Regulation (EU) 2022/2065 Digital Services Act Transparency Report for Hotel Hideaway

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: Hotel Hideaway.

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 billion
Notices (reported by users)	644786 (~0.03 % of total)
Flagged by own initiative	9667527 (~0.48 % 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, Hotel Hideaway 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 terms 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 providers' own initiative

Hotel Hideaway provides several different types of moderation depending on the severity of the offence.

Self Help

Users of Hotel Hideaway have available several different options available to them in order to block other users or speech that they find inappropriate.

- **Block** - This allows users to select specific people that they would like to block. This prevents them from seeing the person in the game rooms, or any chat that person may send to them.
- **Room Block and Kick** - Users are able to block and kick users from their private rooms if they wish to do so.
- **Filtered text options** - users are able to choose if they want filtered text to be restricted further.

Provider Moderation

Hotel Hideaway uses a mix of different kind of moderation in order to ensure that the community can enjoy a safe environment to play in. Sanctions are accumulative and end in a permanent ban. When warranted, an instant permanent ban is given.

- **Filtered Text** - All text is run through our filter, blocking toxicity from the users.
- **Self-detected moderation** - When Hotel Hideaway detects specific wording, we will ban users who break our policies and community rules.This is mostly done by automated means and reviewed by human moderators.
- **User-reported moderation** - Users are able to report on all text based user generated content to Hotel Hideaway for review. A mix of automation and manual moderation is used, and users are able to appeal if they feel like their report was not handled correctly.

User Moderation

Hotel Hideaway has a group of dedicated users, Ambassadors, who also

- **Mute** - Ambassadors are able to mute others in both public and private rooms so that users who are not following the terms and conditions of the game are blocked from speaking for a short period.
- **Kick** - Ambassadors are able to kick disruptive users out of private rooms so that they cannot disrupt other users experience.

Sanctions

The gradual sanction system is as follows, and is listed in this FAQ.

- 15 Min Mute
- 60 Min Mute
- 2 Hour Mute
- 12 Hour Ban
- 24 Hour Ban
- 48 Hour Ban
- 72 Hour Ban
- 7 Day Ban
- 30 Day Ban
- 61 Day Ban
- Permanent Ban

We have been adjusting our moderation system during this year to further comply with all DSA regulations that may have some effect on numbers shown.

Meaningful and comprehensible information regarding content moderation engaged in at the providers' own initiative

Detection and Moderation methods

We use a combination of automated keyword detection, user reporting tools, and real-time human moderation to identify content that is potentially illegal or incompatible with our Terms and Conditions. In-game chat and content interactions are monitored using filters for flagged terms and behavior patterns, while players can also report content or behavior directly within each room.

Actions Following Detection.

There are several actions that are taken once it has been determined that content has breached or terms and conditions or that the content flagged is illegal.

- **Removal of name text**. If a user name, group name or room name has been determined to breach the terms and conditions, or is illegal the name in question is removed. In the case of a user name, this means that the avatar is deactivated, or, if the only avatar on the account, the name is changed. Group names and Room names are also changed to a generic name.
- **Removal of profile, motto or room and group description**. If a profile, motto, room or group description has been determined to breach the terms and conditions, or is illegal, the text in question is removed. Depending on the severity of the offence, the user what wrote the text can also be sanctioned.
- **Sanction of avatar**. If a person has been determined to breach the terms and conditions, or has done something illegal, the whole account of the user in question is removed. This also removes any rooms that they may have published at the time. Most cases are sanctioned on a gradual system, however in extreme cases a permanent ban is given directly.
- **Filtered text**. In certain cases text is filtered before users are even able to see it.
- **Reporting to authorities**. In cases where there is considered to be a risk to other people, Hotel Hideaway will report an avatar to the authorities local to where the user is located.

Exposure to Illegal or Incompatible Content.

Gameplay takes place in rooms that have a room limit of 45 players per room, so exposure to any illegal content is limited in scale by design. During the reporting period, content reported as illegal was estimated to account for approximately 1.04% of all user reports.

Estimated Reach Prior to Moderation:

As each room is limited to 45 users per room, illegal or incompatible content would have been seen by a maximum number of 45 users before a moderation action took place. As the text disappears within a few seconds from the room, exposure is further limited. In most cases, action on the avatar that was typing the text reported as illegal was taken within hours of detection or reporting if warranted. This was checked by a human moderator.

Qualitative description of the automated means

Hotel Hideaway uses the product Active Fence to cover its moderation. Active Fence is a system that uses a combination of AI, machine learning and human review to moderate.

Automated Content Moderation Description

Active Fence is used with Hotel Hideaway to do the following:

- **Filter harmful language -** This is based both on keyword lists, as well as the probability of the text being offensive. Filtered text appears as a line of the following character: ‘*’
- **Prevent inappropriate usernames, room names and group names from being created** - Based on a mix of keyword lists and the probability of being offensive, if a user creates a name that the system deems inappropriate, they are prompted to create a different name.
- **Prevent group, room, profile and motto descriptions from being created** - Based on a mix of keyword lists and the probability of being offensive, if a user creates a name that the system deems inappropriate, they are prompted to create a different description.
- **Sanction of reported chat** - When the automated moderation detects a reported chat as having a high probability of an offense, the reported user will be sanctioned on the gradual sanction system at the level of sanction their specific avatar is currently at.
- **Detection of avatars chat -** When the automated moderation detects a chat (reported or not) as having a high probability of breaching the terms and conditions of the game, the reported user will either be sanctioned for specific offenses set by Hotel Hideaway, or flag the chat to be checked by a Human Moderator.

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

Hotel Hideaway's goal is to maintain a safe and respectful environment without over-restricting user expression. While we rely on automated tools for scalable moderation, we recognize their limitations and actively support them with human oversight, ongoing review, and system updates. We are committed to improving accuracy and minimizing error rates through continuous feedback and adjustment. We use both human moderators and automated moderation to do this.

Automated moderation is handled by a third-party provider, whose tools are designed to detect and flag content that may violate our community standards—such as hate speech, threats, harassment, or spam.

Assessment of accuracy

We do not currently receive detailed accuracy metrics (such as precision, recall, or error rates) from our third-party provider. As a result, we are not able to directly report specific indicators like true positive or false negative rates.

Human oversight and continuous improvement

Our in-house moderation team supports and complements the automated system. Human moderators review all appealed cases, and we have alerted the third party to any mistakes that we see. Our moderators also regularly check automatically sanctioned and non-sanctioned CFH to ensure that the system is working properly. While we cannot modify the underlying model, we can adjust some filtering thresholds or keyword settings, and we report performance issues to the provider as needed.

Specification of the precise purposes to apply automated means

Hotel Hideaway utilizes automated means—specifically through a trusted third-party AI moderation provider—to support our content moderation processes. The use of automation is used both in the initial detection and classification of potentially harmful or inappropriate content in real-time, such as hate speech, harassment, threats, and sexually explicit material, within in-game chat and user-generated content. Our AI moderation is also used as an extra layer of security on top of

The precise purposes for applying automated means are:

- **Real-time risk detection**: To identify high-risk content that may negatively impact player safety or violate community standards.

- **Support human moderation**: To prioritize and flag content for human review based on severity, allowing our moderation team to focus on the most pressing issues.

- **Enhance response time**: To minimize exposure to harmful content by triggering early warnings or temporary, proportionate sanctions based on predefined rules.

- **Enable consistent enforcement**: To help maintain consistency in applying our community guidelines across our player base.

It is important to note that while some user data is shared with the third-party provider; we anonymize as much as we can and data is only kept for a maximum of 30 days within the third party provider's system. Additionally, our moderation system operates on a **gradual sanction framework**, starting with warnings for minor infractions and escalating only in cases of repeated or severe violations, with human oversight at key decision points.

Safeguards applied to the use of automated means

Hotel Hideaway has limited direct control over the automation mechanisms used as we use a third party provider for our moderation. However, to safeguard users, we have implemented the following measures:

- **Human review of flagged content:** When possible, content flagged by automated systems is reviewed manually by our team or the third-party provider before any enforcement action is taken. However, due to limited resources and scale, not all flagged content is reviewed by a human prior to action.

- **Appeals process:** Users can appeal moderation decisions, which triggers a manual reassessment to ensure fairness and accuracy.All appeals are looked at by a human moderator.

- **Regular feedback and escalation:** We maintain an active feedback loop with our third-party provider to report false positives or negatives and improve moderation accuracy over time.

- **User reporting tools:** Our platform enables users to report inappropriate content manually, providing an additional layer of oversight beyond automation.

- **Transparency to users:** We notify users when their content is moderated, including the reason and available steps to contest the decision.

These safeguards aim to balance the efficiency of automated moderation with accountability and fairness for our user community.