Last update: 04/2025

UBISOFT'S PLAYER SAFETY TRANSPARENCY REPORT

The purpose of this document is to present Ubisoft's transparency report, demonstrating the company's compliance with the provisions of Article 15 of Regulation (EU) 2022/2065 of the European parliament and of the council of 19 October 2022 on a Single Market for Digital Services and amending Directive 2000/31/EC (Digital Services Act) (hereinafter "DSA").

Name of the service provider	UBISOFT
Date of the publication of the report	17 april 2025
Date of the publication of the latest previous report	Not applicable
Starting date of reporting period	January 1st 2024
Ending date of reporting period	December 31st 2024

CONTEXT

Ubisoft's mission statement is to create meaningful experiences that enrich players' lives. Making this a reality requires going even beyond the creation of immersive worlds. This involves ensuring these experiences are accessible to as many players as possible, are inclusive, and harness the educational potential of video games. Additionally, it's vital to ensure players have access to the right resources and can tailor their experience to their preferences. And of course, it is about ensuring a safe and enjoyable experience for everyone, whether in solo games or when playing online.

Online gaming is about providing a unique environment where players belong, socialize and collaborate. These connections are key to building a meaningful experience and fueling a sense of community and mutual respect. However, while the majority of players play fair, ensuring consistent positive interactions is challenging. Sometimes the experience can turn sour because of cheating, unfair play, or toxic behavior.

With videogames evolving into global ecosystems, gathering billions of players across the globe, it's never been more critical to ensure positive interactions in the long run. As a videogame creator, we have a duty of care towards our players, a duty to provide everyone with an experience where everyone can be themselves, feel welcome and respected.

Disruptive player behavior is an issue that we take very seriously, but also something that is very hard to solve. Disruptive behavior is inherently human and as such requires nuance to address. Anyone can have a bad day and be out of bounds while playing online once. That still does not make it acceptable, and it will trigger a response, but that does not make them a bad person either.

This deeply human issue is complex and does not have a one-size-fits-all solution. Instead, we believe it requires a global, holistic approach that we will explain in the next pages of this report. We strongly believe in leveraging technology and innovation to continue to improve in identifying, contextualizing, and preempting disruptive behavior and in actively protecting our communities through a range of measures. Most of all, we believe that sanctions are not and should not be our only lever and that education and positive reinforcement are critical components of creating safe and inclusive communities.

Today, we release this report covering 2024 and to be updated annually, in compliance with article 15 of the Digital Service Act. With this document, we wish to openly communicate about this topic, better explain our efforts in addressing this issue, and continue raising awareness. Our goal is to pave the way for more positive, inclusive, and rewarding experiences.

INTRODUCTION

In its best and truest form, interactive online gaming invites players from around the world to engage, compete and communicate around a common interest. Unfortunately, it only takes a handful of players who exhibit disruptive behavior to ruin the experience for everyone else.

Disruptive behavior is complex and far-reaching - it stretches beyond our Ubisoft community and even beyond the gaming industry. While we are fully committed to protecting our players, we recognize that we cannot tackle this problem alone. Over the past years, we have initiated cross-industry collaborations with key players who are equally concerned about this type of behavior. We have also worked closely with independent experts, and international organizations to design initiatives that will generate significant, sustainable results.

We have made it our mission to create a positive first-time experience, sustain digital trust across all our platforms, foster an inclusive online culture, teach our players what it means to be in good standing and comply with laws and regulations around the world.

As we contemplated and researched ways to uproot disruptive behavior and began systematically eradicating it from our communities, it became very clear that our approach needed to be multi-faceted and centered around the players themselves. The framework we developed is built on three pillars: **prevention**, **detection**, and **intervention**.

Prevention: Openly communicating the rules of the game with our players from the outset and educating them on how disruptive behavior impacts the broader community.

- *Our Code of Conduct:* A clear, values-driven guide that defines expected player behavior, promotes a respectful gaming environment, and ensures safety across the Ubisoft ecosystem.
- *Fair Play Program* : An e-learning program designed by psychologists and digital learning experts to better understand what toxic behavior looks like and the importance of respect and fairness.
- *Good Game Playbook* : Partnering with mental health charity Safe in Our World to create a holistic guide

  designed to help players understand the broader implications of disruptive gameplay on others.
- *Reputation Score Rainbow Six Siege:* A behavior evaluation system for players.

  Our guiding philosophy with the Reputation System itself is that each player should know why they have a particular Reputation Standing, what they can do to modify their playstyle or interactions and improve their position in the Reputation System. We want to ensure that players have the proper tools available to them and the opportunities to make any necessary adjustments. The release of the Reputation System in 2022 marked an essential step for Ubisoft, and with player feedback being a crucial part of our strategy, we are listening to their thoughts and concerns.

Detection: Innovating constantly to make sure we accurately detect, identify and counter disruptive behavior. Content moderation can be both reactive and proactive, depending on the nature of the illegal content identified. Reactive moderation refers to the review and action taken following user reports, while proactive moderation involves fully automated or manual detection of potentially harmful or illegal content before it is reported. In a nutshell, reactive moderation refers to actions taken after content has already been seen and reported, either automatically or manually. On the other hand, proactive moderation in our case is fully automated and happens before the content becomes visible to anyone. Both approaches work in tandem to ensure a safe environment for our players.

- Tools
  - For moderating usernames.
  - For moderating text chats.
- Collaborative Project "Zero Harm in Comms" (with Riot Games):

    - Development of an anonymous database to train proactive AI moderation systems.
    - Deployment of proactive moderation tools capable of detecting and addressing toxic and harmful online behaviors.

Intervention: Taking action to protect the gaming community from disruptive behavior in all its forms.

- Enforcing a code of conduct, with ongoing efforts to make it more accessible and visible to players.
- Implementing a sanctions matrix, an internal document used to determine proportionate penalties for identified toxic behaviors. Sanctions can include in-game penalties, warnings, temporary deactivation of communications, temporary or permanent bans at the game or account level.
- Moderating toxic content through voluntary initiatives proactively (automation) and through human review of player reports, and more.
- Collaborating with law enforcement when necessary.

Through these actions, Ubisoft strives to ensure a safe, welcoming, and enjoyable gaming environment for all players.

PART I – QUALITATIVE INFORMATION

1. Summary of the content moderation engaged in at Ubisoft's own initiative

To effectively moderate content within a gaming community, a combination of automated systems and advanced detection techniques is essential. This integrated approach helps manage inappropriate behavior while enhancing the overall user experience.

As a preamble, it is important to note that Ubisoft provides account-level settings that allow players to restrict chat functionalities across Ubisoft connect. By default, these settings are disabled for minors, and parents or guardians have access to robust parental control tools to manage communication features. These safeguards are part of our broader commitment to fostering a safe and respectful environment for all players.

Moderator reviews play a vital role in content moderation by evaluating reported behaviors and content. Moderators use their expertise to make informed decisions, such as reducing privileges or imposing temporary bans. Human oversight is crucial for addressing complex cases that require contextual understanding, ensuring fair and accurate outcomes.

AI detection offers a scalable solution by automating the identification of problematic behaviors. Using machine learning algorithms, AI can analyze large volumes of data to detect patterns of abusive language, harassment, or other rule violations. This proactive approach helps filter out undesirable content more efficiently.

Profanity filters are another key tool, automatically blocking inappropriate words or expressions. These filters can be customized by humans moderators to align with the community's tolerance levels varying per game, minimizing offensive language in game communication tools and fostering a more respectful environment.

Customer support teams act as the first line of defense when community members report issues. They play a critical role in resolving conflicts, providing explanations, and ensuring rule violations are addressed promptly and appropriately.

Threshold automation introduces predefined limits to trigger automatic actions in response to specific behaviors. For instance, thresholds can be set to alert moderators or enforce automatic sanctions after a certain number of reports or inappropriate actions are logged, streamlining the moderation process.

By combining these tools and strategies, gaming communities can implement a robust, proactive moderation system that ensures a safe and engaging environment for all players.

2. Meaningful and comprehensible information regarding content moderation engaged in at the Ubisoft's own initiative

Ubisoft employs advanced detection methods to identify and address inappropriate behavior on its platforms, combining automated systems with human oversight for a comprehensive approach. Automated systems analyze player communications, such as usernames, chats, and voice interactions, to detect abusive language, harassment, or prohibited content using machine learning algorithms trained on anonymized datasets. These systems are designed to adapt to evolving behaviors, with real-time detection methods regularly refined for accuracy. Recognizing the importance of context, Ubisoft trains algorithms to identify problematic behaviors based on the environment of the discussion, ensuring nuanced and accurate moderation. This context is regularly updated to reflect societal changes and current events, maintaining the system's relevance and effectiveness.

All moderation efforts are guided by a clear code of conduct and detailed terms of service, which define acceptable behaviors and community standards. When inappropriate behavior triggers a sanction, actions are executed in alignment with a predefined banning matrix. This ensures that sanctions and responses are proportionate, consistent, and based on the severity and frequency of the violations.

Human moderators play a critical role in reviewing flagged content, applying contextual judgment, and aligning decisions with these established guidelines.

This structured approach ensures that detection and enforcement methods remain fair, effective, and aligned with regulatory and community expectations.

3. Qualitative description of the automated means

Ubisoft's automated systems leverage natural language processing (NLP) and machine learning to detect inappropriate behaviors, such as abusive language and harassment, across usernames, text chats, and voice interactions. Trained on anonymized datasets, these algorithms are designed for accuracy and adaptability, with a focus on contextual understanding to ensure nuanced detection. Regular updates reflect evolving societal contexts, enabling scalable, precise, and responsive moderation.

4. Qualitative description of indicators of accuracy and possible rate of error of automated means

To evaluate the accuracy and potential error rate of automated detection systems, several qualitative indicators and metrics are used.

ACCURACY INDICATORS

Periodic random message sampling: Random samples of messages are periodically reviewed to validate the accuracy of the automated solution. This helps ensure that the system is correctly identifying and categorizing messages.

Total accuracy rate: The main indicator of accuracy is calculated as the total number of correctly detected message statuses divided by the total number of messages. This provides an overall measure of how well the system is performing.

Feedback from customer support: Feedback is gathered from the customer support team by analyzing player tickets related to incorrect detections. This human review process helps identify and understand specific cases where the system may have failed, contributing to system improvements.

ERROR RATE INDICATORS

To assess the rate of error, two primary metrics are used:

False positive rate (FP): This is calculated as the total number of messages that were incorrectly removed or flagged (i.e., they were actually safe but marked as toxic) divided by the total number of messages. The goal is to minimize this rate to avoid unnecessarily blocking safe content.

False negative rate (FN): This measures the total number of messages that were kept but should have been removed or flagged (i.e., they were actually toxic but marked as safe) divided by the total number of messages. A lower FN rate indicates better detection of harmful content.

ACCURACY AND ERROR RATE METRICS

To further refine accuracy and error measurement, the following metrics are utilized:

Recall (True Positive Rate): Recall measures the proportion of toxic elements that were correctly identified as toxic by the model. The objective is to maximize Recall, ideally reaching 100%, which would mean that all toxic elements were accurately detected.

False Positive Rate (FPR): This metric indicates the proportion of safe elements that were mistakenly flagged as toxic. The aim is to minimize this rate, ideally reaching 0%, to ensure that safe content is not incorrectly blocked.

BALANCING ACCURACY AND ERROR RATES

In practice, the model is carefully tuned to find an optimal balance between Recall and False Positive Rate. This involves maximizing the detection of toxic elements (high Recall) while minimizing the incorrect blocking of safe elements (low FPR). This balance is crucial to maintaining both the effectiveness and reliability of the system.

These qualitative indicators and quantitative metrics collectively provide a comprehensive approach to evaluating the accuracy and error rate of automated detection systems.

5. Specification of the precise purposes to apply automated means

Automated means are applied to achieve key objectives in content moderation, including **real-time detection** of toxic behavior, abusive language, and prohibited content to safeguard the community. These systems are designed for **scalability**, enabling efficient moderation across large volumes of interactions on multiple platforms. By incorporating **contextual assessment**, they analyze the intent and environment of communications to improve accuracy. Automated tools also facilitate **proactive intervention** by triggering sanctions and actions based on predefined criteria, reducing harm and deterring repeat violations. They support moderators by providing **actionable insights** and flagged content, allowing human reviewers to focus on complex cases. Additionally, these systems ensure **consistency and fairness** by aligning enforcement with the code of conduct and banning matrix while continuously analyzing **behavioral trends** to refine algorithms and inform policy updates.

6. Safeguards applied to the use of automated means

Ubisoft ensures the responsible use of its automated moderation systems through a range of safeguards. **Transparency** is prioritized by clearly communicating with players about the purpose and functioning of these tools. **Regular accuracy checks** are conducted to update algorithms and minimize false positives and negatives, maintaining reliability. Automated actions are always **supplemented by human oversight**, allowing for fairness and contextual assessment in complex cases. We ensure that **privacy principles are integrated from the design stage of our tools** through a "privacy by design" approach. Privacy Impact Assessments have also been implemented to ensure the compliance of moderation tools with the GDPR (General Data Protection Regulation). Additionally, Data Processing Agreements have been signed with Ubisoft's external partners when they process data on our behalf. Finally, Ubisoft’s systems that rely on automated means **are designed to comply with legal and regulatory requirements,** including compliance with the Digital Services Act (DSA), ensuring both accountability and alignment with best practices.

PART II – QUANTITATIVE INFORMATION

A - Orders from member state authorities [Art. 15.1 a) DSA]

- **Number of orders received:** 31

Member State Issuing the order	**Type of illegal content**	**Reception date**	**Acknowledgment date**	**Response date**	**Median time to respond**
Germany	German Criminal Code, Child Pornography	March 17th, 2025	March 17th, 2025	March 20th, 2025	3 days

Germany	German Civil Code owning juvenile pornographic files	February 18th 2025	March 3rd 2025	March 6th 2025	16 days
Germany	This act violates Section 176 a (Sexual abuse of children without physical contact with the child) of the german criminal code (StGB).	February 13th 2025	February 13th 2025	February 17th 2025	4 days
Germany	DSA / Threat	February 13th 2025	February 13th 2025	February 14th 2025	1 day
Germany	German Criminal Code 263a (fraud)	January 24th 2025	January 29th 2025	January 31st 2025	7 days
Germany	Violence against police officer with TiktoK account	January 23rd 2025	January 29th 2025	January 31st 2025	8 days
France	Judicial req.	January 21st 2025	January 23rd 2025	January 28th 2025	7 days
Germany	Minor protection	January 14th 2025	January 16th 2025	January 20th 2025	6 days
Germany	Account Fraud	January 2nd 2025	January 6th 2025	January 9th 2025	7 days
Germany	Threatening commission of serious criminal offence	December 23rd 2024	January 2nd 2025	January 9th 2025	17 days
Germany	Hate speech	December 9th, 2024	December 13th, 2024	December 13th, 2024	4 days
Germany	Credit card fraud – telephone purchase	December 4th, 2024	December 4th, 2024	December 10th 2024	5 days
Germany	Extorsion and ransomware	November 20th, 2024	November 29th, 2024	December 3rd, 2024	13 days

Italy	Illegal access	November 18th, 2024	November 20th, 2024	November 20th, 2024	2 days
Germany	Distribution, acquisition and possession of child pornographic content	October 28th, 2024	October 31st, 2024	October 31st, 2024	3 days
Poland	Illegal access to multiple accounts	October 15th, 2024	October 28th, 2024	November 6th, 2024	22 days
Germany	Data Spying	October 10th, 2024	October 29th, 2024	October 29th, 2024	19 days
Germany	Suspect of terrorism and crime against humanity	September 5th, 2024	September 11th, 2024	September 11th, 2024	6 days
Germany	Suspect of terrorism	September 7th, 2024	September 9th, 2024	September 9th, 2024	2 days
Germany	Use of symbols of unconstitutional and terrorist organizations	August 16th, 2024	September 4th, 2024	September 4th, 2024	18 days
Germany	Data spying (illegal access to multiple accounts)	August 13th, 2024	September 4th, 2024	September 4th, 2024	21 days
Germany	Illegal access to an account	August 16th, 2024	August 21st, 2024	August 27th, 2024	11 days
Poland	Illegal access to information	July 19th, 2024	July 23rd, 2024	July 25th, 2024	6 days
Germany	Computer fraud: unauthorized registration (connection)	July 16th, 2024	July 29th, 2024	October 4th, 2024	80 days
Germany	Online Hate Crime	July 12th, 2024	July 17th, 2024	July 17th, 2024	5 days

Germany	Insult to the Minister of Interior of M-V, Germany	June 7th, 2024	June 24th, 2024	July 1st, 2024	24 days
Germany	Fraud	June 13th, 2024	June 24th, 2024	June 24th, 2024	11 days
Germany	Computer fraud and Data Spying	May 20th, 2024	May 20th, 2024	June 12th, 2024	23 days
Germany	Dangerous Bodily Harm	May 20th, 2024	May 20th, 2024	June 4th, 2024	15 days
Germany	Computer fraud	May 15th, 2024	May 20th, 2024	May 20th, 2024	5 days
Germany	Drug-related crime	March 21st, 2024	March 21st, 2024	May 22nd, 2024	62 days

B. Notification and actions of allegedly illegal content [Art.15.1 b) DSA]

Notes

- Regarding the number of notifications received, the same content can be reported multiple times. After review, a piece of content may not be confirmed as problematic. Therefore, the number of user reports should not be interpreted as equivalent to the number of confirmed violations.
- Ubisoft did not receive any report submitted by trusted flaggers as defined under Article 22 of the DSA.
- We also want to emphasize that while players may report the same content several times or occasionally misclassify it, we are committed to maintaining a safe and healthy ecosystem for our communities across our games.
- **Total number of notifications received regarding illegal content, via the customer support website: 14 029**
- **Number of notifications from trusted flaggers: 0**
- **Notification handling:** Depending on the type of game, player reports are processed automatically in-game based on the category, with moderators handling the most frequently reported cases and support teams managing reports made outside the game. It's worth noting that most report categories are generally focused on toxicity for in-game processing. It is also important to highlight that there is a possibility to flag illegal content directly on the CRC platform via the Help tool.
- **Actions taken regarding illegal content following the reports:**

Type of Allegedly illegal content	Action Taken	Based on legislation? Y/N	Based on Terms & Conditions? Y/N	Processed automatically ? Y/N	Action Count

Non-consensual behaviour	Remove Content	y	y	N	33755
Illegal or harmful speech	Game Ban	y	y	N	19397
Illegal or harmful speech	Mute	y	y	y	15031
Non-consensual behaviour	Game Ban	y	y	N	14016
Non-consensual behaviour	Game Ban	y	y	y	8466
Illegal or harmful speech	Block Feature	y	y	y	1671
Pornography or sexualised content	Game Ban	y	y	N	820
Non-consensual behaviour	Mute	y	y	y	811
Illegal or harmful speech	Warning	y	y	y	437
Protection of minors	Remove Content	y	y	N	355
Illegal or harmful speech	Game Ban	y	y	y	206
Pornography or sexualised content	Remove Content	y	y	N	186

Non-consensual behaviour	Warning	y	y	N	70
Illegal or harmful speech	Remove Content	y	y	N	44
Illegal or harmful speech	Warning	y	y	N	38
Threats, violence, self-harm or suicide	Escalate to Social Media Platform	y	y	N	38
Illegal or harmful speech	Mute	y	y	N	32
Non-consensual behaviour	Escalate to Social Media Platform	y	y	N	28
Pornography or sexualised content	Warning	y	y	N	20
Non-consensual behaviour	Escalate to Law Enforcement	y	y	N	14
Threats, violence, self-harm or suicide	Escalate to Law Enforcement	y	y	N	4
Protection of minors	Escalate to Internal Security	y	y	N	2
Protection of minors	Escalate to NMEC	y	y	N	1

Protection of minors	Escalate to Internal Security	y	y	N	1
Non-consensual behaviour	Escalate to Internal Security	y	y	N	1
TOTAL					95 444

C. Content Moderation Initiated by the Provider [Art.15.1 c) DSA]

Proactive automated moderation for text chat and username: 16 212 753

- **Use of automated tools: 8 930 130**
    - Tool description **: Username moderation**
    - Specific objectives: Automating the username moderation protects players from seeing usernames that contain toxic or shocking content and encourages them to choose safe usernames.
    - Accuracy and error rate: To measure the accuracy of the systems, we use 2 metrics: the Recall, and the False Positive Rate.

      The Recall (or True Positive Rate) measures the proportion of toxic elements that were rightly detected as toxic by the model. We want to maximize this indicator, the best possible value being 100% of the toxic elements that got rightly caught. The False Positive Rate, on the other hand, measures the proportion of safe elements that the model mistakenly detected as toxic. We want to minimize this indicator, the best possible value being 0% of the safe elements that got mistakenly blocked. In practice, we tune the model to find the optimal balance between the Recall & the False Positive Rate, so we detect as many toxic elements as possible, while limiting the errors.
    - Safeguards: If the player disagrees with the sanction, they can appeal by contacting Ubisoft's Customer Support. The case is then reviewed by a human agent, who can lift the sanction if it was unjustified.
- **Training and support for human moderators:** We frequently use human moderators' feedback to spot model errors and directly use them to re-train the model, so it improves specifically on these mistakes.

Content Type	**Restriction Type/**	**Detection Method**	**Number of Actions**
Pornography or sexualised content	Block	Account Username	3,876,783
Illegal or harmful speech	Block	Account Username	2,365,717
Non-consensual behaviour	Block	Account Username	1,715,626

Risk of public security	Block	Account Username	379,879
Protection of minors	Block	Account Username	370,994
Threats, violence, self-harm or suicide	Block	Account Username	221,131
TOTAL			8 930 130

- **Use of automated tools: 7 282 623**
  - Tool description: **Profanity Filtering - In-Games text chat moderation**
  - Specific objectives: The purpose of the feature is to protect players from highly toxic messages and to motivate players to improve their communication.
  - Accuracy and error rate: Periodical random message sampling and going through to validate the accuracy of the solution.

    The main accuracy indicator will be the total number of correctly detected message status divided by the total number of messages.
  - Safeguards: Automatic kill switch will be put in place to stop impacting players in case of bug or reliability issues. As a failsafe, the new system will be put on top of the previous one using RegEx. If the player disagrees with the sanction, they can appeal by contacting Ubisoft's Customer Support. The case is then reviewed by a human agent, who can lift the sanction if it was unjustified.
- **Training and support for human moderators:**

  We frequently use human moderators' feedback to spot model errors and directly use them to re-train the model, so it improves specifically on these mistakes.

Content Type	**Restriction Type/**	**Detection Method**	**Number of Actions**
Illegal or harmful speech	Block	Text Chat	6,307,427
Pornography or sexualised content	Block	Text Chat	747,419
Non-consensual behaviour	Block	Text Chat	207,340
Risk of public security	Block	Text Chat	9,243
Protection of minors	Block	Text Chat	8,618

Threats, violence, self-harm or suicide	Block	Text Chat	2,576
Total			7 282 623

D. Complaints Received [art15. 1. d) DSA]

- **Total number of complaints:** 198 533
- **Basis of complaints:**

  When your account is permanently locked for toxicity, cheating or fraud, you will be unable to access any online games or services, or login pages on Ubisoft websites.

- **Decisions taken:**

Complaint Category	Number of Complaints	Decisions Made	Number of Decisions overturned	Median Processing Time (days)
Permanent locks	3,521	Lifted	0	
Permanent locks	195,012	Upheld	0	