# Regulation (EU) 2022/2065 Digital Services Act Transparency Report for Habbo

**Name of service provider**: This Report is published by Sulake Oy. in relation to our services, in accordance with the transparency reporting requirements under Articles 15 and 24 of the European Union's Digital Services Act (Regulation (EU) 2022/2065) ('DSA'). In the context of this report, "services" refers to the following, offered by Sulake Oy in the European Union: Habbo.

**Date of the publication of the report**: 17 April 2025

**Starting and ending date of reporting period**: The Report contains information for a reporting period from 17 February 2024 to 17 February 2025.

**Overview**

| | |
|---|---|
| Chat lines and other content total | ~2.8 billion |
| Notices (reported by users) | 180256 (~0.006 % of total) |
| Flagged by own initiative | 3367611 (~0.12 % of total) |

The content classifications in this report are based on individual chat lines and do not represent entire conversations, users, or posts. This granularity, while valuable for moderation and safety efforts, can result in high volume counts that may appear misleading without proper context. Our platform does not allow users to share images, videos, or files of any kind. Additionally, Habbo includes role-playing and fictional storytelling features that can influence language use in chat. These role-play interactions may include simulated scenarios, often between fictional characters, which our system may classify under sensitive categories—despite being part of a moderated, fictional in-game experience.

It's also important to note that our automated moderation tools may operate by detecting specific keywords or phrases, which may trigger flags regardless of the broader conversational context. For example, the use of a sensitive term during a news-related discussion or educational comment (e.g., referencing a serious issue like abuse in a headline or debate) may be interpreted similarly to actual violations such as inappropriate roleplay. In such cases, the system cannot always distinguish between harmful use and contextually appropriate mention. This may lead to over classification in some categories, even when the actual intent or content is benign.

Our moderation systems are designed to maintain a safe and age-appropriate environment for all users. We continuously update and refine these systems to accurately identify and respond to potential violations while avoiding overinflated interpretations of the data.

### Summary of the content moderation engaged in at the provider's own initiative

Habbo maintains a clear set of Community Guidelines and other applicable platform rules that define acceptable behavior and content on its services. As part of its content moderation efforts, Habbo proactively employs a combination of automated systems, human moderation, and user-driven reporting tools to maintain a safe and respectful environment. This section focuses on the content moderation activities undertaken by Habbo on its own initiative, including automated filtering, human review processes, and the enforcement of community standards across the Habbo platform.

- **Community Guidelines & Rule Enforcement:** Clearly defined rules educate players on appropriate behavior.
- **Automated & Human Moderation:** AI-driven content filters work alongside human moderators for accuracy.
- **Reporting & Appeals System:** Users can report misconduct, track report statuses, and appeal moderation decisions.
- **Tiered Sanctions:** Progressive penalties discourage rule-breaking while allowing users to correct behavior.

**Automated Chat Filters (Bobba Filter)** – Habbo has a built-in chat filter that automatically replaces inappropriate or offensive words with "bobba" to prevent toxic behavior.

In *Habbo*, moderators engage in various forms of content moderation at their own initiative to maintain a safe and enjoyable environment for players. Some of these include:

**Moderation Team (Live Moderators)** – A dedicated team of moderators actively monitor in-game chats, user interactions, and reports. They can mute, kick, or suspend accounts violating community guidelines.

**Safety Lock & Account Protection** – To prevent unauthorized access and hacking, Habbo provides security features like Safety Lock, which requires additional authentication for certain actions.

**Room Moderation & Mute Options** – Room owners can set chat rules, mute users, and

remove troublemakers from their rooms.

**User Reporting System (Notice and Action Mechanism)** – Players can submit **notices** about abusive behavior, inappropriate content, or suspicious activity, which are then reviewed by Moderation system and Habbo staff.

**AI Monitoring & Behavior Tracking** – Advanced systems detect and flag harmful content, including scams, threats, or bullying, allowing moderators to take proactive action.

In *Habbo*, self-help moderation tools empower users to manage their own experience and ensure a safer environment. Some of these self-help features include:

1. Ignore Feature – Players can mute or ignore other users to stop seeing their messages and avoid harassment or unwanted interactions.

2. Room Rights & Controls – Room owners can moderate their spaces by setting permissions, kicking disruptive users, and applying word filters.

3. Privacy Settings – Users can adjust who can send them friend requests, trade with them, or enter their rooms, reducing unwanted interactions.

4. Safety Tips & Guides – Habbo provides built-in safety guidelines and help articles to educate players on how to protect themselves from scams, bullying, and inappropriate behavior.

5. Habbo Guardians – Experienced players volunteer to assist in reporting issues and helping newcomers understand the rules.

## Meaningful and comprehensible information regarding content moderation engaged in at the provider's own initiative

Habbo engages in content moderation on its own initiative to ensure a safe, respectful, and age-appropriate environment for its users. This includes the use of automated filtering systems and a dedicated team of human moderators who proactively monitor, review, and remove content that violates Habbo's Community Guidelines. The moderation process addresses offensive language, inappropriate behavior, scams, and safety risks without requiring external notification. Additionally, Habbo empowers users with in-game reporting tools and room moderation controls to further support a safe community experience.

Habbo's moderation efforts include:

- **Automated filtering** of offensive language, hate speech, spam, and sharing of personal information. Inappropriate messages are replaced with "bobba" or blocked before being sent.
- **Warnings & Temporary Sanctions:** Minor offenses trigger a warning or temporary mute.
- **Human moderation** by a multilingual team who review flagged content and appeal user reports.
- **Warnings & Temporary Sanctions:** Minor offenses trigger a warning or temporary mute.
- **Suspensions & Termination:** Repeated or severe violations (e.g., threats, scams, sexual content) lead to bans.
- **User-driven reporting tools** that empower players to report misconduct.
- **Progressive Gradual sanctions system**, from warnings to termination.

  The biggest change in comparison to the old moderation system is that most of the sanctions are now gradual. The new suspension chart progresses like so:

  - Alert
  - 1 hour mute
  - 18 hour suspension
  - 7 days suspension
  - 30 days suspension I
  - 30 days suspension II
  - Termination
- **Room moderation tools** empowering users to mute, kick, or suspend disruptive players. Users with room rights can mute, kick, or termination disruptive players.
- **Decision Transparency:** Players can see if their reports were handled manually or automatically through the **"My Report Status"** feature.

Additionally, Habbo emphasizes **user education and self-protection** through safety guidelines, an **Ambassador Program**, and direct parental support.

### Qualitative description of the automated means

Habbo's automated moderation systems operate based on **predefined rules, thresholds, and behavioral patterns** that assess the severity and frequency of potential violations. These parameters help determine when automated actions, such as termination or chat restrictions, are applied.

#### Conditions for Termination and Penalties

Automated penalties may be triggered when: Repeated Rule Violations, High-Severity Offenses, Bypassing Moderation Filters,Scam & Fraud Patterns, Bot-Like Behavior, Self-Harm or Dangerous Behavior content related to self-harm or endangerment may trigger interventions, including support notifications, chat restrictions, or account suspension for severe cases.

## Escalation & Review Process

Automated actions are **tiered**, meaning minor violations **start with warnings or temporary restrictions**, escalating to **longer suspensions or permanent bans (termination)** for repeated offenses. **User Appeals**: Users can appeal automated bans (suspensions), and flagged cases are reviewed manually by Habbo moderators to **correct potential false positives**. **Ongoing System Calibration**: If a high number of appeals are successful, moderation parameters are reviewed and adjusted to improve accuracy.

Habbo employs automated moderation technologies on its own initiative to proactively detect, block, and remove content that may violate its Community Guidelines and other applicable platform policies. These automated systems are designed to identify harmful, inappropriate, or unsafe content before it is widely disseminated across the platform.

The automated moderation tools used by Habbo primarily operate on in-game chat messages, user-generated content such as room names and descriptions, and user profiles. These systems analyze text-based content to assess the likelihood that it contains offensive language, hate speech, personal information, or attempts to bypass moderation filters through obfuscation techniques.

The primary purpose of these automated systems is to enable real-time detection and prevention of content that may be harmful to the community, including offensive speech, scam attempts, spam, and bot-generated activity. In certain cases, when the automated systems identify a high probability of a policy violation, the content may be immediately filtered (e.g., replaced with a placeholder term like “bobba”) or trigger automated penalties such as chat restrictions or suspension.

Habbo's automated moderation systems include:

- **AI-powered chat filters** that censor offensive or inappropriate language.

- **Pattern recognition algorithms** that detect obfuscated offensive terms, symbols, or disguised words.
- **Trade & scam prevention tools** to detect fraudulent behavior.
- **Anti-bot systems & CAPTCHA implementation** to prevent bot-related abuse.
- **Automated penalties** like chat restrictions and suspension for repeated violations.

## Qualitative description of indicators of accuracy and possible rate of error of automated means

Habbo monitors the accuracy of its automated moderation systems through a combination of human review, user feedback, and internal audits. Key indicators of accuracy include the rate at which automated actions are overturned following user appeals or manual review. Regular audits of moderation decisions help identify patterns of false positives and false negatives, allowing Habbo to improve its systems over time. Safeguards such as human oversight, an appeals process, and moderation quality checks are in place to minimize errors and ensure compliance with Habbo's Community Guidelines.

While no specific statistical accuracy rates are provided, Habbo acknowledges that:

- The **automated filters are updated regularly** based on Notices and behavior trends.
- **Mapped characters, delimiters, and word matching techniques** improve detection accuracy.
- To minimize errors, human moderators are involved in reviewing flagged content, and an **appeals system** is available to users.

## Specification of the precise purposes to apply automated means

Habbo applies automated moderation systems to proactively detect, block, and limit the spread of harmful, inappropriate, or rule-breaking content. The primary purposes of these automated means are to prevent the use of offensive language, hate speech, personal information sharing, scams, and bot-related abuse; and to enforce community guidelines efficiently at scale. Automated tools are also used to trigger alerts for repeated misconduct and apply chat restrictions or suspension where necessary, helping to maintain a safe and positive environment for users.

Automated moderation in Habbo is used to:

1. **Filter and block offensive, explicit, or harmful language.**
2. **Detect and prevent scam attempts and fraud.**
3. **Identify and act against bot accounts or automated abuse.**
4. **Detects attempts to bypass word filters through obfuscation.**
5. **Trigger alerts and automatic penalties for repeated rule-breaking behavior.**
6. **Detecting and addressing context related to self-harm to provide appropriate support.**

## Safeguards applied to the use of automated means

When automated systems flag content or behavior, Habbo's moderation team has the ability to manually review and reassess these decisions. The platform operates an **appeals system**, allowing users to challenge automated moderation actions. Appeals and human moderation outcomes are regularly analyzed to identify and correct potential false positives or inaccuracies in the automated systems.

In addition, Habbo uses **transparent indicators** such as the "My Report Status" feature, which informs users whether moderation actions were taken automatically or manually. Abnormal patterns in automated enforcement actions, including a high number of overturned decisions following appeals, trigger internal reviews by Habbo's moderation and engineering teams.

### Identified Risks Addressed by Automated Moderation

Habbo's automated moderation systems are designed to mitigate a range of risks that could negatively impact user safety, platform integrity, and compliance with legal obligations. These include:

- Risk of exposure to harmful content: Automated filtering prevents users from encountering offensive language.
- Risk of scams and fraud: Automated scam detection tools help prevent phishing attempts, account takeovers, and fraudulent in-game transactions.
- Risk of child exploitation and grooming: AI-driven moderation assists in detecting inappropriate interactions.
- Risk of bot-related abuse: Automated systems identify and restrict bot accounts used for spam, unfair trading practices, and automated harassment.

**Habbo applies multiple safeguards:**

- **Regular AI system updates** to improve detection accuracy.
- **User appeals system** to challenge automated moderation decisions.

- **Manual review of appeal cases** to prevent unfair penalties.
- **Transparency through “My Report Status” feature** so users can track if their report was handled automatically or manually.
- **Training of moderators** to oversee automated moderation outcomes and address false positives.