XBOX

2022 H2

TRANSPARENCY REPORT

TABLE OF CONTENTS

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.

What players don't often see are the expansive set of tools and investments that help us create safer experiences. Our safety suite includes AI-powered tools, as well as human-moderation tools that enable speed, precision and breadth to catch content before it ever reaches players. The data detailed in this Transparency Report demonstrates how our technology and proactive processes support our community – for example, almost 80% of all enforcements in the period were a result of proactive detection and we saw a 16.5x increase in proactive enforcements vs. this same period last year.

Our team frequently experiments and innovates on new features and methods that will further support our community. And as these roll out, we'll see the impact of that captured in future editions. This will help us learn and iterate, while also remaining transparent about our approach.

We are committed to an ongoing journey of learning and constant improvement. We continue to work closely with industry partners, associations, regulators, and community members to improve our multi-faceted safety strategy.

Key takeaways from the report

01 Proactive measures are a key driver for safer experiences

In this period, **80%** of our total enforcements issued were through our proactive moderation efforts. Our proactive moderation approach includes both automated and human measures that filter out and stop content before it can reach and impact players. Our use of AI-powered content moderation tooling, such as Community Sift, helps to identify offensive content within milliseconds, and in the last year alone has assessed over **20 Billion** human interactions on Xbox. Proactive efforts continue to be integral for providing safer online experiences.

02 Increased focus on inappropriate content

During this last period, we increased our definition of vulgar content to include offensive gestures, sexualized content, and crude humor. This type of content is generally viewed as distasteful and inappropriate, detracting from the core gaming experience for many of our players. This policy change, in conjunction with improvements to our image classifiers, have resulted in a **450%** increase in enforcements in vulgar content, with **90.2%** being proactively moderated (**+5.7%** from last report). These enforcements sometimes result in just removing the inappropriate content, which is reflected in the **390%** increase in "content-only" enforcements during this time period.

03 Continued emphasis on inauthentic accounts

Our proactive moderation, up **16.5x** from the same period last year*, allows us to shield players from negative content and conduct. The Xbox Safety team issued more than **7.51M** proactive enforcements against inauthentic accounts, representing **74%** of the total enforcements in the reporting period (up from **57%** last reporting period). Inauthentic accounts are typically automated or bot-created accounts that create an unlevel playing field and can detract from positive player experiences. We continue to make investments to address these accounts so players can have safe, positive, and inviting experiences.

* Jul-Dec 2022 time period vs Jul-Dec 2021

H2 2022 High Level Safety Data Summary (July – Dec 2022)

Player Reports

27.47M

12.97M (47%) Communications
11.40M (41%) Conduct
3.11M (11%) User Generated Content

Enforcements Issued

10.19M

8.08M (79%) Proactive[1]
2.11M (21%) Reactive[2]

NCMEC[3] Reports

549

Crisis Text Line Referrals

1,361

Appeals (Case Review)

229.87k

215.48k (94%) Non-Reinstatements
14.39k (6%) 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

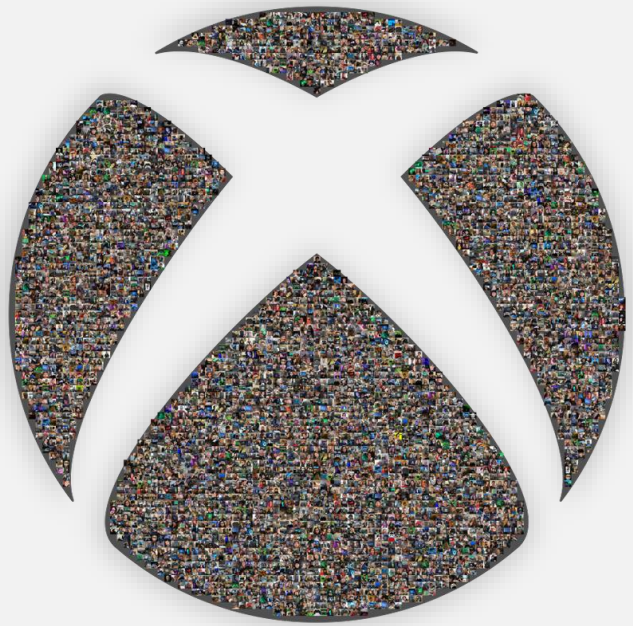
OUR VISION

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
- 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 dedicated team ensure safety is, and will always be, a priority for everyone.

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

Player Choice via Settings

We know that 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

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

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 block 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

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 helps to find resolution sooner, reduce the need for human review, and further reduce the impact of toxic content on human moderators. 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.

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 temporary 1-day, 3-day, 7-day, 14-day, or permanent suspension. The length of suspension is primarily based on the offending conduct or content, with repeated violations of the policies resulting in lengthier suspensions, an account being permanently banned from the service, or a potential device ban.

Learn about types of enforcements

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)

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.

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

SHARING OUR SAFETY DATA

The data that we'll be sharing below covers the time period between **July 1 – Dec 31, 2022** and was collected in accordance with Microsoft's commitment to privacy.

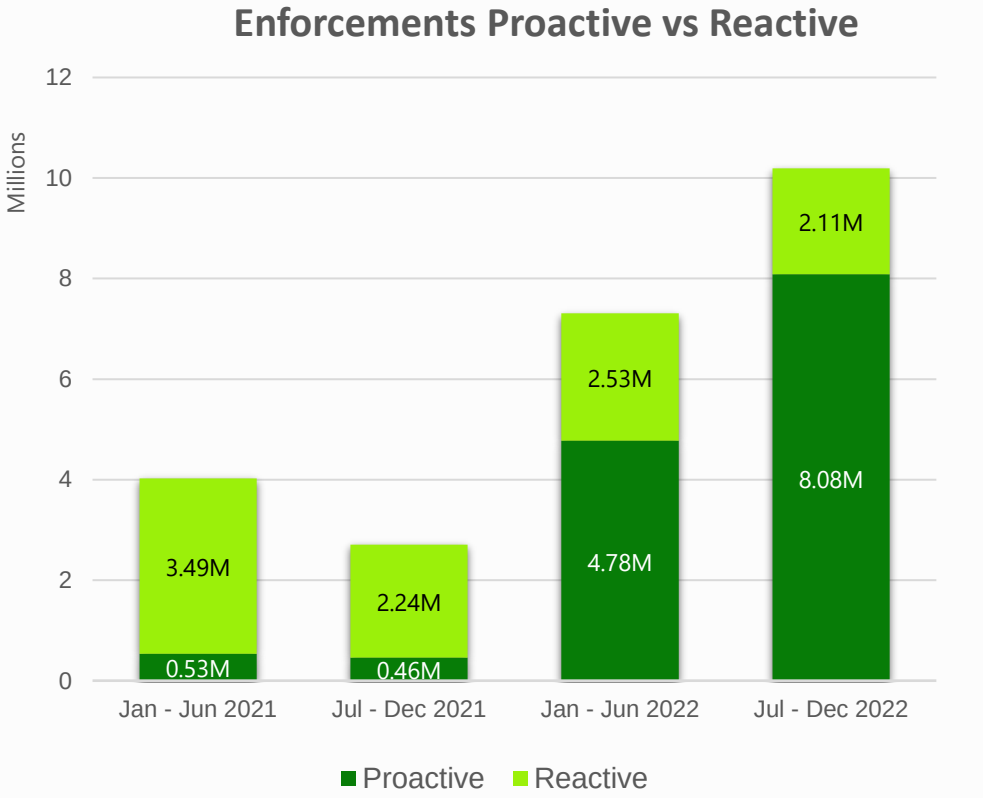
Proactive Moderation Data

Proactive enforcements are 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 last reporting period, we saw a 16.5x increase from last year in terms of proactive enforcements.

In looking at our proactive work over the last period, 7.51M of 8.08M enforcements (93%) were centered around detecting accounts that have been tampered with or are being used in inauthentic ways.

These accounts impact players in a myriad of ways including the production of unsolicited messages (spam), facilitation of cheating activities that disrupt play, improper inflation of friend/follower numbers, and other actions that ultimately create an unlevel playing field for our players or detract from their experiences.

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

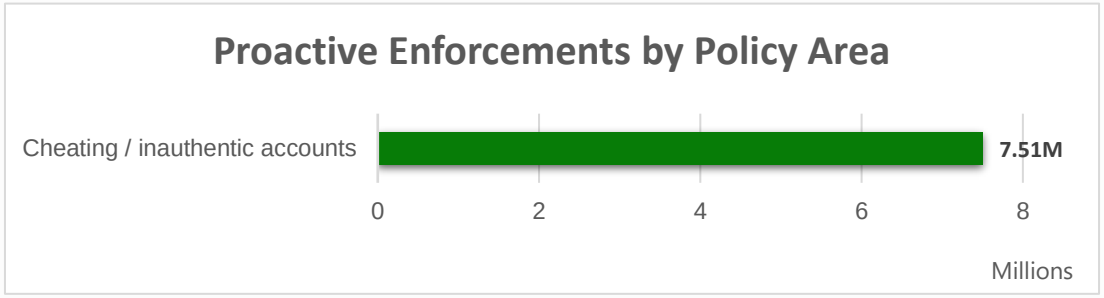


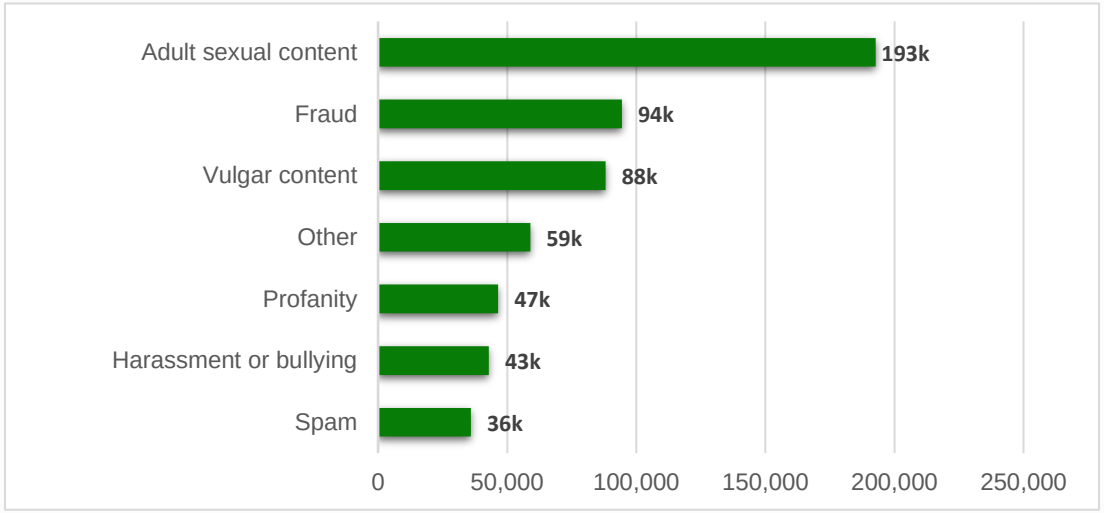
Approximately 80% of the enforcements issued in the second half of 2022 were a result of proactive detection.

Data shown above covers the time period of Jul-Dec 2022

SHARING OUR SAFETY DATA

We can break this down into policy areas for the previous 6-month period:



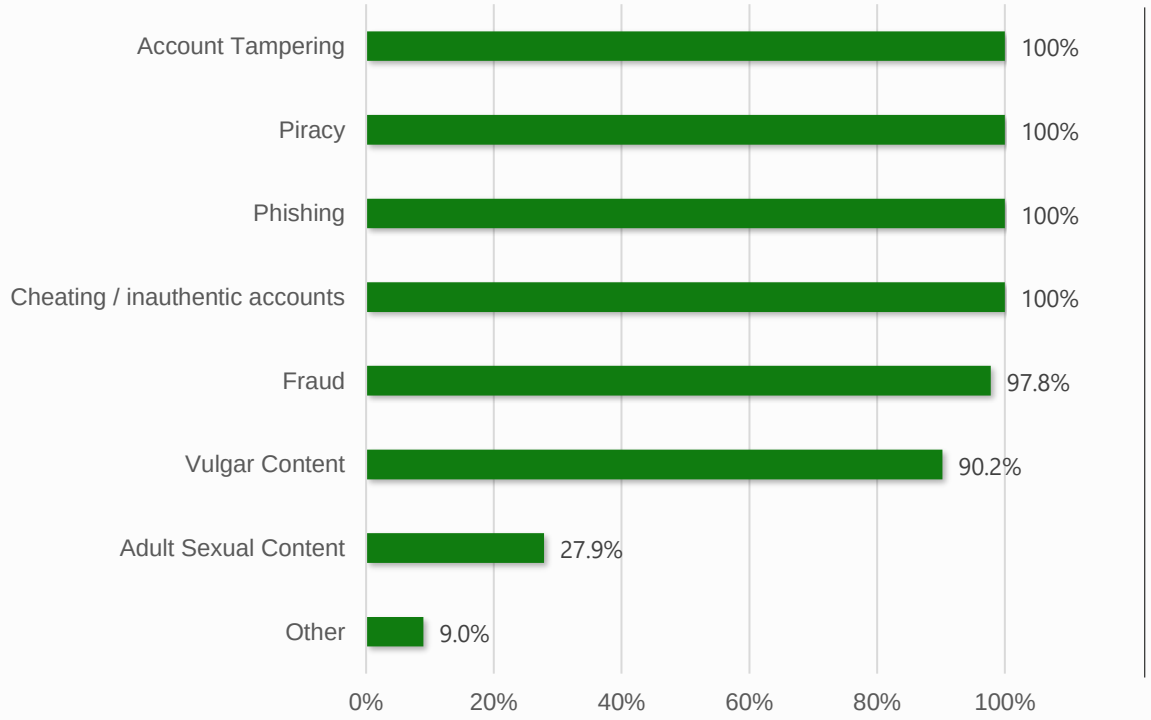


Beyond our focus on stopping inauthentic accounts as soon as they're created, the other areas that see high numbers of proactive enforcements include adult sexual content, fraud, vulgar content, profanity, harassment or bullying, and spam. The Other category includes smaller volume areas such as piracy,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

We must consider several factors when examining the number of proactive enforcements per policy area including the amount of that type of content on the platform, the efficacy of our proactive technologies at detecting that content, and whether we offer users personalized controls to self-govern their experiences in those areas (which produces reactive enforcements).

Data shown above covers the time period of Jul-Dec 2022

We can break this down further by looking at the % of enforcements that were issued proactively (before a player brought the issue to our attention) for the previous 6-month period:



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, phishing, and vulgar content. The Other category includes areas such as drugs, profanity, hate speech, harassment or bullying, spam, advertising, or solicitation.

Data shown above covers the time period of Jul-Dec 2022

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.

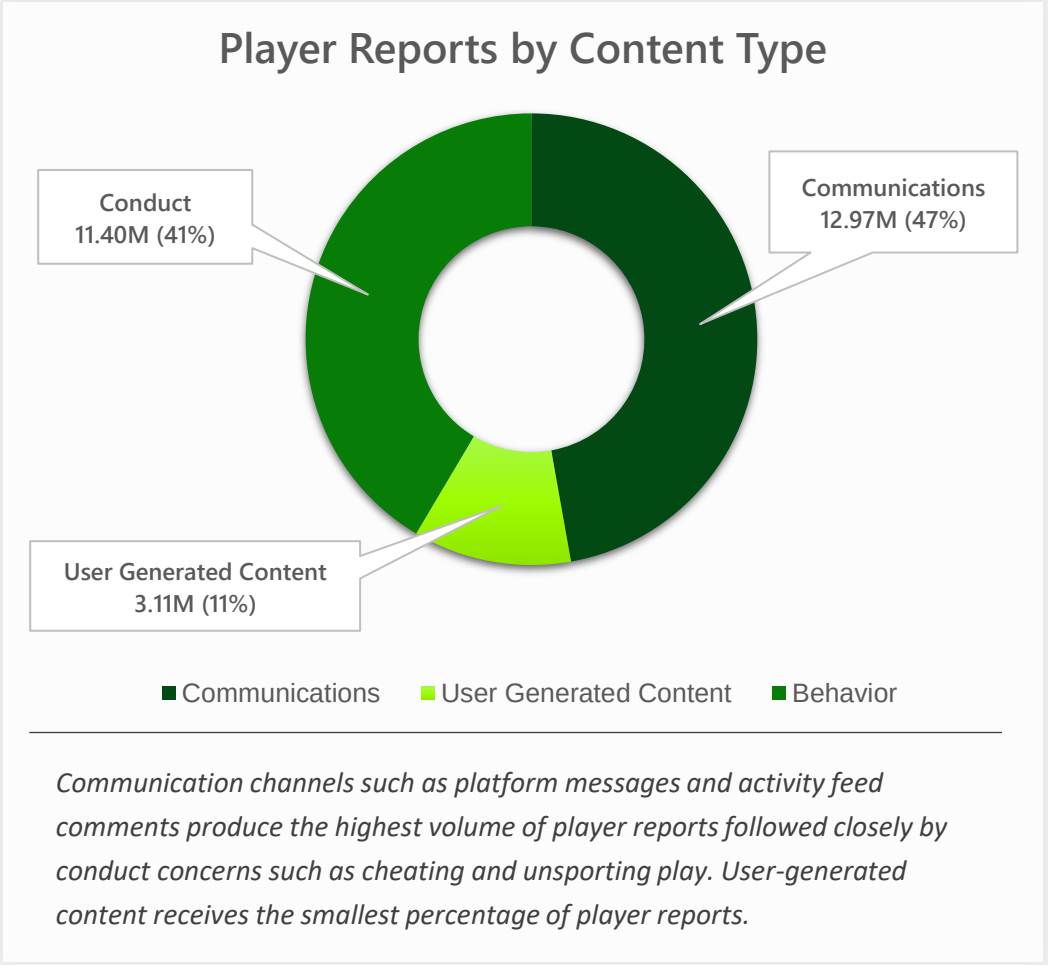
We classify player reported content into three main categories:

01 **Conduct** – The ways in which a player acts on Xbox including cheating, unsporting conduct, 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 user reports based on the category of report:



Communication channels such as platform messages and activity feed comments produce the highest volume of player reports followed closely by conduct concerns such as cheating and unsporting play. User-generated content receives the smallest percentage of player reports.

Data shown above covers the time period of Jul-Dec 2022

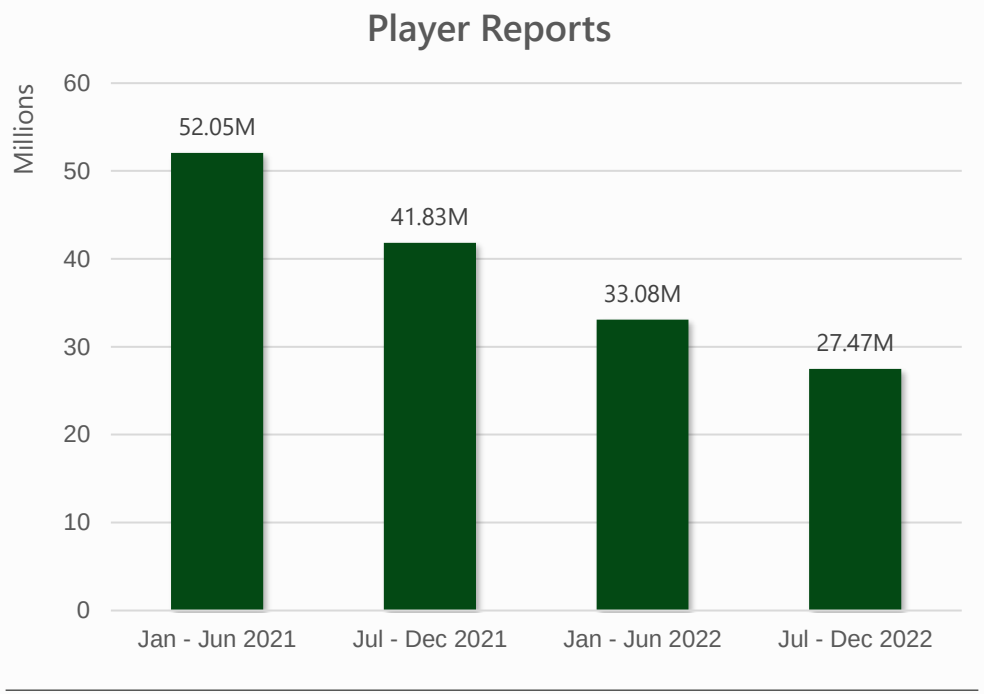
Player Reports

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.

This most recent reporting period saw a 34% decline in player reports from the same period in 2021.

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



The rate at which players file reports on the Xbox platform has been on a steady decline over the last several years, despite a growing player base.

Data shown above covers the time period of Jul-Dec 2022

Enforcements Data

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

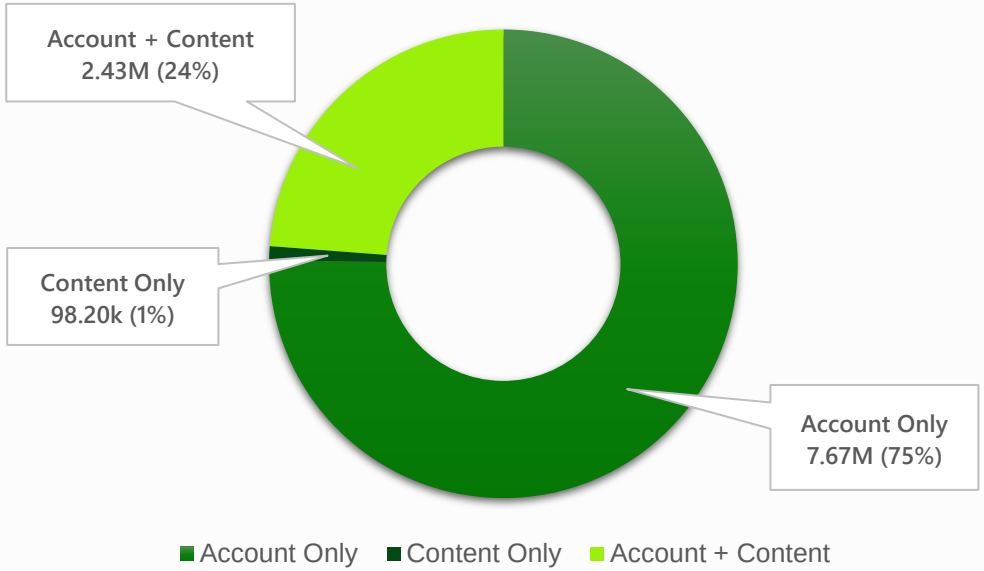
01 The content is removed (Content Only Enforcement)

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 last six months of 2022:



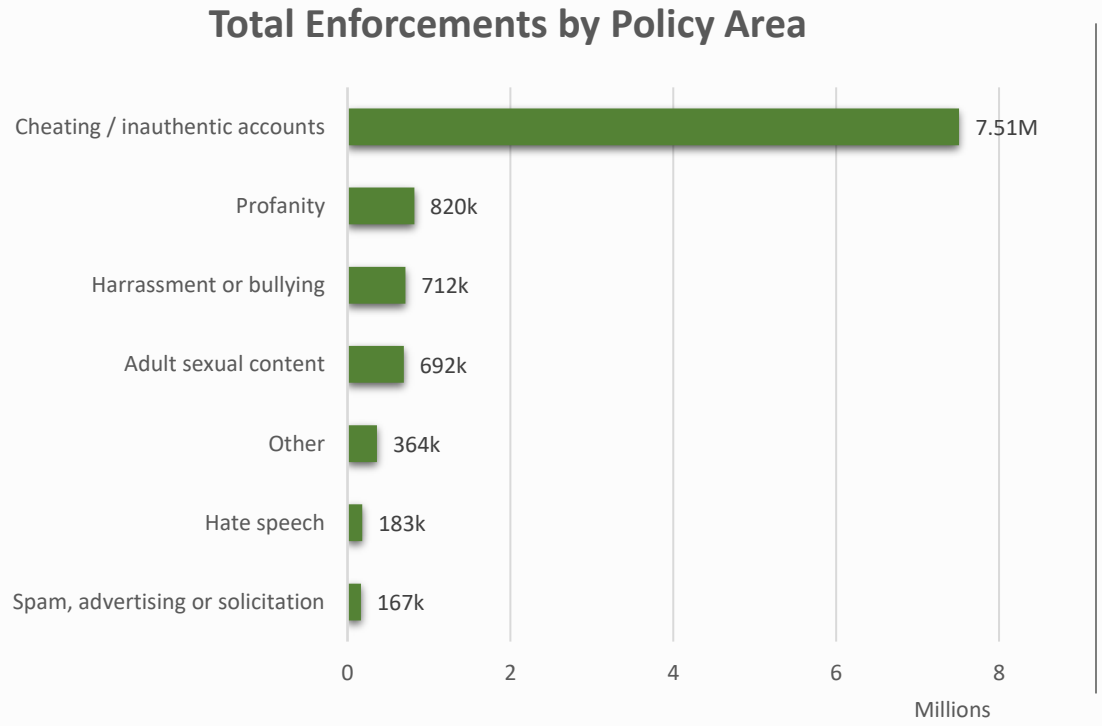
We've seen a significant increase in account-only enforcements issued in the first half of this year as we've ramped up proactive enforcements dealing with inauthentic accounts. These inauthentic accounts are often issued enforcements before they can add harmful content to the platform.

Data shown above covers the time period of Jul-Dec 2022

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:



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

Data shown above covers the time period of Jul-Dec 2022

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, 549 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 1,361 Crisis Text Line messages to players.

Appeals (Case Review) Data

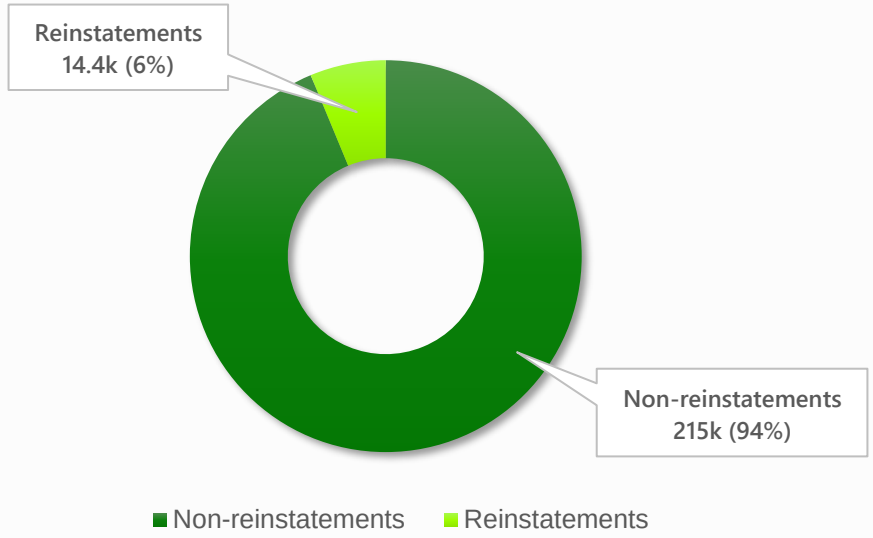
When a player receives an enforcement beyond a certain length of time, they can dispute or ask for clarification through an appeal, otherwise known as a case review.

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 over **230k** Appeals cases, up **51%** from the previous time period. This increase is partially driven by the increased focus on inappropriate content discussed in Key Takeaways point #2.

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

Appeals (Case Review) Volume & Reinstatement %



We handled over 230k appeals (case reviews) during this last period, with a reinstatement rate of approximately 6.3%. Reinstatements are issued when an error is uncovered or if the player deserves reconsideration specific to their enforcement. A non-reinstatement is when the original enforcement action was found to be warranted and upheld after review.

Data shown above covers the time period of Jul-Dec 2022

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

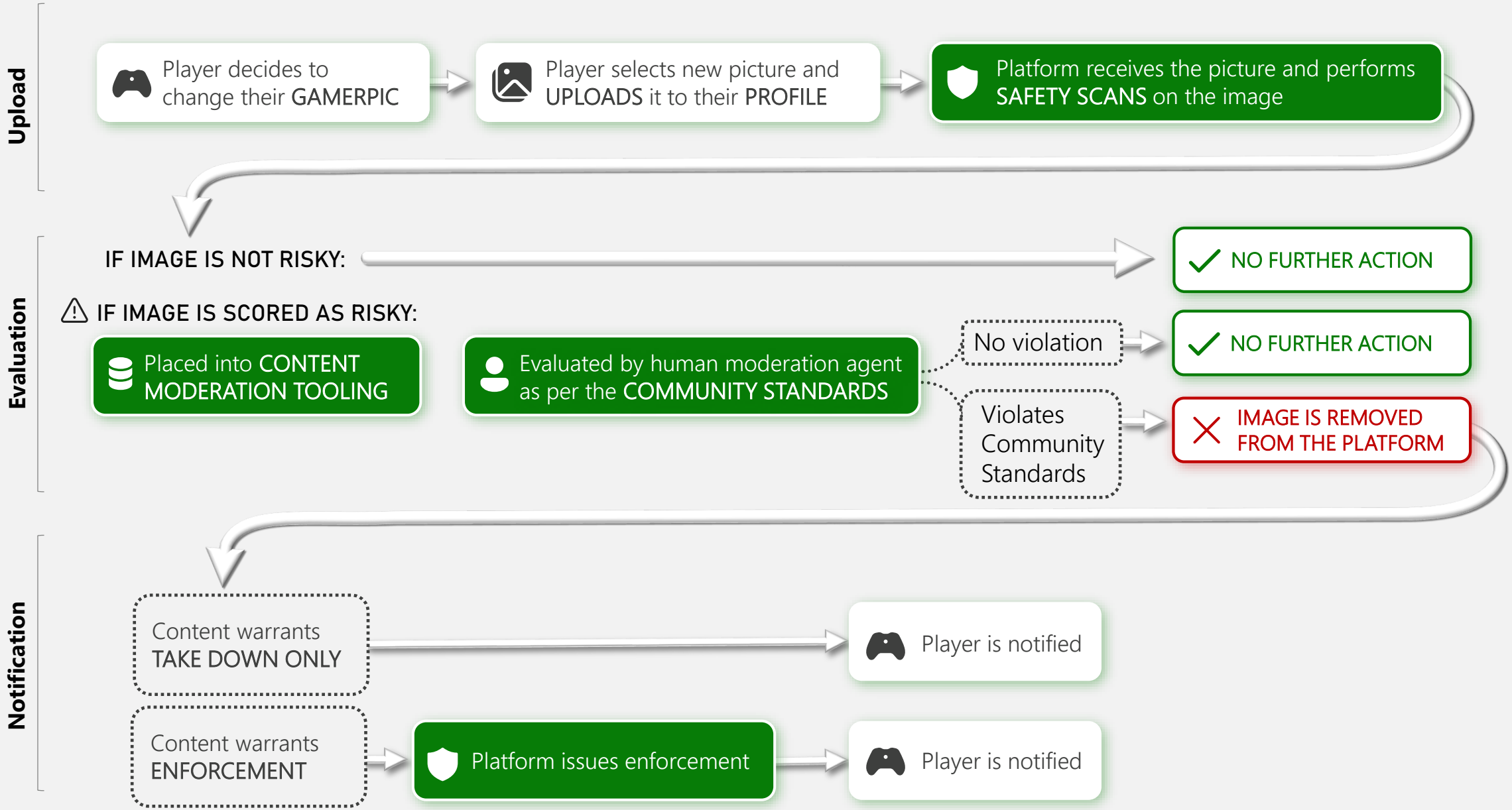
Glossary of Definitions

- 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

PLAYER IMAGE UPLOAD INFOGRAPHIC



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, or cheating

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 JOURNEY INFOGRAPHIC