



2024 H1

TRANSPARENCY REPORT



TABLE OF CONTENTS

<u>03</u>	EXECUTIVE SUMMARY
<u>05</u>	KEY TAKEAWAYS
<u>06</u>	ADVANCED INNOVATIONS IN SAFETY
<u>07</u>	HIGH LEVEL DATA SUMMARY
<u>08</u>	OUR VISION
<u>11</u>	OUR APPROACH
<u>13</u>	Community Standards
<u>14</u>	Player Choice via Settings
<u>15</u>	Parental Controls
<u>16</u>	Enforcement
<u>16</u>	Microsoft Digital Safety Content Report
<u>17</u>	Proactive Moderation
<u>17</u>	Reactive Moderation
<u>18</u>	Help When Players Need It
<u>19</u>	Appeals / Case Reviews
<u>20</u>	SHARING OUR SAFETY DATA
<u>21</u>	Proactive Moderation Data
<u>24</u>	Reactive Moderation Data (Player Reported)
<u>26</u>	Enforcements Data
<u>28</u>	Microsoft Digital Safety Content Report Data
<u>28</u>	Crisis Text Line Data
<u>29</u>	Appeals Data
<u>30</u>	Toxicity Prevented
<u>31</u>	POLICIES AND PRACTICES
<u>32</u>	Supplemental Information
<u>33</u>	GLOSSARY OF TERMS
<u>34</u>	Glossary
<u>35</u>	APPENDIX
<u>36</u>	Player Journey Infographic
<u>37</u>	Player Image Upload Infographic
<u>38</u>	Enforcement Strike System Infographic

EXECUTIVE SUMMARY



EXECUTIVE SUMMARY

At Xbox, our mission is to bring the joy and community of gaming to everyone on the planet.

When you come to play, you deserve the opportunity to experience a place free from fear and intimidation, safe within the boundaries that you set.



As the Xbox community grows and technology continues to advance, protecting players from disruptive online behavior becomes increasingly complex. Our approach to creating more inclusive, approachable, and safer gaming experiences is constantly evolving to meet these challenges. In our fifth transparency report, we will share some of the ways we are combining player-centric solutions with the responsible application of AI to ensure we are creating positive gaming experiences for Xbox players and upholding our [Community Standards](#) together.

Players are at the center of everything we do. As stewards of our community, their use of our reporting feature continues to be a crucial way in which communities are kept fun, safe, and welcoming. We have also started deploying advanced AI solutions to both proactively address unwanted communication and enhance our ability to respond to player reports so our human moderators can prioritize more nuanced and complex issues. By applying AI responsibly with a focus on amplifying human expertise, we have been able to effectively scale our safety solutions to meet the growing needs of our community.

As always, we are deeply committed to working with industry partners, regulators, and our players to improve our multifaceted safety strategy. We will continue to learn, iterate, and be transparent about our approach in keeping the Xbox community safe and fun for everyone.

EXECUTIVE SUMMARY

Key takeaways from the report

01 Balancing safety and authenticity in messaging

We introduced a new approach to detect and intercept harmful messages between non-friends, contributing to a significant rise in disruptive content prevented. From January to June, a total of **19M pieces of Xbox Community Standards-violating content were prevented** from reaching players across text, image, and video. This new approach balances two goals: safeguarding players from harmful content sent by non-friends, while still preserving the authentic online gaming experiences our community enjoys. We encourage players to use the [New Xbox Friends and Followers Experience](#), which gives more control and flexibility when connecting with others.

02 Safety boosted by player reports

Player reporting continues to be a critical component in our safety approach. During this period, players helped us identify an uptick in spam and advertising on the platform. We are constantly evolving our strategy to prevent creation of inauthentic accounts at the source, limiting their impact on both players and the moderation team. **In April, we took action on a surge of inauthentic accounts (1.7M cases, up from 320k in January) that were affecting players in the form of spam and advertising.** Players helped us identify this surge and pattern by providing reports in Looking for Group (LFG) messages. **Player reports doubled to 2M for LFG messages and were up 8% to 30M** across content types compared to the last transparency report period.

03 Our dual AI approach

We released two new AI tools built to support our moderation teams. These innovations not only prevent the exposure of disruptive material to players but allow our human moderators to prioritize their efforts on more nuanced and complex issues. The first of these new solutions is Xbox AutoMod, a system that launched in February and assists the moderation of reported content. So far, **it has handled 1.2M cases and enabled the team to remove content affecting players 88% faster.** The second AI solution we introduced launched in July and proactively works to prevent unwanted communications. We have directed these solutions to detect Spam and Advertising and will expand to prevent more harm types in the future.

ADVANCED INNOVATIONS IN SAFETY

Player-Supported AI Solutions for Increasingly Complex Challenges

The gaming and technology landscape is constantly shifting and evolving, requiring a dynamic and comprehensive approach to community safety. With our community of players growing, not just on Xbox consoles but across platforms like mobile, PC, and beyond, our solutions to keep players safe need to continue evolving and advancing across all the ways players engage with each other.

At Xbox, we believe the technological advances of AI, combined with human expertise, play crucial and complementary roles in effectively identifying, reporting, and preventing harms at scale, especially as these online harms become more technologically sophisticated. We continue to improve upon our existing solutions, including Community Sift, so we can best meet the current and anticipated needs of players and communities.

The application of AI continues to be a preliminary step in categorizing and labeling text content, allowing teams at Xbox to both prevent known toxicity and focus human moderation efforts on prioritizing more nuanced and complex issues. To increase our efficiency in detecting unwanted messages, we have launched two new AI-enabled tools that take a multi-faceted approach to protecting players:

- **Xbox Automod** reactively classifies and takes action on text, once a player uses the reporting feature, by identifying words or phrases that are likely to be unwanted or harmful. This tool uses a sophisticated language model to analyze the reported text and make quick decisions.
- **The second AI solution** proactively classifies messages. This tool is powered by an efficient and targeted language model that allows it to review millions of messages every day to prevent players from seeing unwanted and harmful content.

These solutions are directed to detect Spam and Advertising and are rapidly expanding to more harm types in the future. Underpinning these new advancements is a safety system that relies on both players and the expertise of human moderators to ensure the consistent and fair application of [Xbox Community Standards](#), while improving our overall approach through a continuous feedback loop.

We continue to build a safe, inclusive, and fun gaming community, and we remain committed to creating responsible AI by design, guided by Microsoft's [Responsible AI Standard](#).

H1 2024 High Level Safety Data Summary (Jan – Jun 2024)

Player Reports

29.63M

14.31M (48%) Communications

10.85M (37%) Conduct

4.47M (15%) User Generated Content

Enforcements Issued

9.57M

6.76M (71%) Proactive¹

2.81M (29%) Reactive² (From Player Reports)

NCMEC³ Reports

606

Crisis Text Line Referrals

2,210

Appeals (Case Review)

171.31k

143.63k (84%) Non-Reinstatements

27.67k (16%) Reinstatements

¹**Proactive Enforcement** – When we action on inappropriate content or conduct before a player brings it to our attention

²**Reactive Enforcement** – When we action on inappropriate content or conduct via a player bringing it to our attention

³**NCMEC** – National Center for Missing & Exploited Children

OUR VISION



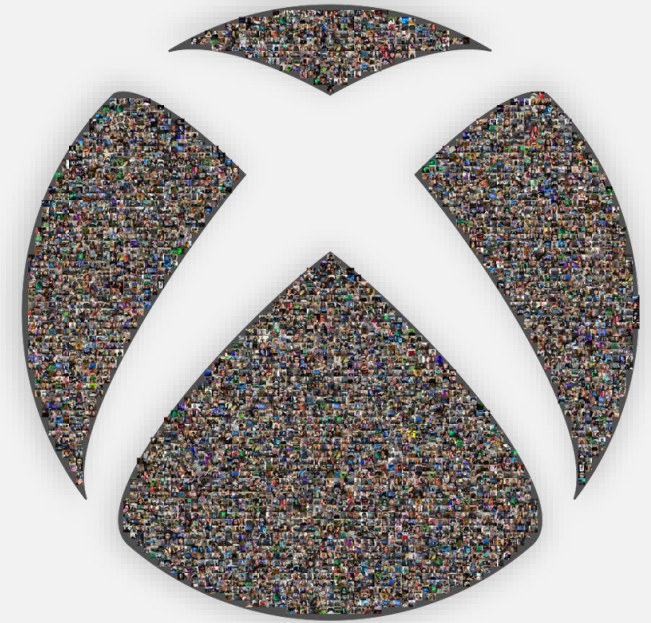
OUR VISION

The Xbox community is yours.

We all bring something unique, and that uniqueness is worth protecting.

Whether you are new to gaming or have been playing for decades, you are stewards of this place, protecting each other even as you compete.

Because when everyone plays, we all win.



OUR VISION

Our [Xbox Community Standards](#) outline the conduct and content that are acceptable within our community. We acknowledge that negative activity can and has taken place. This conduct is not okay and goes against the community we strive to create – a place that is vibrant, safe, and welcoming.

We want you to feel confident that we are listening and acting upon your feedback – we use that feedback to test and implement new features, and better understand the activity and conduct of our players. One way to help us deliver the best gaming experience possible is to [provide feedback](#) and by taking part in our [Xbox Insider Program](#).



OUR APPROACH



OUR APPROACH

Our multifaceted approach

- Working to create a strong community of gamers who are thoughtful about their conduct and guided by comprehensive [Community Standards](#)
- Giving players controls to customize their settings across the entire Xbox ecosystem from console to PC to Xbox Cloud Gaming (Beta), including comprehensive [parental controls](#) so children can engage in safer experiences that are appropriate for them
- Using proactive technology and tools to detect and remove problematic content before it is seen and to reduce conduct that runs counter to our Community Standards
- Enabling useful [reporting tools](#) for our players to identify issues
- An [Appeals](#) process to educate our users about the Community Standards
- A new [Enforcement Strike System](#) to help players better understand their enforcements and prevent them from repeating
- Continued learning and investment in our safety measures

↔ [Learn about our shared commitment to safer gaming](#)

Protecting our community requires constant work and diligence. Our foundational approach to safety-by-design and a dedicated team ensures safety is, and will always be, a priority for everyone.



OUR APPROACH



Community Standards

The [Microsoft Services Agreement's](#) Code of Conduct section applies to Xbox and its players. Our [Xbox Community Standards](#) offers an additional level of explanation, providing specifics on our expectations for player conduct on our network. They also reflect the policies we have in place to moderate conduct and, when necessary, impose consequences for players that violate our policies.

⇒ [Learn about the Xbox Community Standards](#)

OUR APPROACH

Player Choice via Settings

We know that when it comes to preferences on content and experiences, it is not one-size-fits-all. Content or language that is fine for one player may not be suitable for others.

We offer our players choices about the types of content they want to see and experience on our network, which include:

- [Automated text, media and web link filtration](#) so you can decide what text-based messages you would be comfortable receiving
- [Filter flexibility](#), allowing players to configure safety settings along a spectrum from most filtered to least so you can choose what is best for you
- Customizable [parental controls](#), including a convenient [Xbox Family Settings App](#) on mobile devices
- [Mute and block](#) other players and their messages
- [Real name sharing](#) if players want to share their real name with friends
- [Customizable privacy and notification settings](#) that let players manage their social experience.

Every player has the opportunity to adjust and select their privacy and safety settings at any time, with those settings being effective across all the ways players access Xbox.

↔ [Learn about safety settings for Xbox messages](#)

↔ [Learn about managing Xbox safety and privacy settings](#)

OUR APPROACH



Parental Controls

Xbox offers a robust set of [parental controls](#) that help children on our platform have safer experiences on our services, including a convenient [Xbox Family Settings App](#) for mobile devices. Child accounts on Xbox come with default settings that prevent children from viewing or playing games that have mature ratings and require parental permission for other actions such as playing multiplayer games, chatting with other players, and making purchases. Parents can also receive [weekly activity reports](#) about their children's time on Xbox, including games played, time spent on each game, and purchases made.

We care deeply about what our Xbox Community wants. That is why we've continued to add to our capabilities since the debut of our Xbox Family Settings App. Because of direct feedback from parents of gamers, we've added more options to [prevent unauthorized purchases](#) and the ability for caregivers to [set good screen time habits](#). These options also help spark conversation between parents and children to help younger players build stronger digital skills and safely navigate their online presence.

⇒ [Download the Xbox Family Settings app](#)

⇒ [Learn more about Parental Controls](#)

⇒ [Learn more about the Xbox Family Settings App](#)

OUR APPROACH

Enforcement

When a player's conduct or content has been found to violate our policies, the content moderation agents or systems will take action - we call this an enforcement. Most often this comes in the form of removing the offending content from the service and issuing the associated account a suspension.

The length of suspension is primarily based on the type of offending conduct or content while taking into consideration the account's previous history. Repeated violations of the policies result in lengthier suspensions and can culminate in a 12-month suspension of social features. Particularly egregious violations can result in permanent account suspensions or device bans.

We recently introduced a [new strike system](#) to our Enforcement approach designed to better educate players about enforcements and to further empower players to engage positively and appropriately on Xbox and with the community.

↔ [Learn about types of enforcements](#)

↔ [Enforcement strike system FAQ](#)

↔ [Enforcement action FAQ](#)

Microsoft Digital Safety Content Report

For several years, Microsoft has published a bi-annual [Digital Safety Content Report \(DSCR\)](#), which covers actions Microsoft has taken against terrorist and violent extremist content ([TVEC](#)), non-consensual intimate imagery ([NCII](#)), child sexual exploitation and abuse imagery ([CSEAI](#)), and grooming of children for sexual purposes across its consumer services, including Xbox.

At Xbox, violations of our CSEAI, grooming of children for sexual purposes, or TVEC policies will result in removal of the content and a permanent suspension to the account, even if it is a first offense. These types of cases, along with threats to life (self, others, public) and other imminent harms are immediately investigated and escalated to law enforcement, as necessary.

↔ [Learn about the Digital Safety Content Report \(DSCR\)](#)

OUR APPROACH

Proactive Moderation

To reduce the risk of toxicity and prevent our players from being exposed to inappropriate content, we use proactive measures that identify and stop harmful content before it impacts players. For example, [proactive moderation](#) allows us to find and remove inauthentic accounts so we can improve the experiences of real players.

For years at Xbox, we've been using a set of content moderation technologies to proactively help us address policy-violating text, images, and video shared by players on Xbox. With the help of these common moderation methods, we've been able to automate some of our processes. This automation enables us to achieve greater scale, elevate the capabilities of our human moderators, and reduce exposure to sensitive content. If content that violates our policies is detected, it can be proactively blocked or removed.

Reactive Moderation

Proactive blocking and filtering are only one part of the process in reducing toxicity on our service. Xbox offers robust reporting features, in addition to [privacy and safety controls](#) and the ability to [mute and block](#) other players; however, inappropriate content can make it through the systems and to a player.

Reactive moderation is any moderation and review of content that a [player reports to Xbox](#). When a player reports another player, a message, or other content on the service, the report is logged and sent to our moderation platform for review by content moderation technologies and human agents. These reactive reports are reviewed and acted upon according to the relevant policies that apply. We see players as partners in our journey, and we want to work with the community to meet our [vision](#).

OUR APPROACH

Help When Players Need It

We also look to help our players when they need it. If a player's communications are flagged as concerning (including content associated with suicide ideation or self-harm), either by our system or by other players, we may provide Crisis Text Line information to the player so they can reach out to resources who can help.

Crisis Text Line is a US-based nonprofit organization that [Xbox has been partnering with since 2018](#), which provides free, text-based 24/7 support.



OUR APPROACH



Appeals / Case Reviews

Our [appeals](#) process enables a player to get more information about any enforcements they have received including account suspensions or content removals. A player can launch an appeal, otherwise known as a case review, to provide us with more information if they disagree with our determination that a policy was violated. Based on the appeal, the original decision may be confirmed, modified, or overturned and the account reinstated.

⇒ [How to file a case review](#)

⇒ [Learn about types of enforcements](#)

⇒ [Enforcement action FAQ](#)

⇒ [Enforcement Strike System FAQ](#)

SHARING OUR SAFETY DATA



The data that we'll be sharing below covers the time period between Jan 1 – Jun 30, 2024 and was collected in accordance with [Microsoft's commitment to privacy.](#)

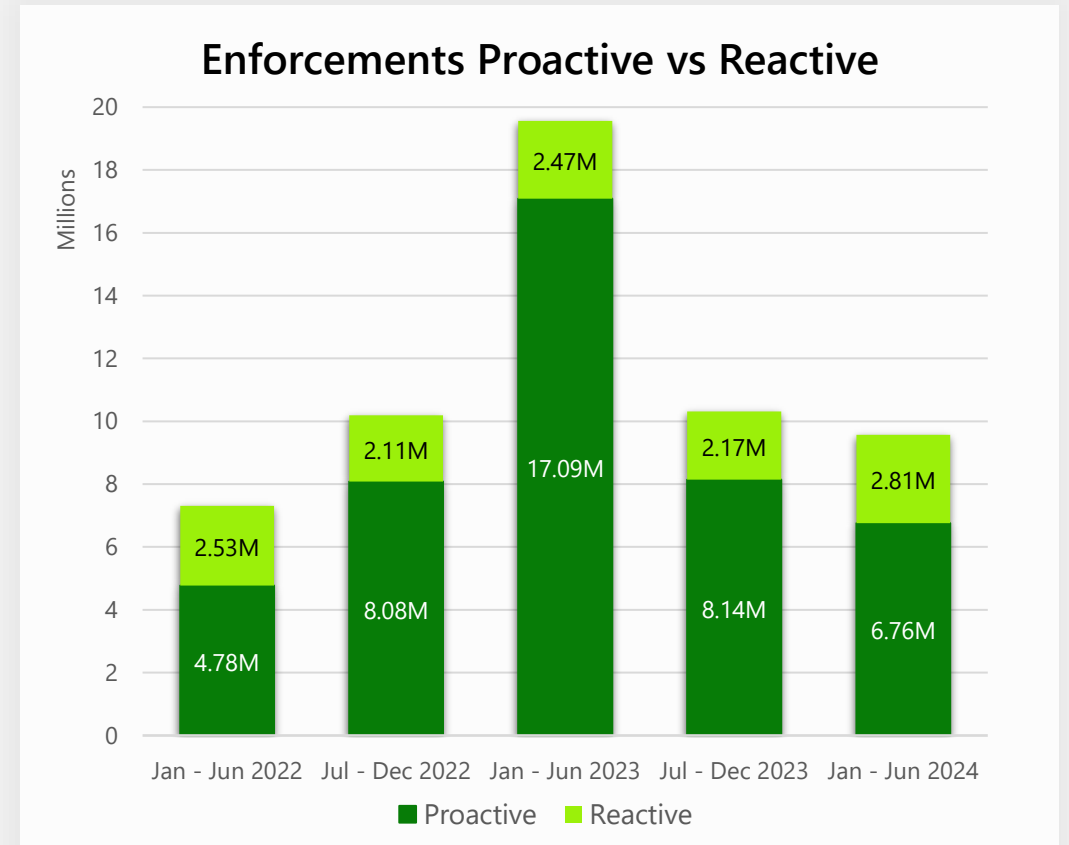
SHARING OUR SAFETY DATA

Proactive Moderation Data

Proactive enforcements happen when we use our portfolio of protective technologies and processes to find and manage an issue before it is brought to our attention by a player. In this reporting period, we handled a total of **6.76M** proactive enforcements, representing **71%** of all enforcements.

We are committed to the detection and removal of accounts that have been tampered with or are being used in inauthentic ways to prevent impact to players. **89%** of proactive enforcements over the last period were directed at cheating and inauthentic accounts. Working with the Microsoft accounts team to stop these accounts from being created before they can reach Xbox players remains a top priority.

Below is a breakdown of proactive vs. reactive enforcements over time:

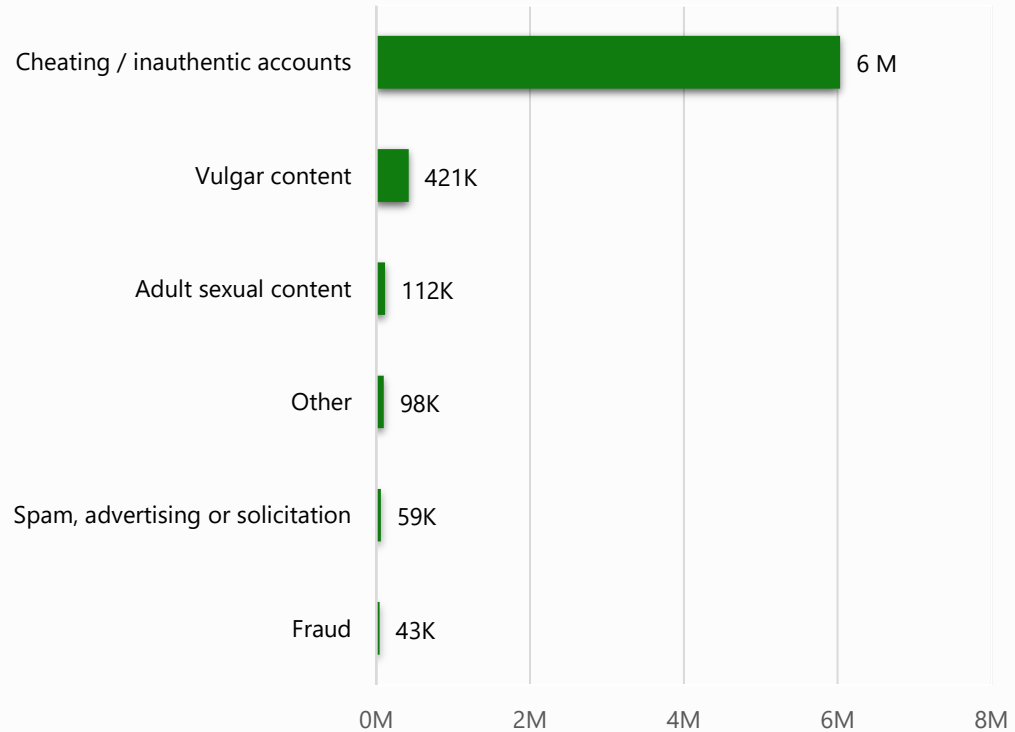


A note on the increase in proactive enforcements and subsequent drop: in collaboration with the Microsoft accounts team, many inauthentic accounts are now prevented from being created as Microsoft accounts in the first place, which no longer appear as Xbox enforcements.

SHARING OUR SAFETY DATA

We can break our proactive enforcements down into policy areas for the previous 6-month period:

Proactive Enforcements by Policy Area

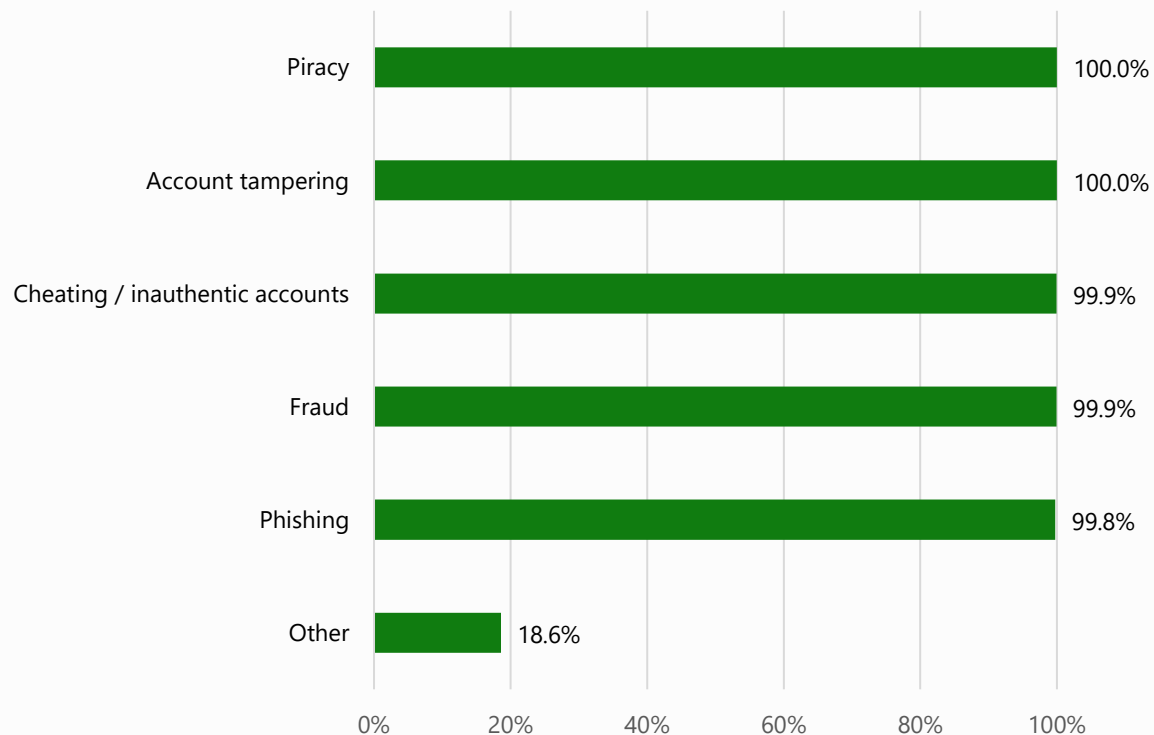


Beyond our focus on stopping inauthentic accounts as soon as possible, the other areas that see high numbers of proactive enforcements include vulgar content, adult sexual content, fraud, and spam. The Other category includes smaller volume areas such as hate speech, violence, profanity, piracy, account tampering, harassment or bullying, and drugs in aggregate.

SHARING OUR SAFETY DATA

We can continue to examine enforcements by looking at the % that were issued proactively (before a player brought the issue to our attention) by policy area for the previous 6-month period:

% of Proactive Enforcements by Policy Area



Dealing with inappropriate conduct and content before it is reported to us by players is an important element to creating a healthy and competitive gaming environment.

In addition to our focus on stopping inauthentic accounts, the other areas that see high percentages of proactive enforcements include account tampering, piracy, fraud, and phishing. The Other category includes areas such as vulgar content, drugs, profanity, hate speech, harassment or bullying, and spam, advertising, or solicitation.

SHARING OUR SAFETY DATA

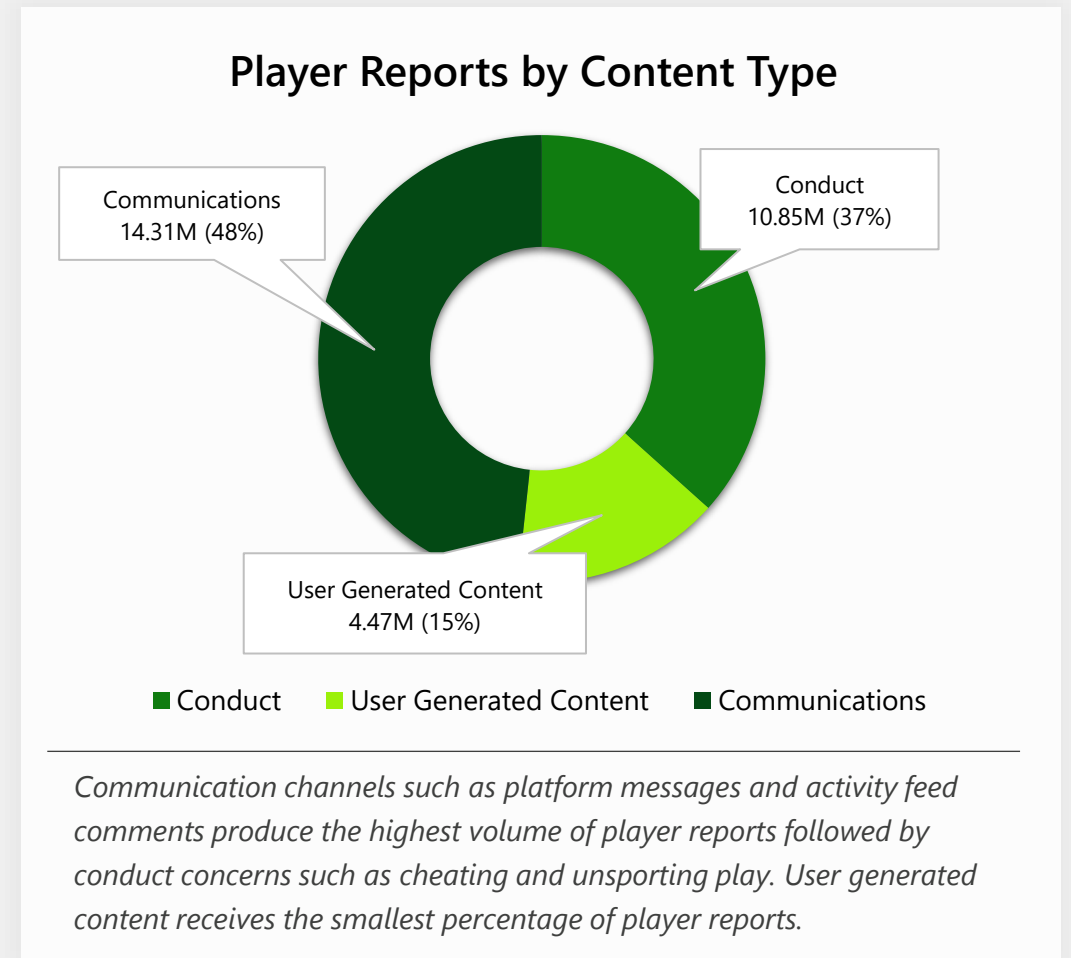
Reactive Moderation Data (Player Reported)

When a player brings something to our attention instead of being detected by our system, we consider that report to be reactive.

We classify player reported content in three main categories:

- 01** **Conduct** – The ways in which a player acts on Xbox including cheating, unsporting conduct such as griefing, teamkilling, etc.
- 02** **User Generated Content (UGC)** – Any content created by a player that isn't messaging related, such as a gamertag, club logo, or an uploaded screenshot or video clip.
- 03** **Communications** – Content related to communicating with other players such as a platform message or comment left on an activity feed post.

Below is a view of player reports based on the category of report:



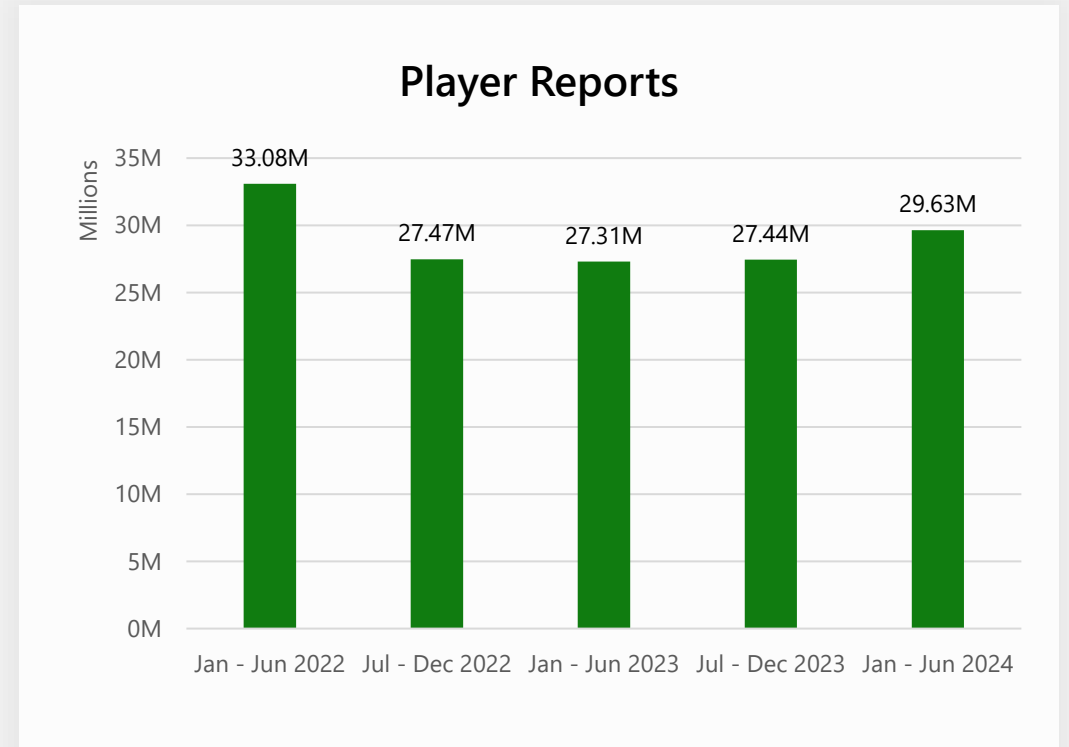
SHARING OUR SAFETY DATA

Player Reports

As player reports enter the system, they are often first evaluated by content moderation technologies to see if a violation can be determined, with the remainder reviewed by human content moderation agents for decision-making.

Content moderation agents are on-staff 24 hours a day, 7 days a week, 365 days a year to make sure the content and conduct found on our platform adheres to our [Community Standards](#).

Below you can find the number of reports submitted by players:



Driven by an uptick in spam and advertising, the rate at which players filed reports on the Xbox platform increased by 8% in the most recent reporting period.

SHARING OUR SAFETY DATA

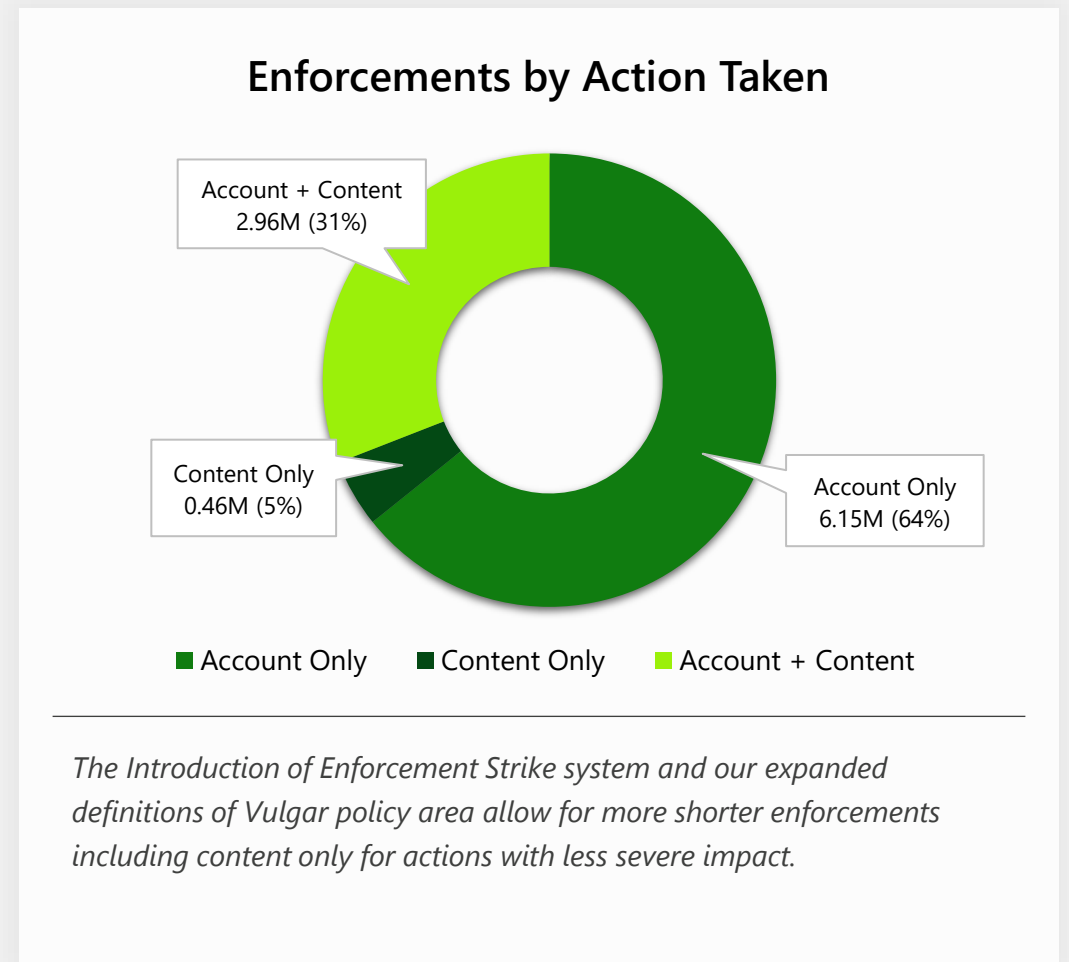
Enforcements Data

When a violation of our Community Standards is determined to have taken place, one of three things happens:

- 01** The content is removed (Content-Only Enforcement)
- 02** The player account is suspended (Account-Only Enforcement)
- 03** A combination of the two occurs (Account + Content Enforcement)

These actions are referred to as an enforcement.

Here we look at the types of enforcement actions taken during the first half of 2024:

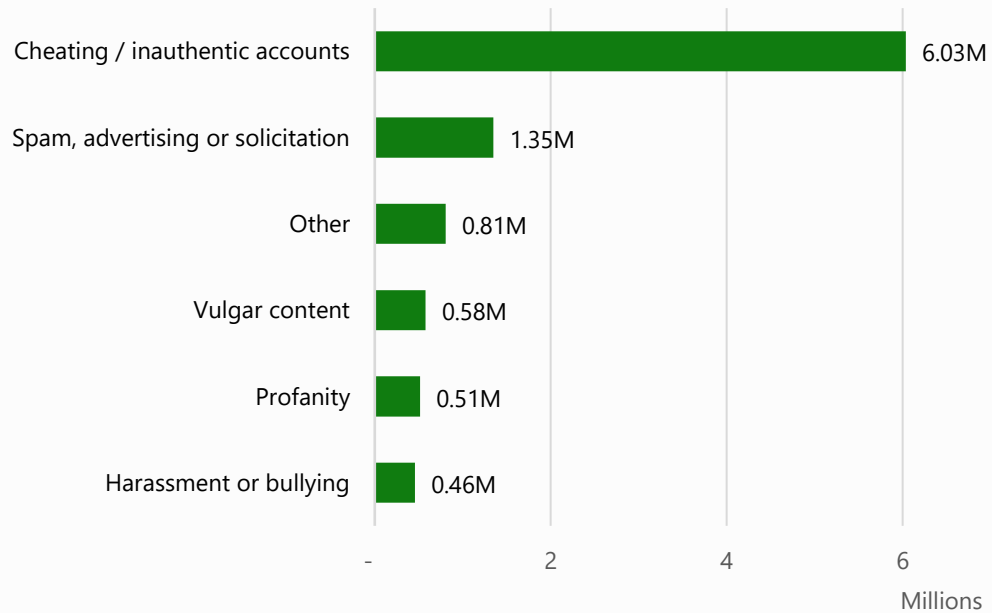


SHARING OUR SAFETY DATA

Most enforcements are categorized by the policy area where the violation occurred.

A breakdown of the most common areas of policy violation (from both proactive and reactive sources) can be seen below:

Total Enforcements by Policy Area



Cheating / inauthentic accounts is the area with the largest number of enforcements. Spam, vulgar content, profanity, and harassment or bullying are the other policy types that round out our top five. The Other category includes smaller volume areas such as piracy, adult sexual content, phishing, account tampering, or drugs.

SHARING OUR SAFETY DATA

Microsoft Digital Safety Content Report Data

As a US-based company, Microsoft reports all apparent Child Sexual Exploitation or Abuse Imagery ([CSEAI](#)) or grooming of children for sexual purposes to the National Center for Missing and Exploited Children ([NCMEC](#)) via the [CyberTipline](#), as required by US law.

In the period covered by this report, **606** of Microsoft's reports were from Xbox.

More information on Microsoft's efforts regarding CSEAI, grooming of children for sexual purposes, and terrorist and violent extremist content ([TVEC](#)) can be found in the [Digital Safety Content Report](#).

Crisis Text Line Data

The most common real-world concerns that we see on the platform have to do with threats of self-harm, which are handled with a referral to counseling services via the [Crisis Text Line](#).

In the period covered by this report, we sent **2,210** Crisis Text Line messages to players.

SHARING OUR SAFETY DATA

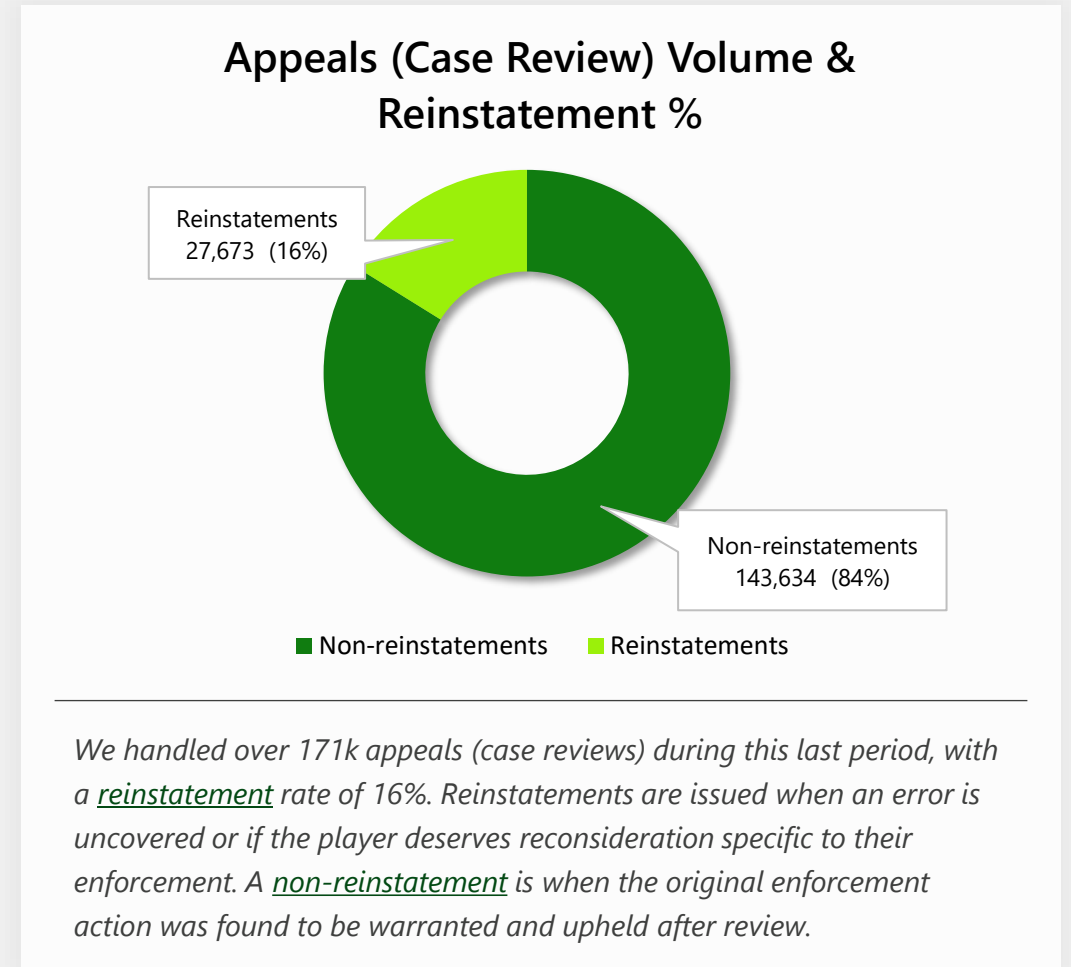
Appeals (Case Review) Data

When a player receives an enforcement beyond a certain length of time, they can dispute or ask for clarification through an appeal, otherwise known as a case review.

When filing a case review, the player can explain their actions and a moderation agent will review the case to see if an error was made or if special reconsideration is warranted.

During the last period, we handled **171k** appeal cases, down **4%** from the previous period. The reinstatement rate increased to **16%** due to an expanded policy for addressing unique circumstances and promoting positive engagement with the community.

Here we look at the volume of appeals handled and the associated percentage of accounts that were reinstated:



Data shown above covers the time period of January to June 2024

SHARING OUR SAFETY DATA

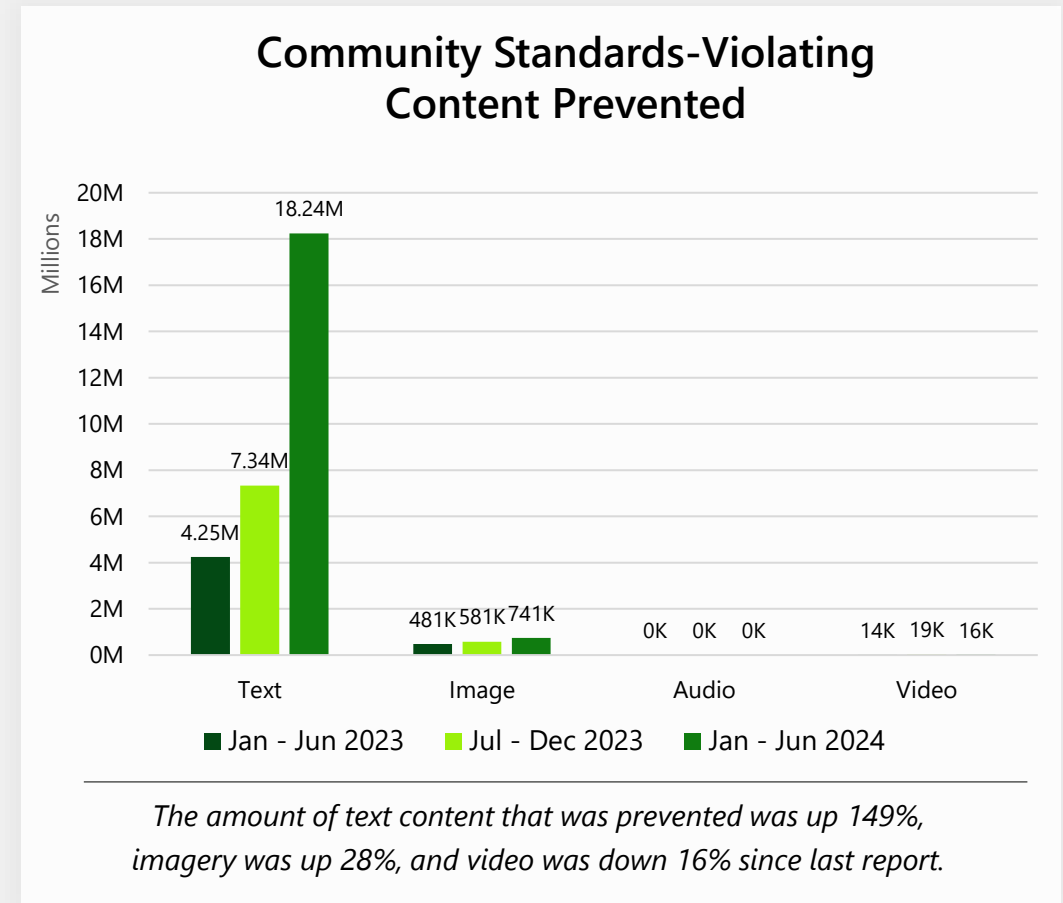
Toxicity Prevented

As we focus on creating safer experiences for our players, measuring the impact of the safety technologies we bring to bear becomes an important consideration. Keeping track of the amount of content that we prevent from entering or proliferating on our platform is one of the better indicators of the performance of the technologies we use in the safety space.

- **Text** – The amount of text, such as an inappropriate comment on someone’s activity feed, that we prevented.
- **Imagery** – The amount of imagery, such as an inappropriate gamerpic that was uploaded, that we prevented.
- **Video** – The amount of video, such as an inappropriate GameDVR clip, that we prevented.

We introduced a new approach to detect and intercept harmful messages between non-friends which led to a **149% increase** in inappropriate text prevention. The method strikes a balance between protecting players from harmful content sent by strangers and maintaining authentic experiences without imposing uniform enforcement.

Below you can find the amount of violating content that was prevented from entering or proliferating on our platform, broken down by content type:



POLICIES AND PRACTICES



POLICIES AND PRACTICES

Here is some supplemental information that may help you better understand the content of this report:



Policy & Standards

- [Xbox Community Standards](#)
- [Microsoft Services Agreement](#)



Reporting Process

- [How to report a player](#)



Appeals Process (Case Review)

- [How to submit a case review](#)



Glossary of Definitions

- [Definitions](#)



Additional Resources

- [Family & Online Safety](#)
- [Privacy & Online Safety](#)
- [Parental Controls](#)
- [Family Hub](#)
- [Responsible Gaming for All](#)
- [Learn about the Xbox Family Settings app](#)
- [Learn about safety settings for Xbox messaging](#)
- [Xbox Family Settings app](#)
- [Xbox Insiders Program](#)
- [Privacy dashboard](#)
- [Enforcement Strike System FAQ](#)
- [Reactive Voice Reporting](#)

GLOSSARY OF TERMS



GLOSSARY OF TERMS

Appeals (Case Review) – A mechanism through which a player that received an enforcement can find out more information as to the circumstances and appeal to have the enforcement removed or shortened

Case Review – See Appeals

CSEAI – Child Sexual Exploitation or Abuse Imagery

CyberTipline – The nation’s centralized reporting system for the online exploitation of children

DSCR (Digital Safety Content Report) – A half yearly report published by Microsoft that covers digital safety concerns. Found [here](#)

Enforcement – Action taken against a player, usually in the form of a temporary suspension which prevents the player from using certain features of the Xbox service

Inauthentic accounts – Throwaway accounts that are commonly used for purposes such as spam, fraud, cheating, or other actions that ultimately create an unlevel playing field for our players or detract from their experiences

NCII – Non-consensual intimate imagery

NCMEC – National Center for Missing & Exploited Children

Non-reinstatement – When a player appeals an enforcement action on their account and the original enforcement was found to be warranted

Player Report – When a player files a complaint or brings a policy violation to the attention of the Safety Team

Proactive Enforcement – When we action on inappropriate content or conduct before a player brings it to our attention

Reactive Enforcement – When we action on inappropriate content or conduct via a player bringing it to our attention

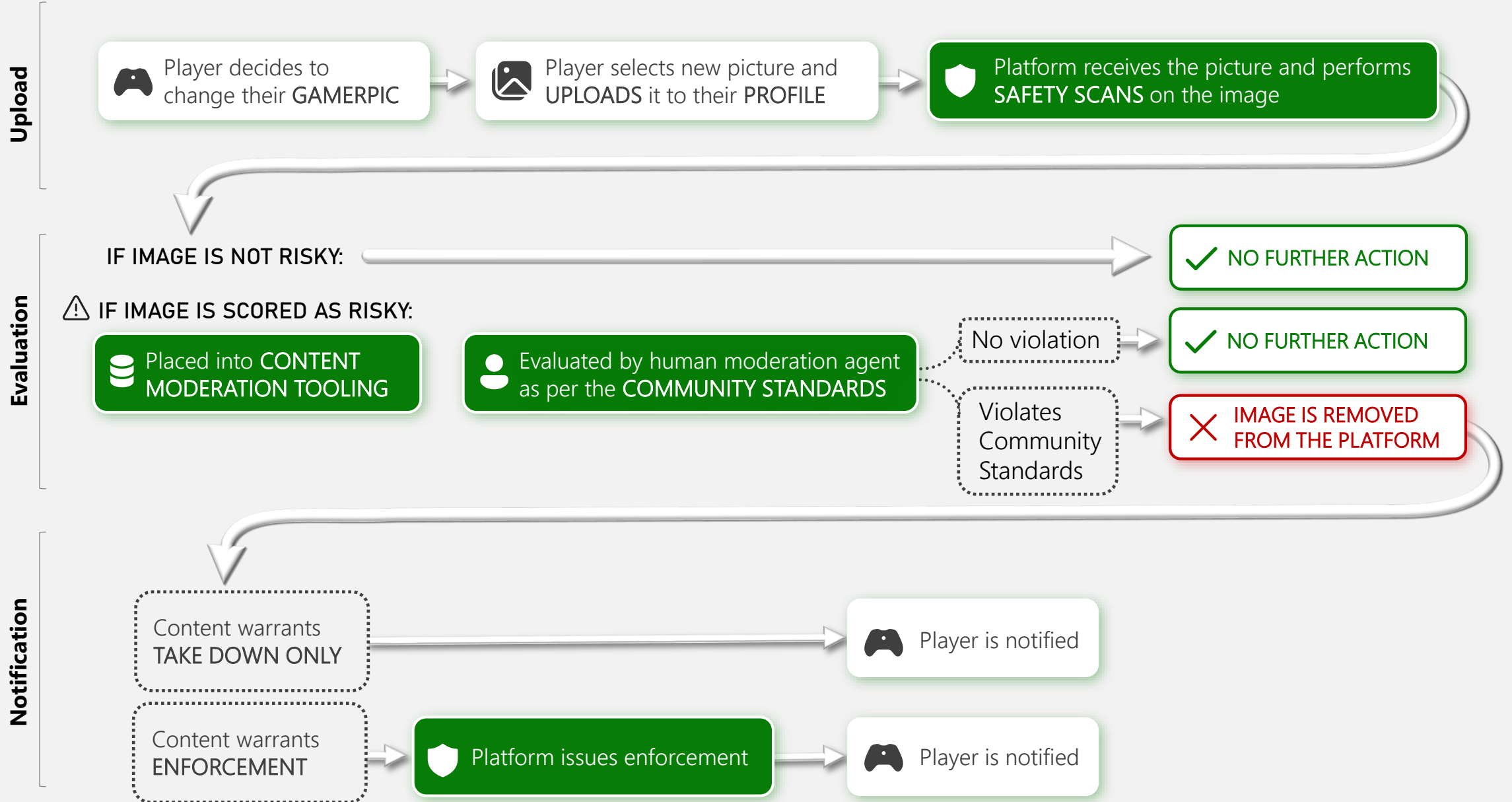
Reinstatement – When a player appeals a received enforcement, and their account is reinstated (enforcement is removed). This usually occurs due to an error, extenuating circumstances, or when compassion is shown

TVEC – Terrorist and Violent Extremist Content

APPENDIX



PLAYER IMAGE UPLOAD INFOGRAPHIC



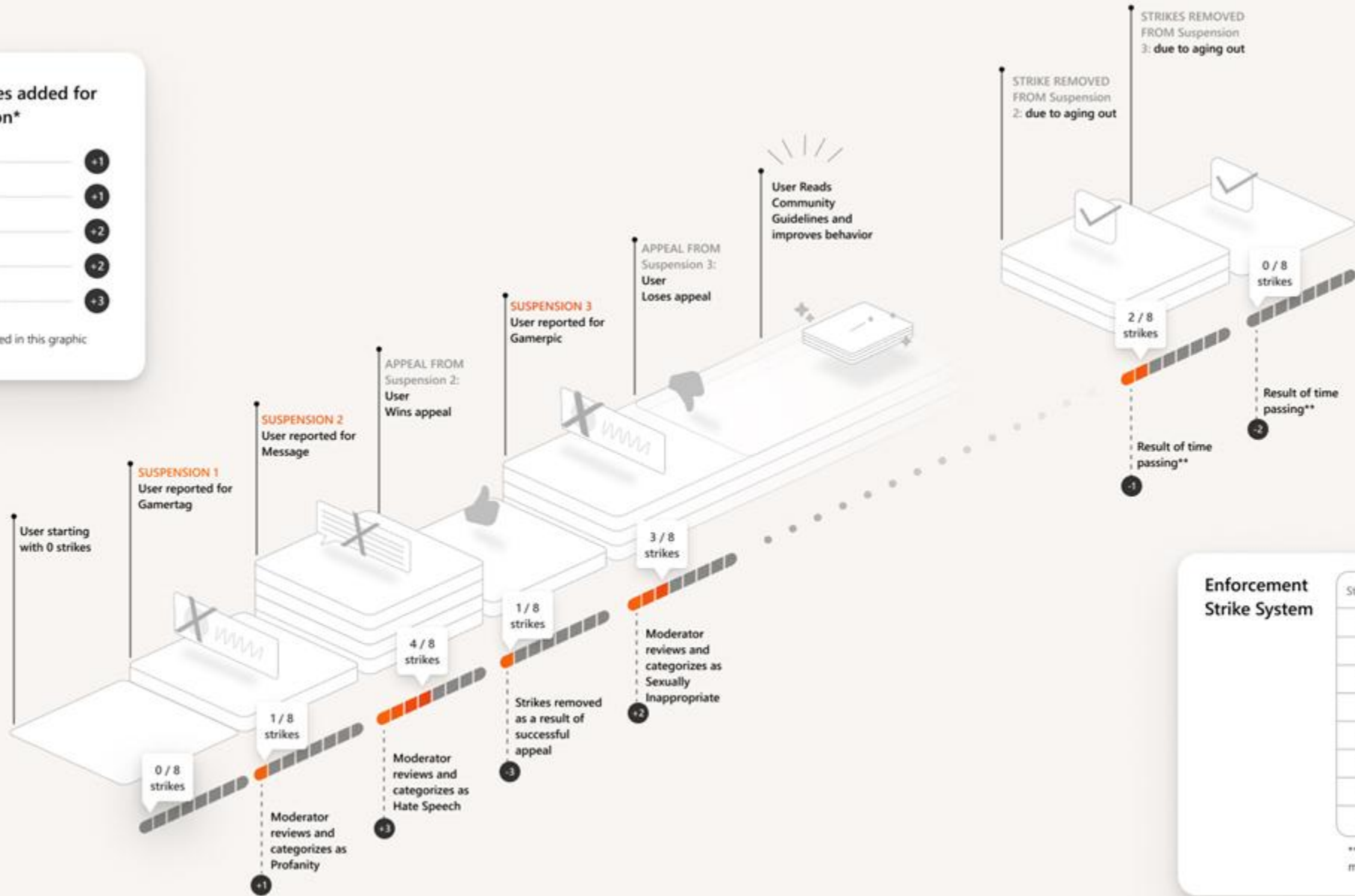
ENFORCEMENT STRIKE SYSTEM | USER JOURNEY INFOGRAPHIC



Examples of strikes added for each type of action*

- Profanity +1
- Cheating +1
- Sexually Inappropriate +2
- Harassment or Bullying +2
- Hate Speech +3

*not all actions are represented in this graphic



Enforcement Strike System

Strike	Suspension Length
1	1-day
2	1-day
3	3-days
4	7-days
5	14-days
6	21-days
7	60-days
8	365-days

**All strikes stay on record for 6 months