content in-game for our review, and they did so actively. About We 1,100 instances of content were removed or users sanctioned in response to player flags in 2024:

Sexual Content – ~288 cases (e.g. explicit or pornographic language and imagery in chats)
Violence and Threats – ~246 cases (e.g. violent threats or graphic content)
Political/Religious Speech – ~179 cases combined (content related to political or religious topics that violated our rules or local laws)
Hate Speech & Harassment – ~162 cases labeled *Illegal or Harmful Speech* (e.g. slurs, extreme toxicity, or hate content targeting identity)
Extremism/Radicalization – 65 cases (promotion of terrorism or extremist ideology)
Inappropriate Usernames – 80 cases (offensive or impersonating player names)
Ban Evasion – 18 cases (attempts to return after prior account bans)
Other Violations – ~17 cases (misc. categories such as “Negative civic discourse” or minor safety issues)

These figures show that sexually explicit and violent content were the most common problematic contents our team dealt with, together making up more than 500 incidents in 2024. This trend is consistent with what other game communities report.

Harassment and offensive content (including sexual or hateful speech) also frequently led to reports and penalties, indicating our community's intolerance for abusive behavior.

Fraudulent or scam content was comparatively rare in Wildlife's games (we did not register notable in-app fraud cases in 2024), likely due to our games' design and proactive scam prevention measures.

Also, when a player submits a report, we strive to acknowledge and address it as quickly as possible. In 2024, the median time to process a user report (from the moment a report was received to the time a decision was made) was approximately 36 hours.

In many cases – especially for clear-cut violations like overt hate speech or sexual predation – our response was much faster, often within a few hours. However, more complex cases or off-peak reports could take longer to evaluate, bringing the overall median to under two days.

We have 24/7 moderation coverage and prioritize reports by severity; urgent safety issues are typically addressed well within 24 hours.

Automated Moderation

Wildlife Studios employs a hybrid moderation model that leverages both automated systems and human expertise. In 2024, the balance was skewed toward automation, with the vast majority of content decisions made by our AI-driven tools.

While user reports help us take targeted action against players who violate our Terms of Service and Community Guidelines, automated moderation works in real time to filter out inappropriate content before it ever reaches other players, for instance by removing inadequate images or messages.

Also, unlike user reports - which can directly lead to account penalties such as warnings, suspension and bans - our automated filters act as a first line of defense by blocking chat messages or usernames deemed unsuitable. As a result, it prevents the offending player's content from being displayed to others in the game.

The table below shows the total number of chat messages or images blocked by our automated filter in 2024, broken down by violation category.

CSAM – 22 cases

Doxxing – 13.340 cases
Hate speech – 7.790 cases (slurs or hateful language)
Insults – 150.070 cases (general abusive or insulting remarks directed at other players)
Personal data – 3,170 cases (sharing personally identifiable information)
Profanity – 9,227,130 cases (swearing or coarse language)
Sexual content – 67,140 cases (explicit sexual content)
Vulgar – 7,247,640 cases (strongly offensive language or gestures not classified directly as hate speech or insults)

Profanity and vulgar content make up the largest share of automatically blocked messages—together numbering in the millions over the course of 2024. These may include casual or unintentional violations by players who might be venting frustration or trash-talking. Meanwhile, categories like doxxing or CSAM are extremely rare but taken very seriously.

Appeals

Wildlife Studios launched an enhanced internal complaints and appeals system in 2024 to ensure users have recourse if they believe a moderation action was unjustified. We strive to be fair and transparent, and we recognize that mistakes or misunderstandings can happen. Players have two primary avenues for appeal:

- **Content Removal Appeals:** If a player reported content and we decided not to remove it (i.e. we judged it was not a violation), the reporting player can appeal that decision within 6 months if they strongly disagree and believe we missed something. Essentially, this lets players ask for a second look at content they find illegal or harmful that initially was left up.
- **Sanction Appeals**: If a player received a punishment (such as a chat ban or account suspension) and they think it was a mistake, they can submit an appeal explaining their case. Our team will then review the context and evidence again.

By the end of the year, we recorded 4,804 formal appeals through our system. After thorough re-examination, our moderators overturned the initial decision in 114 of those appeals.

Law Enforcement

Wildlife Studios received 0 orders from government agencies and courts to remove content during the reporting period.

Internal Complaints

Wildlife Studios received 0 internal complaints about content moderation decisions during the reporting period.

Enforcement Trends

Throughout 2024, we observed a steady improvement in our moderation effectiveness. Thanks to better detection algorithms and community reporting, the rate of enforcement increased modestly each quarter. For instance, actions against cheating programs spiked mid-year after we implemented new detection software, and content removals for hate or self-harm speech also rose following policy updates spurred by emerging online behaviors.

We also fine-tuned our moderation rules as new trends appeared – for example, early in the year “in-app dating” (players using game chats to seek romantic exchanges) was identified as a misuse of our platform; we addressed 13 such cases and clarified our policies to the community.

Overall, our moderation efforts in 2024 kept pace with challenges: we responded rapidly to new forms of toxicity and scaled up enforcement where needed, which helped contain issues before they spread widely.

Temporary Chat Suspension – ~ 892 cases (caused by light sexual, hate, political and religious speech)
Permanent Chat Suspension– ~ 65 cases (caused by Severe Toxicity such as Radicalization and Extremism)
Permanent Account Suspensions – ~ 18 cases (caused by Ban Evasion attempts from prior permanent account suspensions)

Geographical Trends

The largest share of user reports and moderation actions came from North America and Europe, which together accounted for roughly 60% of enforcement cases. European players in particular have become more active in flagging content. North America similarly saw a high reporting rate, especially for harassment and hate speech cases, aligning with broader online trends in those countries. In contrast, Latin America (including our home country of Brazil) and Asia-Pacific regions contributed slightly fewer user reports per capita.

Note: This report is based on data collected by Wildlife Studios. The data may not be exhaustive and may not reflect the full scope of content moderation activity on our platform.

Contact: For any questions or comments regarding this report, please contact legal@wildlifestudios.com.

* * *