Roblox: The Structural and Legal Controversy Over the Safety of Minors

Rafael Verde Vazquez[1], Jose Angel Juarez Austria[2], Diego Armando Bautista Hernandez[3], Emiliano Sosa Ugalde[4], Jesus Armando Gonzalez Rodriges[5], Roman Resendiz Vargas[6]

[1,2,3,4,5,6] Instituto Tecnológico de Pachuca, Hidalgo, Mexico

***Corresponding Author**

DOI: https://doi.org/10.47772/IJRISS.2026.102500033

Received: 29 October 2026; Accepted: 03 November 2026; Published: 24 November 2026

ABSTRACT

Roblox brings together in one place a store of virtual items, a social network and a system for publishing games created by the users themselves, and a large share of its audience is young. For this reason, this article studies, on the basis of documents, how three parts of the platform relate to one another: its economy, with a virtual currency that is bought at one price and redeemed at a much lower one; the mechanisms that keep players connected, such as limited-time events and daily rewards; and a chat moderation system that reviews messages one by one. Nineteen sources are reviewed and classified according to how reliable their evidence is. As a result, an independent audit of more than two million messages found risky contact with minors (grooming), sexual content and requests to continue the conversation on other apps, all without being blocked by the filters. In addition, sanctioned users change the way they write in order to evade the filter. On the other hand, there are institutional signals pointing in the same direction, such as more than 65,000 reports to NCMEC in 2025 and several lawsuits, although their level of verification differs. By contrast, the evidence on design aimed at retaining players is the weakest. The article also examines the case of the YouTuber Schlep, banned from the platform after exposing alleged abusers: the company justifies the ban on safety grounds, while its critics suspect it is protecting its image, and the available evidence does not allow a decision between the two versions. Consequently, the article explains why word-based filters are insufficient and proposes design measures to improve them. It concludes that the evidence justifies constant vigilance, but does not by itself establish who is responsible.

Keywords: Roblox; player retention; virtual economy; online child safety; content modera tion; grooming; lawsuits against platforms.

INTRODUCTION

First of all, Roblox is not an ordinary video game. It is a platform that brings together millions of “experiences” created by users, with text and voice chat, its own currency and a store of virtual items. In fact, a recent academic audit places it among the largest multiplayer platforms aimed at minors, with about 144 million daily active users (Kaushik et al., 2026; Journey BJJ Academy, 2026). Nevertheless, other sources speak of 132 million in early 2026 and, using data that Hindenburg Research attributed to the company itself, indicate that between 20 % and 21 % of users are under 9 years old (Journey BJJ Academy, 2026; Mi3, 2024). This discrepancy arises because the sources use different dates and definitions of “active user,” and that lack of clarity is part of the controversy.

Given that scale, what in a small game would be a design decision of little consequence becomes a fundamental problem. On Roblox, the system that allows a creator to earn money also influences how much interest that creator has in a minor continuing to play, the social value of exclusive items, and how easy it is for an adult with bad intentions to win a child’s trust by giving them one. This practice is called *grooming*: an adult’s approach to a minor in order to gain their trust for purposes of abuse. In this regard, guides for families note that buying Robux causes many conflicts between parents and children and link it to limited-time events, exclusive items

and peer pressure (Common Sense Media, 2024; Screenwise, n.d.). For their part, safety audits describe conversations in which an expensive item is used as a currency of pressure and in which moving to another app happens before the harm occurs (Kaushik et al., 2026; Allchorn & Orofino, 2026).

Now, the controversy has already reached the courts. According to the available coverage, Los Angeles County sued Roblox on February 19, 2026, dozens of federal lawsuits were consolidated into a single proceeding (MDL No. 3166), and several state attorneys general have brought their own actions (Hoodline, 2026; IBTimes, 2026). At the same time, Hindenburg Research's 2024 report accused the company of inflating its figures and of failing on child safety, accusations that Roblox rejected (Game World Observer, 2024; Mi3, 2024). To this is added the case of the YouTuber Schlep, whom Roblox banned and sent a cease-and-desist letter with a threat of legal action after, according to his lawyers, his undercover operations helped lead to six arrests; the company responds that vigilantes put other users at risk (Gould Grieco & Hensley, PLLC, 2025; Rolling Stone, 2025). However, all of this consists of allegations or one party's version: none of the sources consulted reports a final judgment. This article therefore maintains that distinction from beginning to end.

Despite the attention the topic receives, the available information is scattered across technical studies, news reports, parent guides and legal press releases. Indeed, there is a first independent, large-scale evaluation of Roblox's chat, and there is also the argument that deleting content is not enough when the harm is built up gradually through a relationship (Kaushik et al., 2026; Allchorn & Orofino, 2026). Consequently, this article brings these pieces together starting from the basics: it describes how the platform works, explains why certain chat filters are easy to evade, and organizes the evidence according to its level of verification. The general objective is to analyze how economic design, player retention and moderation relate to one another to create risks for minors. The specific objectives are: (a) to describe the platform's economy and the difference between the purchase price and the redemption price; (b) to explain the mechanisms for retaining players and how solid their evidence is; (c) to examine the failures of moderation and the ways of evading filters; (d) to organize incidents and lawsuits according to their level of verification; and (e) to analyze how the company responded to outside scrutiny, based on the Schlep case.

LITERATURE REVIEW

Sources Reviewed and Summary of Approaches

Before continuing, it is worth explaining how the sources were chosen. The review follows a logic similar to that of the PRISMA method in order to show how the material was selected, but it is not a systematic review: no academic databases were consulted and there was no independent second review. Thus, we started from five sources indicated in the assignment (Kaushik et al., 2026; Allchorn & Orofino, 2026; Journey BJJ Academy, 2026; Brooks & Otte, 2025; Gould Grieco & Hensley, PLLC, 2025) and added fourteen more, found through targeted searches on October 23, 2026, to cover the economic, retention and legal topics and the Schlep case. Of those nineteen sources, four were read in full, thirteen were consulted through search excerpts, and two were known only indirectly: The Guardian, verified through reproductions, and the study by Kou et al., which is known from the citation made of it by the chat audit (Brooks & Otte, 2025; Kou et al., 2025). In addition, seven reference works are cited (behavioral psychology, mental accounting, automata theory, approximate string matching and Unicode standards) that are not part of the sources analyzed and are used only to support concepts (Ferster & Skinner, 1957; Festinger, 1954; Prelec & Loewenstein, 1998; Hopcroft et al., 2006; Navarro, 2001; Unicode Consortium, n.d.-a,n). In this sense, Table 1 summarizes the approach of each source and assigns it a level of evidence (A: clear method; B: third-party analysis or news; C: party material, commercial or guidance material).

As the table shows, two features of the set of sources affect everything that follows. First, only two sources use a method with clear data: an audit of chat and an experiment with fictitious accounts (Kaushik et al., 2026; Brooks & Otte, 2025). Second, the sources that support the economic and retention-related claims are, almost always, guides or third-party sites (Common Sense Media, 2024; RoLearn, 2026). Consequently, the quality of the evidence varies greatly by topic.

Player Retention and Virtual Economy

Within that landscape, the information on player retention is the weakest. Family guides describe spending on Robux as a source of conflict and point out that limited events, exclusive items and social pressure lead to purchases (Common Sense Media, 2024); others add rewards for logging in each day, streaks and advantages that are paid for with money (Screenwise, n.d.). However, these sources describe the design, not its effect on minors, and one clarifies that playing for long periods is not, by itself, an addiction: what matters is whether it affects family, social or school life (Addiction Help, n.d.; Common Sense Media, 2024). Therefore, in order to understand these descriptions, it is useful to draw on classic ideas that are not exclusive to Roblox: the reinforcement schedules of behavioral psychology (Ferster & Skinner, 1957), comparison with others as a driver of behavior (Festinger, 1954), and the "pain of paying" felt with cash, which an intermediate currency could reduce (Prelec & Loewenstein, 1998).

Safety, Grooming and Moderation

In contrast to the above, in safety the strongest evidence comes from the audit of nearly two million messages from four popular experiences. That audit found grooming, sexual content, harassment, violence, self-harm and requests for personal data that remained visible despite the filters (Kaushik et al., 2026). Similarly, the Revealing Reality experiment, published by The

Table 1: Summary of the literature analyzed and level of evidence.

Source	Type	Approach	Finding / limitation	Level
(Kaushik et al., 2026)	Data-based audit (preprint)	~2 M messages from 4 games; topic analysis supported by a language model	Grooming, sexual content and evasion get past the filter; only public servers and English text	A
(Brooks & Otte, 2025)	Press report on an experiment	Fictitious accounts aged 5 to 40+ (Revealing Reality)	5-year-olds were able to talk to adults; the original could not be opened	A
(Kou et al., 2025)	CHI study (indirect)	Developers' views on earnings, safety and moderation	Tension between earning money and protecting; not consulted directly	A
(Allchorn & Orofino, 2026)	Analytical framework (GNET/GIFCT)	Education, pauses, redirection and intervention	Deleting content is necessary but not enough; focuses on extremism	B
(Game World Observer, 2024; Fortune, 2024; Mi3, 2024)	News	Hindenburg's report and Roblox's response	Figures allegedly inflated by 25 to 42%; the author was betting against the stock	B
(Rolling Stone, 2025; TweakTown, 2025; KPLC, 2025)	News and company statements	Schlep case; Roblox's response; Louisiana lawsuit	Ban and cease-and-desist letter confirmed; Roblox explains its reasons	B
(Hoodline, 2026; IBTimes, 2026)	Digital media	Federal, municipal and state lawsuits	Allegations without a final judgment	B
(Common Sense Media, 2024;	Guides for families	Spending, status and use	Limited events, streaks and peer pressure; playing a lot	B/C

Screenwise, n.d.; Addiction Help, n.d.)			is not equivalent to addiction	
(Journey BJJ Academy, 2026)	Commercial blog (gym)	Summary of NCMEC figures and lawsuits	More than 65,000 reports in 2025; second-hand data	C
(Gould Grieco & Hensley, PLLC, 2025)	Law firm press release	Schlep case	Six arrests in seven confrontations; one party's document	C
(Generalist Programmer, n.d.; RoLearn, 2026)	Third-party sites	Robux purchase and redemption price	Redemption equals 28–30% of purchase; figures that do not match	C

Guardian, reaches a comparable result: children registered as five-year-olds were able to talk to adults and the safety controls worked poorly, even after Roblox restricted direct messages for accounts of users under 13 in November 2024 (Brooks & Otte, 2025; Kaushik et al., 2026).

On the other hand, the work of Allchorn and Orofino offers another perspective. They argue that removing content is necessary but insufficient when influence is built through friendship, humor and a sense of belonging, and they propose four measures: education, pauses, redirection and intervention (Allchorn & Orofino, 2026). Their analysis focuses on extremism and presents its ideas as proposals, not as evaluations of systems that already exist; even so, their British case of contact on Roblox and a move to Discord coincides with the path described by the chat audit (Allchorn & Orofino, 2026; Kaushik et al., 2026).

Financial Reports, Self-Directed Actions and Gaps

In addition to the audits, Hindenburg Research's report of October 2024 made two kinds of accusations: that key figures, such as daily active users, were inflated by between 25 % and 42 %, and that the platform was a space with grooming and abusive content (Game World Observer, 2024; Fortune, 2024). Roblox rejected these accusations (Mi3, 2024) and, in turn, one of the sources recalls that Hindenburg had taken a position to profit if the stock fell (Journey BJJ Academy, 2026). Hence, because that firm had a financial interest, the report should be taken as an idea to be checked and not as proof. In a similar way, the accusations about the Schlep case come from the lawyers of one of the parties (Gould Grieco & Hensley, PLLC, 2025), although the ban and the cease-and-desist letter are confirmed by independent media and by the company itself, which explained its reasons (Rolling Stone, 2025; TweakTown, 2025).

In sum, this review leaves two gaps. First, there is no independent evidence on how the virtual economy and retention affect minors on Roblox: there are only guides and third-party calculations (Common Sense Media, 2024; RoLearn, 2026). Second, the safety evidence is solid, but it is limited to text on public servers, so it does not cover voice, private servers or other languages (Kaushik et al., 2026; Allchorn & Orofino, 2026).

METHODOLOGICAL PROPOSAL

Documentary Design and Levels of Evidence

Given these gaps, a documentary, exploratory and qualitative design was chosen, which compares several sources and analyzes the platform's weak points. No new data were collected and no minors were spoken to. Unlike the chat audit, whose work required the approval of an ethics committee, and the experiment with fictitious accounts, this study uses only sources that have already been published (Kaushik et al., 2026; Brooks & Otte, 2025). Consequently, risks to people are avoided, but it is not possible to calculate original figures. In addition, each claim is marked as a finding (level A), news or third-party analysis (level B), or an accusation or datum from a party (level C), according to Table 1, and the figures are read as orders of magnitude and not as exact percentages of the whole platform (Kaushik et al., 2026; Allchorn & Orofino, 2026).

Framework for Analyzing Weak Points

On that basis, a structural weak point is defined as a feature of the platform's design that allows or increases harm without anyone having to break the system's rules; that is, a problem of incentives, design or platform governance, and not a simple programming error (Kaushik et al., 2026; Allchorn & Orofino, 2026). It should be made clear that this article does not analyze Roblox's code or look for flaws that someone could exploit; it only reasons about the design that can be known from the sources. In this way, the framework distinguishes six types of failure, each with partial support in the documents:

• **F1, moderation without memory:** message-by-message decisions that do not consider how the conversation progresses (Kaushik et al., 2026; Allchorn & Orofino, 2026).

• **F2, filters that are easy to evade:** filters that depend on how a word is spelled and not on what it means (Kaushik et al., 2026; Navarro, 2001).

• **F3, move to other platforms:** contacts that begin in a moderated place and continue in one with less oversight (Allchorn & Orofino, 2026; Journey BJJ Academy, 2026).

• **F4, weak age control:** minors and adults who interact without effective verification (Brooks & Otte, 2025; Journey BJJ Academy, 2026).

• **F5, misaligned incentives:** a price gap that rewards spending and playtime (RoLearn, 2026; Kou et al., 2025).

• **F6, no formal channel for outside help:** those who detect risks outside the company have no clear way to collaborate, which can lead to conflicts such as the Schlep case (KPLC, 2025; Kaushik et al., 2026).

Limitations

All in all, this design has limits that must be stated clearly. The search was not systematic, thirteen supplementary sources were consulted only through excerpts, and the video about Hindenburg that was mentioned in the assignment could not be processed, so what relates to that report relies on written notes (Game World Observer, 2024; Fortune, 2024). In addition, the DevEx rates come from third-party sites that do not agree with one another, even on the minimum for redemption, and must be checked against official documentation (Generalist Programmer, n.d.; RoLearn, 2026). Likewise, for the Schlep case only press coverage and public statements were consulted, not court documents or the original letter (Rolling Stone, 2025; KPLC, 2025). Finally, the formal reasoning in Section 4.4 is the authors' own analysis based on automata theory and on Unicode standards, not findings of the sources about Roblox (Hopcroft et al., 2006; Unicode Consortium, n.d.-b).

RESULTS AND DISCUSSION

How chat works and where the risk lies

To begin with, the risk depends on how users are grouped. On public servers, which are managed by the platform, the system brings players together automatically and each server has a maximum number of players that, in the audited games, ranges from 15 to 48; private servers, by contrast, belong to users (Kaushik et al., 2026). Thus, games with tens of billions of visits, such as Brookhaven RP (38.7 billion) and Adopt Me! (34.5 billion), produce a constant flow of conversations among strangers of different ages (Kaushik et al., 2026; Brooks & Otte, 2025). Therefore, since players come together through the game they choose and not through a verified age, failures F3 and F4 become important.

Next, it is of interest where the filter acts. According to the platform documentation cited by the audit's authors, messages are reviewed before being displayed and blocked text is replaced with hash marks; moreover, the company states that the system reviews billions of messages per day and that it uses a language model that runs on ordinary processors (Kaushik et al., 2026). Since the review must be done before the message is delivered, the available time is very short, so it is reasonable to think that quick reviews of each message are preferred over slower analyses of the whole conversation; this is a deduction about the design, not a published fact. Along these

lines, it is warned that automatic signals, on their own, cannot know the intention of the person writing and need review by a person who understands the context (Kaushik et al., 2026; Allchorn & Orofino, 2026).

Finally, what the sources do not measure also matters. The audit covers only text on public servers and in English, so voice chat and private channels remain outside the available evidence (Kaushik et al., 2026). Likewise, the AI voice detection systems mentioned in the literature are general examples for enforcing conduct rules, not systems proven to detect grooming (Allchorn & Orofino, 2026; Kaushik et al., 2026). Ultimately, that something has not been measured does not mean there is no risk, but neither does it allow us to assert that there is none.

The Virtual Economy

Turning to the economic aspect, let p be the purchase price of one Robux, d what a creator receives when redeeming it, and $\rho = d/p$ the share of the money paid that returns to the creator. With $p \approx 0.0125$ USD and d between 0.0035 and 0.0038 USD, according to the available third-party sources (Generalist Programmer, n.d.; RoLearn, 2026):

$$\rho = \frac{d}{p} \in [0.280,\ 0.304], \qquad C = R\,d = \rho\,G, \tag{1}$$

where G is what the player spends in dollars, $R = G/p$ is the Robux purchased and C is the gross money obtained upon redemption. Consequently, a spend of 10 USD equals 800 Robux and, when redeemed, yields between 2.80 and 3.04 USD before fees; moreover, one source calculates that, after store and platform fees, the developer keeps between 24 % and 29 % of each dollar (RoLearn, 2026; Generalist Programmer, n.d.). Table 2 summarizes these data and the differences between sources.

Table 2: Economic data reported by third-party sources.

Datum	Value	Source	Remark
Purchase price	≈ 0.0125 USD per Robux	(Generalist Programmer, n.d.)	Common rate; varies by package
Standard DevEx redemption	0.0035 (2025) to 0.0038 (2026) USD per Robux	(RoLearn, 2026; Generalist Programmer, n.d.)	Sources do not agree
Redemption/purchase ratio (ρ)	28.0% to 30.4%	Own calculation	Less than one third
Rate for Robux earned with verified adults in the U.S.	0.0054 USD per Robux (since June 8, 2026)	(Generalist Programmer, n.d.)	43.2% of purchase; no official confirmation
Share kept by the developer after fees	≈ 24% to 29%	(RoLearn, 2026)	Third-party calculation
Minimum for redemption	30,000 or 100,000 Robux	(Generalist Programmer, n.d.; RoLearn, 2026)	Sources contradict each other

Now, this difference does not by itself prove that there is a harmful incentive, although it does create the conditions for one to exist. Indeed, if only between a quarter and a third of the money returns to the creator, the pressure to get each player to spend more increases, and the most widely cited parent guide in the set of sources places the purchase of Robux at the center of family conflicts (Common Sense Media, 2024; RoLearn, 2026). In this sense, a study presented at CHI 2025 gathers developers' views on the tension between earning money, protecting children and moderating content (Kou et al., 2025; Kaushik et al., 2026). Additionally, if it is

confirmed that the redemption rate is higher for Robux earned with verified adults in the U.S., that would indicate that the economic design already takes verified age into account, which links the economy with failure F4 (Generalist Programmer, n.d.; Journey BJJ Academy, 2026).

In turn, the intermediate currency adds another aspect. The idea of "financial dissociation" holds that exchanging money for Robux reduces what is felt when paying, something similar to what the mental accounting literature calls the "pain of paying" of cash (Prelec & Loewenstein, 1998). However, the documents reviewed do not verify this idea in minors, and the most widely cited parent guide records only spending pressure (Common Sense Media, 2024; Prelec & Loewenstein, 1998). For this reason, it is treated here as a hypothesis that should be tested with ethical studies in the future.

Mechanisms for Retaining Players

With respect to retention, the sources describe mechanisms that coincide with one another: limited-duration events, exclusive items and social pressure (Common Sense Media, 2024), in addition to daily rewards, streaks and advantages that can be bought (Screenwise, n.d.). As our own reading, limited events and streaks set a deadline for obtaining the reward and, therefore, exclusive items turn progress into a status signal that other players see. This combination is consistent with what the theory of reinforcement schedules and of social comparison says about why a behavior is repeated (Ferster & Skinner, 1957; Festinger, 1954); nevertheless, no source measures that effect in children, so this is a theoretical explanation and not a verification with data.

Moreover, the figures that support the idea that the platform retains players for long periods are in dispute. According to Hindenburg, a consultant's review of 30.4 million daily users found about 22 minutes of play per day, compared with the 2.4 hours per day that Roblox reported for 2023 (Game World Observer, 2024; Fortune, 2024). If that reading were correct, part of the playtime shown to the public would not be real play by children; on the other hand, if it were not, retention would be even higher. One detail should be added: the documentation cited by the audit's authors indicates that Roblox removes inactive users after 20 minutes, which limits the possibility of accumulating connection time without playing (Kaushik et al., 2026). In any case, therefore, the headline figure has not been verified by anyone independent.

Ways of Evading Filters

We thus arrive at the central technical point of failure F2, which can be explained as follows. Let Σ^* be the set of all possible text strings, $L \subset \Sigma^*$ the set of prohibited patterns and N a function that orders and cleans the text before it is reviewed. A word-based filter decides

$$F(m) = 1[\, N(m) \in L] \qquad \textbf{(2)}$$

A regular expression describes a regular language, which can be recognized by a finite automaton, that is, a mechanism that remembers nothing other than its current state (Hopcroft et al., 2006; Kaushik et al., 2026). It is therefore worth asking how many ways of writing a word it would have to cover. Let $E(w)$ be the set of forms that a person understands as the word w. If each of the n characters of w has $k_i \geq 2$ variants that can be read, then

$$|E(w)| = \prod_{i=1}^{n} k_i \geq 2^n, \qquad \textbf{(3)}$$

so writing out all those forms in a list of patterns is impossible in practice, and the filter can always be evaded if some form $e \in E(w)$ satisfies $N(e) \notin L$. Consequently, this is not an error of one particular implementation, but a consequence of depending on how a word is spelled. This explains why sanctioned users try new variants again and again (Kaushik et al., 2026; Navarro, 2001).

As for the techniques observed in the data, these fall into six types: splitting the message across several lines, rephrasing what was blocked, changing the spelling, using code words, replacing letters with symbols or numbers (*leet speak*), and trying variants to see what the filter blocks (Kaushik et al., 2026). In that order of ideas, Table 3 describes them in general terms, without examples that could be used to imitate them, and relates them to the filter assumption they break and to the corresponding defensive measure. For its part, standard Unicode

normalization (NFKC forms) and the tables of look-alike characters reduce some of the letter substitutions, although they do not eliminate users' creativity in writing (Unicode Consortium, n.d.-a,n).

Table 3: Types of evasion observed, the filter assumption they break and defensive measure.

Type	**General description**	**Assumption it breaks**	**Defensive measure**
Splitting across lines	Blocked content is reassembled across consecutive messages (Kaushik et al., 2026)	That each message is judged separately	Review conversation windows rather than isolated messages (Kaushik et al., 2026; Allchorn & Orofino, 2026)
Rephrasing what was blocked	The phrase is changed but the intent is kept (Kaushik et al., 2026)	That the pattern covers all equivalent forms	Detect repeated attempts after a block (Kaushik et al., 2026)
Altered spelling	Words cut up, written as they sound, with spaces or marks (Kaushik et al., 2026)	That the word is written exactly the same	Clean the text and compare approximately (Navarro, 2001; Unicode Consortium, n.d.-a)
Code words	Substitutes that are understood from context (Kaushik et al., 2026)	That the meaning lies only in the word	Updated lists and models that understand context (Kaushik et al., 2026)
Character substitution	Letters replaced by similar symbols or numbers (Kaushik et al., 2026)	That the letters are the expected ones	Look-alike character tables and text cleaning (Unicode Consortium, n.d.-b)
Testing variants	Several forms are tried and the user observes what the filter blocks (Kaushik et al., 2026)	That the filter's response gives no information	Show fewer signals and limit retries (Kaushik et al., 2026; Allchorn & Orofino, 2026)

However, splitting the message across several lines shows the filter's deepest limit. Let $H(m_{1:T})$ be a rule that indicates whether a conversation of T messages is harmful. Evasion is possible when $H(m_{1:T}) = 1$ and, at the same time, $F(m_t) = 0$ for every t, because the intent is spread across several messages and none of them contains it on its own. In fact, the audit indicates that this type was the most frequent in its sample and that grooming progresses gradually, with steps that look innocent when taken separately (Kaushik et al., 2026; Allchorn & Orofino, 2026). Therefore, no improvement to the word list eliminates this type of evasion: what is analyzed must change, from isolated messages to entire conversations.

Likewise, testing variants adds another aspect. When the platform replaces blocked text with a visible mark, the user immediately knows what was detected, and the audit shows users who talk about the filter as an obstacle and test variants in an orderly way (Kaushik et al., 2026). Seen as a systems problem, that mark works as a response that can be queried almost for free. Consequently, one possible solution, which is a proposal of this article and not a measure already applied, would be to combine the visible sanction with pauses and limits on consecutive retries, like the pauses during conversation described by the prevention literature (Allchorn & Orofino, 2026; Kaushik et al., 2026).

That said, the audit's authors recommend combining fixed methods with models that estimate probabilities, judging conversations rather than isolated messages, tracking users over time and across games, and improving reports (Kaushik et al., 2026; Kou et al., 2025). To illustrate, Listing 1 turns those recommendations into a defensive scheme written in pseudocode, which shows the general idea without being an actual implementation.

Listing 1: Layered defensive scheme (illustrative pseudocode).

```
function evaluate_message(m, history, user):
    t = normalize(m)              # NFKC, lowercase, collapse repetitions,
                                  # remove separators and invisible characters,
                                  # map look-alike characters
    s_lex = approximate_match (t, updated_lexicon, edit_threshold)
    window = last_k(history) + [t]
    s_ctx = conversational_classifier(window)    # trust, isolation,
                                                 # app switching, personal data
    s_usr = accumulated_risk(user)    # recidivism, retries after a block
    risk = combine(s_lex, s_ctx, s_usr)
    if risk >= high_threshold:
        block(m); send_to_human_review(window, user)
    elif risk >= medium_threshold:
        apply_pause(m)            # notice to recipient, pause, options
    else:
        allow(m)
```

Nevertheless, each layer of the scheme has a cost. The audit's own language-model pre-filter had a precision of 31.39 % and detected 71.43 % of unsafe messages, which shows the dilemma between letting harmful messages through and flooding the people who review with false alarms (Kaushik et al., 2026). To this is added the fact that human capacity is limited: the coverage of Hindenburg quotes the testimony of a moderator who speaks of hundreds of child-safety reports every day and the accusation that part of the moderation is outsourced at low cost in Asia, claims by one of the parties that Roblox does not confirm in the sources consulted (Fortune, 2024; Game World Observer, 2024). Therefore, a technical improvement, by itself, does not solve a problem of lack of resources.

Report Figures and Incident Indicators

To organize the evidence with numbers, Table 4 gathers the figures from the sources and their level of verification, and distinguishes the audit's measurements, the figures from institutions relayed through secondary sources, and the accusations.

In the first place, the increase in reports to NCMEC, from 24,500 to more than 65,000 between 2024 and 2025 (about 165 % more), may be due to several causes: more incidents, better detection or greater compliance with the obligation to report (Journey BJJ Academy, 2026). In addition,

Table 4: Reported figures on safety, platform size and lawsuits.

Indicator	Value	Source	Remark
Messages and users in the audit	2,040,839 messages; 105,214 users	(Kaushik et al., 2026)	4 games, public servers, 336 hours of video
Lines with partially blocked text	88,273 (4.33%)	(Kaushik et al., 2026)	By game: from 2.96% to 6.34%
Language-model pre-filter	Detects 71.43% (recall); precision 31.39%	(Kaushik et al., 2026)	Actual figures are at least this high
Manual review	322 hits and 118 false alarms (73.2%)	(Kaushik et al., 2026)	Categories may repeat within the same case
Evasion sample	12,612 messages from 94 users	(Kaushik et al., 2026)	Users previously sanctioned

Reports to NCMEC	675 (2019); 24,500 (2024); more than 65,000 (2025)	(Journey BJJ Academy, 2026)	Second-hand figure
Report growth	+165% (2024–2025); ≈ 96 times (2019–2025)	Own calculation	May reflect more detection
Users' age	Under 16: 56%; under 13: 35%; under 9: 20%	(Journey BJJ Academy, 2026; Mi3, 2024)	Mi3 reports 21% under 9
Daily active users	≈ 144 M; 132 M (early 2026)	(Kaushik et al., 2026; Journey BJJ Academy, 2026)	Different definitions
Lawsuits in MDL No. 3166	≈ 80 at the start; ≈ 146 by May 2026	(IBTimes, 2026; Journey BJJ Academy, 2026)	No judgments; counts vary
Allegedly inflated figures	25–42%: 22 minutes versus 2.4 hours per day	(Game World Observer, 2024; Fortune, 2024)	Accusation by someone betting against the stock

the same source mentions criticism attributed to NCMEC for insufficient reporting of sadistic exploitation by Roblox and speaks of three settlements with states (Alabama, West Virginia and Nevada) for a reported total of 35.8 million dollars, without independent verification (Journey BJJ Academy, 2026). Consequently, that reports are increasing does not by itself measure how much harm there is, but neither can it be dismissed as unimportant, because reports are the backup mechanism when moderation fails (Kaushik et al., 2026).

In the second place, the audit's figures are the most solid, but they must be read with the precautions indicated by its authors. The 4.33 % of lines with blocked text shows that the filter acts frequently on parts of a message and not on whole conversations; likewise, since the pre-filter detected 71.43 % of unsafe messages, about one in four unsafe conversations was not detected, so the figures are a minimum (Kaushik et al., 2026; Allchorn & Orofino, 2026). Because categories may repeat within the same case, their counts serve to compare how frequent some types are relative to others, not to calculate exact percentages. Even so, that grooming is the most frequent category and that almost all requests to switch apps turned out to be genuine in the manual review shows a pattern that is hard to attribute to chance (Kaushik et al., 2026; Brooks & Otte, 2025).

The Schlep Case: Protecting the Platform or Caring for its Image?

To begin with, it is worth separating what is confirmed from what is not. According to the press release from his lawyers, the YouTuber Schlep, aged 22 and a survivor of grooming on the platform, was banned and received a cease-and-desist letter with legal threats based on a computer fraud law, after undercover operations that, they state, led to six arrests in seven confrontations (Gould Grieco & Hensley, PLLC, 2025). The ban and the letter do not depend solely on that source: they are also reported by independent media and the company itself justified them publicly (Rolling Stone, 2025; TweakTown, 2025). However, it is important to be precise on one point: the sources consulted speak of a cease-and-desist letter and a threat of legal action, not of a lawsuit filed by Roblox against Schlep (Rolling Stone, 2025; KPLC, 2025). According to the lawyers, the company did sue another critic, Ruben Sim, earlier under the same law, but that is a claim by one of the parties that could not be verified here (Gould Grieco & Hensley, PLLC, 2025).

For its part, Roblox explained its decision through statements from its chief safety officer. According to those statements, vigilante groups pose as minors, approach other users and lead them to other platforms to have sexual conversations, which goes against the terms of use; moreover, they sometimes withhold their reports until they organize in-person meetings, and their actions can increase the risk to other users and delay the work of the company and the police (TweakTown, 2025; Rolling Stone, 2025). The company added that other apps, such as Discord, TikTok and Snapchat, have similar rules against these groups (Rolling Stone, 2025; TweakTown, 2025). Table 5 compares both versions.

This explanation from the company coincides in part with the prevention literature, which maintains that players should not investigate or publicly confront possible recruiters and that the assessment belongs to trained teams (Allchorn & Orofino, 2026; TweakTown, 2025). Nevertheless, those who criticize the company see something else: Schlep's lawyers argue that Roblox appears more concerned with silencing criticism than with addressing the problem of predators (Gould Grieco & Hensley, PLLC, 2025), and the YouTuber's followers argued that he had intervened in a problem the company preferred to hide (Rolling Stone, 2025). This reading, according to which the company prioritizes protecting its image over supporting whistleblowers, is an interpretation by one of the parties and its sympathizers, not an established fact.

Table 5: Two versions of the Schlep case.

Topic	Version of Schlep and his lawyers	Roblox's version	Sources
What happened	Ban from his accounts and cease-and-desist letter after operations that allegedly led to six arrests (Gould Grieco & Hensley, PLLC, 2025)	Confirms the ban for violating its rules on vigilantes (Rolling Stone, 2025)	(Gould Grieco & Hensley, PLLC, 2025; Rolling Stone, 2025)
Reason	Retaliation for criticizing and for exposing safety flaws (Gould Grieco & Hensley, PLLC, 2025)	Vigilantes pose as minors, approach users and lead them to other platforms (TweakTown, 2025)	(Gould Grieco & Hensley, PLLC, 2025; TweakTown, 2025)
Effect on safety	Helped arrest alleged abusers (no judicial verification here) (Gould Grieco & Hensley, PLLC, 2025)	Increases the risk to other users and may delay the company and the police (Rolling Stone, 2025)	(Gould Grieco & Hensley, PLLC, 2025; Rolling Stone, 2025)
Prior contact	Says the letter was the company's first official contact (KPLC, 2025)	Not recorded in the sources consulted	(KPLC, 2025)
Legal route	Cease-and-desist letter and legal threat: they mention a computer fraud law (Gould Grieco & Hensley, PLLC, 2025)	No lawsuit against Schlep was found in the sources consulted (Rolling Stone, 2025)	(Gould Grieco & Hensley, PLLC, 2025; Rolling Stone, 2025)

Therefore, with the available evidence it is not possible to know the company's true intention. What can be observed are two facts that help explain why the controversy grew. First, according to Schlep, the first official contact he received from Roblox was the cease-and-desist letter (KPLC, 2025), which suggests that there was no formal channel for collaborating with him. Second, the case occurred amid lawsuits from several states, among them that of Louisiana, whose attorney general maintained that the platform's decisions aim to increase user engagement and profits (KPLC, 2025; IBTimes, 2026). Consequently, beyond intent, the case shows failure F6: the lack of a clear route for those who detect risks to collaborate with the company. A proposal of this article, which should be studied before being applied, is to create a formal channel for verified whistleblowers, with rules that prevent both posing as minors and delaying reports, so that outside contributions do not depend on public confrontation (Allchorn & Orofino, 2026; Kaushik et al., 2026).

Company Responses and Recent Changes

In contrast to the criticism, Roblox has made changes that cannot be ignored. In November 2024 it restricted direct messages for accounts of users under 13 (Brooks & Otte, 2025) and, according to one of the sources, in November 2025 it announced age estimation using a photo of the face in order to use chat, with a worldwide rollout on January 7, 2026, in addition to new account types for ages 5 to 8 and 9 to 15 (Journey BJJ Academy, 2026). The same source records Hindenburg's accusation, based on internal documents that could not be

reviewed here, that years earlier the company declined to require parental approval for chat for children under 13 for fear of lowering user engagement (Journey BJJ Academy, 2026; Game World Observer, 2024). In addition, the company has mentioned an artificial intelligence model called Sentinel for monitoring text and voice, although the sources consulted do not allow its effectiveness to be evaluated (KPLC, 2025; Kaushik et al., 2026). Consequently, because the audit data were collected in 2024 and 2025, they predate age verification and do not allow its effect to be measured (Kaushik et al., 2026; Journey BJJ Academy, 2026).

General Discussion

In bringing the three topics together, the working idea is a circuit and not a single cause. Let S be the social value of virtual items, T the trust that an adult can gain with them, M the move to spaces with less moderation and D the harm:

$$S \longrightarrow T \longrightarrow M \longrightarrow D. \qquad (4)$$

In that circuit, the first step depends on the status and retention mechanisms (Common Sense Media, 2024; Screenwise, n.d.); the second, on observed exchanges in which an expensive item is used as a currency of pressure (Kaushik et al., 2026); and the third, on the documented path from Roblox to other apps, which the British case takes as far as blackmail (Allchorn & Orofino, 2026; Kaushik et al., 2026). Therefore, moderation without memory detects all three steps poorly, because the signal lies in how the conversation progresses and not in an isolated message.

To this is added a difference in costs that helps explain why the pattern repeats. For an adult, giving away an item costs little; in contrast, for a minor within a status hierarchy, that item can be worth a great deal socially (Common Sense Media, 2024; Festinger, 1954). Hence, this difference makes the virtual economy a cheap means of gaining influence, while the filter sanctions texts and not the relationship between people (Kaushik et al., 2026; Allchorn & Orofino, 2026). This is a deduction consistent with the sources, not a verified result.

All the same, the strength of the circuit is uneven. The second and third steps have level A and B evidence; the first depends on guides and third-party calculations; and the complete circuit is a deduction that no study in the set of sources has put to the test (Kaushik et al., 2026; Common Sense Media, 2024). On the legal front, the controversy revolves around whether design decisions and what the company told the public about its safety constitute negligence or deceptive practices, as the lawsuits by families and state attorneys general maintain (IBTimes, 2026; Hoodline, 2026). That will be decided by the courts; therefore, this article only distinguishes which technical evidence is relevant to that discussion and which is merely an accusation.

What can be Improved in Design and Policy

In light of the above, in moderation the evidence supports a multilayer approach: cleaning the text and comparing it approximately to handle the different ways of writing, using models that understand context to detect grooming, and tracking sanctioned users over time, sharing that information across games (Kaushik et al., 2026; Kou et al., 2025). To this are added pauses at moments of risk, such as a notice to the recipient when a new contact insists on switching apps, and intervention routes with human review and with options to correct errors (Allchorn & Orofino, 2026; Kaushik et al., 2026). It should be made clear that pauses and education do not replace detection: they complement it, and how well they work is a question that has not yet been answered with data.

On the other hand, in the economic sphere and in platform governance, the most important thing is transparency. Given that the figures on users and playtime support both the company's value and the debate on retention, having someone independent verify them is necessary for a quality public discussion: the company itself acknowledges that daily active users do not measure unique people (Fortune, 2024; Game World Observer, 2024). In this framework, a proposal of this article, which should be evaluated before being applied, is to condition part of the earnings from redemption on safety indicators for each game, in order to align the incentives that today reward spending (RoLearn, 2026; Kou et al., 2025). Finally, moderation that does not generate trust in the reporting system loses its safety net, because users see that the filter detects messages and, even so, the

person who abuses keeps writing (Kaushik et al., 2026; Brooks & Otte, 2025).

CONCLUSIONS AND OUTLOOK

Ultimately, the findings allow four conclusions. First, there is independent evidence that Roblox's chat moderation lets through grooming, sexual content and requests to switch apps, and that those who break the rules adapt their tactics again and again (Kaushik et al., 2026; Brooks & Otte, 2025). Second, the evidence on economic design and retention is much weaker: it rests on family guides and third-party calculations, and the usage figures are in dispute (Common Sense Media, 2024; Game World Observer, 2024). Third, the legal landscape is moving quickly, but almost everything available consists of pending allegations and no judgment allows responsibility to be assigned (IBTimes, 2026; Hoodline, 2026). Fourth, the Schlep case pits two genuine arguments against each other: that of the company, which speaks of the risks of self-directed vigilance, and that of its critics, who suspect it prioritizes its image; with the available evidence it is not possible to decide between the two, but the lack of a formal channel for collaboration is evident (Rolling Stone, 2025; Gould Grieco & Hensley, PLLC, 2025).

Hence, the impact of these findings is twofold. Technically, the analysis shows that word-based filters and filters that look at isolated messages not only fail in practice, but are insufficient against users who adapt and against an intent spread across several messages, so any improvement to the word list will be followed by new variants (Kaushik et al., 2026; Hopcroft et al., 2006). On the regulatory side, the combination of a price gap, status mechanisms and moderation without memory suggests that responses that only block words would not be sufficient, although that conclusion depends on a circuit that has not yet been verified (Allchorn & Orofino, 2026; Common Sense Media, 2024).

For all of the above, seven lines of future research are proposed:

1. Independent audits that include voice chat, private servers and languages other than English, which today are not in the evidence (Kaushik et al., 2026; Allchorn & Orofino, 2026).

2. Before-and-after evaluations of photo-based age estimation and of the new account types (Journey BJJ Academy, 2026; Brooks & Otte, 2025).

3. Ethical studies on spending, status and excessive use among minors that replace guides as the evidence base and that test the idea of financial dissociation (Common Sense Media, 2024; Prelec & Loewenstein, 1998).

4. Verification with primary sources of the purchase and redemption prices and of the user figures (Generalist Programmer, n.d.; Fortune, 2024).

5. Transparent tracking of court proceedings in order to compare the allegations with the rulings (Hoodline, 2026; IBTimes, 2026).

6. Controlled tests of the proposed measures (text cleaning, conversation review, per-user tracking and pauses) to measure their precision, how many cases they detect and how much extra work they generate for reviewers (Kaushik et al., 2026; Allchorn & Orofino, 2026).

7. Studies on formal protocols for collaboration with verified whistleblowers and on the effects, good and bad, of self-directed vigilance, based on court documents and not only on the press (Rolling Stone, 2025; Allchorn & Orofino, 2026).

In short, Roblox is at once a space for creativity and togetherness for millions of young people and a case study of how the economy, retention and moderation relate to one another on platforms with user-generated content. For this reason, recognizing the strength of the evidence on moderation without extending the conclusions to where the evidence is weak is the basis for a serious technical and regulatory debate (Kaushik et al., 2026; Allchorn & Orofino, 2026).

REFERENCES

1. Addiction Help. (n.d.). *Is Roblox bad for kids?* https://www.addictionhelp.com/video-gam e-addiction/roblox/is-roblox-bad-for-kids/
2. Allchorn, W., & Orofino, E. (2026, September 21). *Beyond the ban hammer: Building resilience against youth exploitation in gaming spaces*. GNET. https://gnet-research.org/2026/09/ 21/beyond-the-ban-hammer-building-resilience-against-youth-exploitation-in-g aming-spaces/
3. Brooks, L., & Otte, J. (2025, April 14). Risks to children playing Roblox 'deeply disturbing', say researchers. *The Guardian*. https://www.theguardian.com/technology/2025/apr/14/ris ks-children-roblox-deeply-disturbing-researchers
4. Common Sense Media. (2024). *Parents' ultimate guide to Roblox*. https://www.fullertonsd. org/cms/lib/CA50010905/Centricity/Domain/1188/Parents%20Guide%20to%20Roblox% 20Common%20Sense%20Media.pdf
5. Ferster, C. B., & Skinner, B. F. (1957). *Schedules of reinforcement*. Appleton-Century-Crofts. Festinger, L. (1954). A theory of social comparison processes. *Human Relations, 7* (2), 117–140. Fortune. (2024, October 9). [Coverage of Hindenburg Research's report on Roblox]. https://fo
6. rtune.com/2024/10/09/roblox-hindenburg-research-report-pedophile-hellscape Game World Observer. (2024, October 9). Hindenburg Research drops bomb on Roblox, accusing platform of inflated DAU, "zombie" engagement, and more. https://gameworldobserver.co m/2024/10/09/hindenburg-research-roblox-report-inflated-dau-child-safety Generalist Programmer. (n.d.). *Robux to USD converter*. https://generalistprogrammer.com /tools/robux-to-usd/17000
7. Gould Grieco & Hensley, PLLC. (2025, August 14). *GGH and Milberg defend YouTuber Michael 'Schlep,' banned from Roblox after exposing child predators leading to six arrests* [Press release]. https://gghlaw.com/press-releases/sggh-and-milberg-defend-youtuber-michael-s chlep-banned-from-roblox-after-exposing-child-predators/
8. Hoodline. (2026, February 19). L.A. County drags Roblox into court over child safety scandals. https://hoodline.com/2026/02/l-a-county-drags-roblox-into-court-over-child-s afety-scandals/
9. Hopcroft, J. E., Motwani, R., & Ullman, J. D. (2006). *Introduction to automata theory, languages, and computation* (3rd ed.). Pearson.
10. IBTimes. (2026). Roblox allegedly failed to protect kids, new lawsuits claim negligence. https: //www.ibtimes.com/roblox-allegedly-failed-protect-kids-new-lawsuits-claim-n egligence-3796723
11. Journey BJJ Academy. (2026, May 19). *Roblox logged 65,000 child-safety reports. The bigger* 16
12. *story is what kids run to next*. https://journeybjjacademy.com/blog/roblox-logged-650 00-child-safety-reports-the-bigger-story-is-what-kids-run-to-next/ Kaushik, P., Hasan, R., Brown, S., & Rahaman, S. (2026). *An evaluation of chat safety modera tions in Roblox*. arXiv. https://arxiv.org/html/2605.04491v2
13. Kou, Y., Hernandez, R. H., & Gui, X. (2025). "The system is made to inherently push child gambling in my opinion": Child safety, monetization, and moderation on Roblox. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems* (pp. 1–18). [Cited in Kaushik et al., 2026; not consulted directly].
14. KPLC. (2025, August 29). Roblox under fire after La. Attorney General opens lawsuit, ban of popular YouTuber. https://www.kplctv.com/2025/08/29/roblox-under-fire-after-l a-attorney-general-opens-lawsuit-ban-popular-youtuber/
15. Mi3. (2024, October 11). $27bn games giant Roblox refutes activist hedge fund allegations of misleading user metrics, and jeopardising child safety. https://mi-3.com.au/11-10-2024/ 27bn-games-giant-roblox-refutes-activist-hedge-fund-allegations-misleading-u ser-metrics
16. Navarro, G. (2001). A guided tour to approximate string matching. *ACM Computing Surveys, 33* (1), 31–88.
17. Prelec, D., & Loewenstein, G. (1998). The red and the black: Mental accounting of savings and debt. *Marketing Science, 17* (1), 4–28.
18. RoLearn. (2026). *Roblox developer revenue share explained (2025–2026)*. https://rolearn.de v/insights/roblox-developer-revenue-share-2026/
19. Rolling Stone. (2025). Roblox banned a creator for 'hunting' pedophiles — and critics are pissed. https://www.rollingstone.com/culture/culture-features/roblox-schlep-pedophile -vigilante-lawsuits-1235414218/

20. Screenwise. (n.d.). *Why kids ask for Robux again and again*. https://screenwiseapp.com/gu ides/why-kids-ask-for-robux-again-and-again
21. TweakTown. (2025). Roblox officially responds to banning a YouTuber that exposes child predators. https://www.tweaktown.com/news/107084/roblox-officially-responds-to-banning-a-youtuber-that-exposes-child-predators/index.html
22. Unicode Consortium. (n.d.-a). *Unicode Standard Annex #15: Unicode normalization forms*. https://www.unicode.org/reports/tr15/
23. Unicode Consortium. (n.d.-b). *Unicode Technical Standard #39: Unicode security mechanisms*. https://www.unicode.org/reports/tr39/