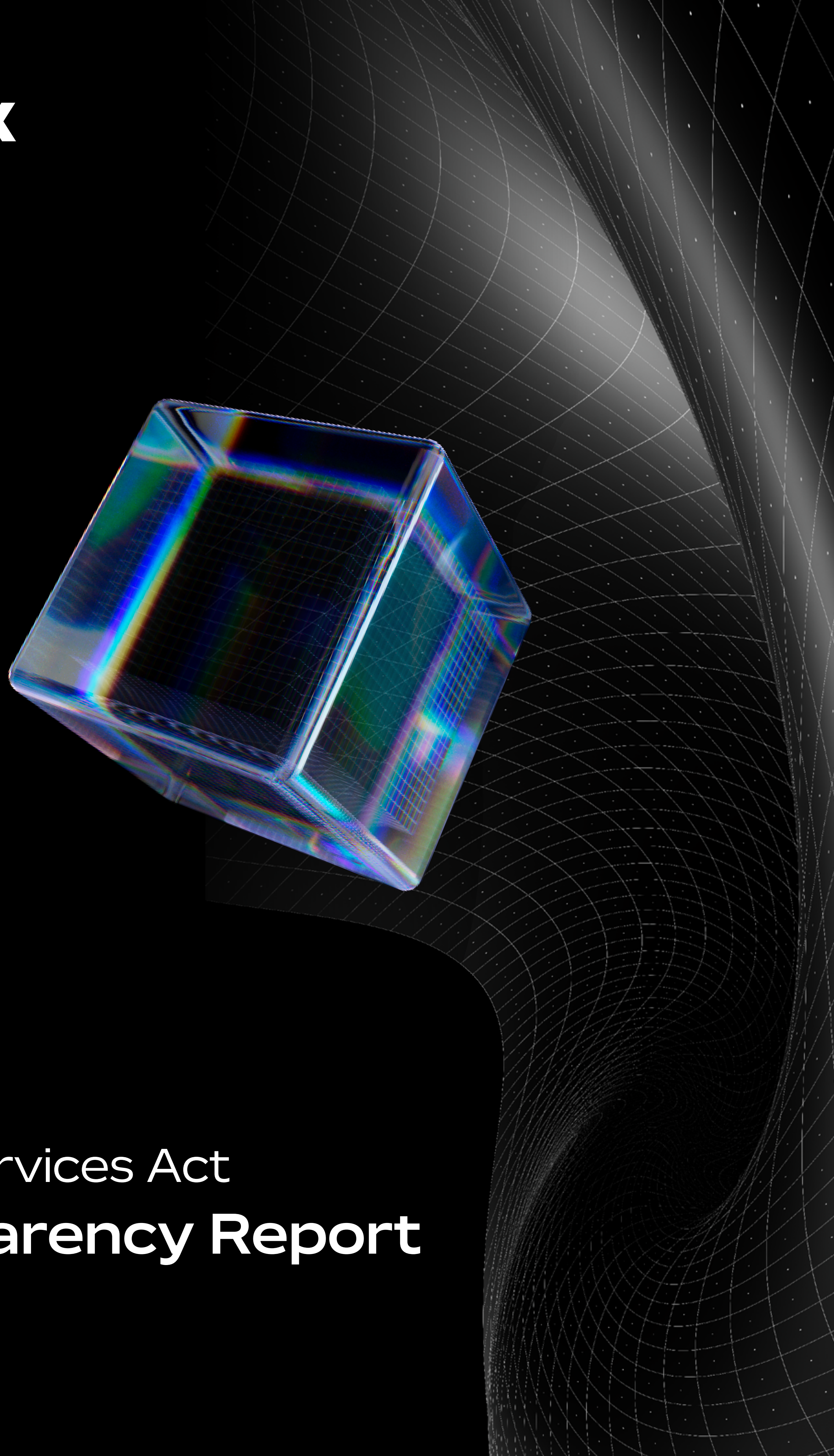
ROBLOX



Digital Services Act

Transparency Report

2024

Roblox is an immersive gaming and creation platform. Every day, millions of users create, play, learn, and connect with each other through online games and other virtual 3D experiences built by a global community of creators.

Roblox has spent almost two decades working to make the platform one of the safest online environments for our users, particularly our youngest users. Our guiding vision is to create the safest and most civil community in the world. As our platform evolves and scales, forging a new future for communication and connection, our investment in preventative measures remains fundamental. With each passing year, we implement new strategies and technology to improve the speed and effectiveness of our safety and moderation systems.

Transparency reports demonstrate our deep commitment to keeping all Roblox users safe and complying with our obligations under the EU's Digital Services Act ("**DSA**"). This document provides additional information about our approach to moderation on Roblox; you can find a machine-readable CSV file of our data on our Transparency Site.

Roblox Terms of Use and Community Standards

Our mission is to connect a billion people with optimism and civility. Roblox's Community Standards, which are incorporated into our Terms of Use, govern user content, and conduct on the platform.

Our Community Standards are broken down into four main categories and cover the following topics:

1. Safety
   a. Child Exploitation
   b. Terrorism and Violent Extremism
   c. Threats, Bullying, and Harassment
   d. Suicide, Self Injury, and Harmful Behavior
   e. Discrimination, Slurs, and Hate Speech
   f. Harmful Off-Platform Speech or Behavior
2. Civility
   a. Real-World Sensitive Events
   b. Violent Content and Gore
   c. Romantic and Sexual Content
   d. Illegal and Regulated Goods and Activities
   e. Profanity
   f. Political Figures and Entities
3. Integrity
   a. Cheating and Scams
   b. Spam
   c. Intellectual Property Violations
   d. Advertising
   e. Roblox Economy

4. Security
    a. Sharing Personal Information
    b. Directing Users Off-Platform
    c. Misusing Roblox Systems

Content moderation practices

Automated and human content moderation

As set out in the CSV file, a combination of automated and human content moderation systems play a crucial role in enforcing our policies by proactively identifying and removing violating content. Content uploaded on the platform such as images undergo a comprehensive review process before it's published, while content such as voice-chat communications are assessed for policy violations in real-time.

We have reported on our own initiative content moderation on tab 4 of the CSV file. 4.4 categorizes the types of violations according to our Community Standards.

Content moderator training and wellfare

Moderators are trained on the implementation of our policies. Training includes coaching on the application and interpretation of Roblox's policies, practice implementing those policies on example sets of potential violations, followed by evaluations. Moderator decisions are regularly evaluated for accuracy and moderators are provided with additional remedial training as necessary.

Roblox supports its moderators with regular well-being check-ins, 24/7 on-demand support from a licensed clinician, peer support groups, wellness workshops, and individual or group counseling sessions. Roblox evaluates these wellness offerings semi-annually for effectiveness and to identify areas for improvement.

Moderation with ML powered AI models

Our ML models detect policy-violating text, images, speech, audio, and 3D content comprehensively across Roblox; this content is reviewed by both our automated and manual systems. These models are trained on Roblox-specific language and abbreviations, and are able to consider the context of potential violations that would likely be missed by other, non-Roblox-specific, ML models.

For text, we use techniques such as hash matching and keyword lists to identify novel violating content, as well as content that's been previously removed from the platform; this helps ensure that once a piece of content is removed once, it won't appear on the platform again.

For visual assets, including avatars and avatar accessories, we use computer vision. One technique involves taking photographs of 3D assets from multiple angles. The system then reviews those photographs to determine what the next step should be. If nothing seems amiss, the item is approved.

For voice chat, we use Automatic Speech Recognition to transcribe voice into text, then use an in-house AI model to classify and detect policy-violating language. This is done in real-time, allowing us to take action such as temporarily suspending or banning users violating policies in voice chats on Roblox.

Responding to user reports

Roblox has a robust reporting process in place to address user reports of policy violations. Users have the ability to easily mute or block other community members and report inappropriate content or behavior on the Roblox application and our website. Roblox also provides users with a direct channel to our Customer Support team to report concerns or other issues.

Enforcement actions

Once a policy violation has been determined, we take action in accordance with our policies. The actions taken can vary based on the severity and impact of the violation, and they can include:

- Warnings
- Removing content
- Roblox account-level or feature-level restrictions
- Reporting users to relevant authorities in cases presenting an imminent risk of harm

In addition to considering the specific violation, we also take into account a user's historical use of the platform and whether they have repeatedly violated our policies. Repeated violations of our policies may increase the severity of the enforcement actions.

DSA obligations

Below we set out some additional contextual information relating to our DSA obligations and some of the data points included in the above-referenced CSV file, for information. Further information about how Roblox is complying with its DSA obligations can be found here.

Illegal content

Content can be reported as illegal using our dedicated illegal content reporting form. The categories at 3.2 of the CSV file are those presented to users completing the form to facilitate our review of the reports. The "other" category is a category we make available to users reporting illegal content who do not consider their report to fall within the categories listed. We review all illegal content reports against our Community Standards first. The vast majority of reports are removed in accordance with our Community Standards and in the reporting period, only one illegal content report was actioned on the basis of local law.

Trusted flaggers

Trusted flaggers accredited in accordance with the Digital Services Act can submit notices of illegal content via our dedicated illegal content reporting form. Reports from trusted flaggers are prioritized by our team of moderators.

Appeals

Whenever we make a decision to remove content on the platform, we notify users of our decision and provide them with an opportunity to appeal by contacting the Roblox Appeals team. Moderators holistically evaluate appeal requests, considering the severity of the violation, the user's reason for appealing, and their behavior on the platform.

EU users can appeal an initial moderation action within a 6-month period. EU reporters that file DSA illegal content reports also have the right to appeal Roblox's decision related to their report.

The median time needed for taking decisions under Art 15(1)(d) was calculated based on the time from receipt of submission to issuance of decision.

Out-of-court dispute settlement

Under the DSA, EU users also have the right to select a certified out-of-court dispute settlement body to resolve disputes relating to a content moderation decision we take, and may also seek judicial redress. EU reporters that file DSA illegal content reports also have the right to appeal Roblox's decision related to their report.

As reported, 34 disputes were submitted to the out–of-court dispute settlement bodies. This figure does not include cases sent to us which were duplicates of existing cases, nor inadmissible cases.