2023 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.

With our fourth transparency report, we'll highlight some of the ways we're innovating to assure an even safer and more welcoming Xbox community. Microsoft's infrastructure, partnerships, and commitment to the ethical and responsible application of AI are providing us with the tools and expertise to evolve our approach.

As we continue to prioritize creating more inclusive, approachable, and safer gaming experiences for all players, our application of new AI advances continues to improve our ability to detect and prevent harm. At Xbox, we remain dedicated to advancing the responsible application of AI to amplify our human expertise in the detection of potential toxicity.

As always, we are deeply committed to working with industry partners, regulators, and our players to improve our multifaceted safety strategy. As we bring new technologies and features online, we'll cover their impact in future editions of this report. We remain committed to learning, iterating, and being transparent about our approach.

A Human-Centered Approach to Safety, Powered by AI

The evolution of the gaming landscape requires a dynamic and comprehensive approach to community safety. With the growing amount of content coupled by growing communities of players, solutions need to evolve and advance to continue to safeguard players across all the ways we communicate.

At Xbox, we believe the technological advances of generative AI, combined with human expertise, play complimentary roles in effectively identifying, reporting, and preventing harms at scale. We continue to improve upon our existing solutions, including Community Sift, so we can best match the current and anticipated needs of players and communities.

The application of generative AI is a preliminary step in categorizing and labeling text content, allowing teams at Xbox to both prevent known toxicity while focusing human moderation on more nuanced and complex communication. For images, we are implementing pattern matching capabilities that allow moderators to find and remove harmful images with specificity at large scale.

- **Auto Labeling** helps to classify conversational text by identifying words or phrases that match criteria and characteristics of potentially harmful content. This approach uses AI to analyze reported text content and helps community moderators quickly sort out false reports so human moderation efforts can be focused on content that is the most critical.
- **Image Pattern Matching,** powered by advanced database and image matching techniques enables rapid removal of known harmful content and identification of emergent toxic imagery.

Underpinning these new advancements is a system that relies on the expertise of humans to ensure the consistent and fair application of 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.

Key takeaways from the report

01 Effective New Ways of Protecting Players

Player behavior on Xbox voice chat has meaningfully improved since the launch of our new voice reporting feature. The feature has been effective in enabling players to report inappropriate verbal behavior with minimal impact to their gameplay. Since its launch, **138k** voice records have been captured utilizing our 'capture now, report later' system. When those reports resulted in an Enforcement Strike, **98%** of players showed improvement in their behavior and did not receive subsequent enforcements. Additionally, we have updated our proactive approach to more effectively prevent harmful content from reaching players (**3.2 million** more lines of text than last report, or **67%**), allowing players to engage in a positive way. We will continue to invest in features that protect and enhance the Xbox player experience.

02 Understanding of Enforcements Leads to a Safer Community

The Enforcement Strike System was launched last year to promote positive play while helping players understand the severity of a violation. Since its launch, **88%** of those who received an Enforcement Strike did not engage in activity that violated our Community Standards to receive another enforcement. We also reduced suspension lengths overall for minor offenses. Of enforcements that would have previously resulted in a 3-day or more suspension, **44%** were given a reduced length. The combination of these results shows that the majority of players choose to improve their behavior after only one suspension, even when it is short.

03 Blocking Inauthentic Accounts Before They Have Impact

We have been continuously investing in ways to spot inauthentic accounts, which has allowed us to quickly block many of them as soon as they're created, preventing them from affecting our players. Our improved methods prevented millions of inauthentic accounts from being used as soon as they are created and have led to a decline in the number of proactive enforcements on inauthentic accounts for the first time in two years, with enforcements dropping from 16.3 million in the last report to **7.3 million**.

H2 2023 High Level Safety Data Summary (Jul – Dec 2023)

Player Reports

27.44M

12.74M (47%) Communications
11.34M (41% Conduct
3.36M (12%) User Generated Content

Enforcements Issued

10.31M

8.14M (79%) Proactive[1]
2.17M (21%) Reactive[2] (From Player Reports)

NCMEC[3] Reports

674

Crisis Text Line Referrals

2,076

Appeals (Case Review)

179.35k

167.62k (93%) Non-Reinstatements
11.73k (7%) 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

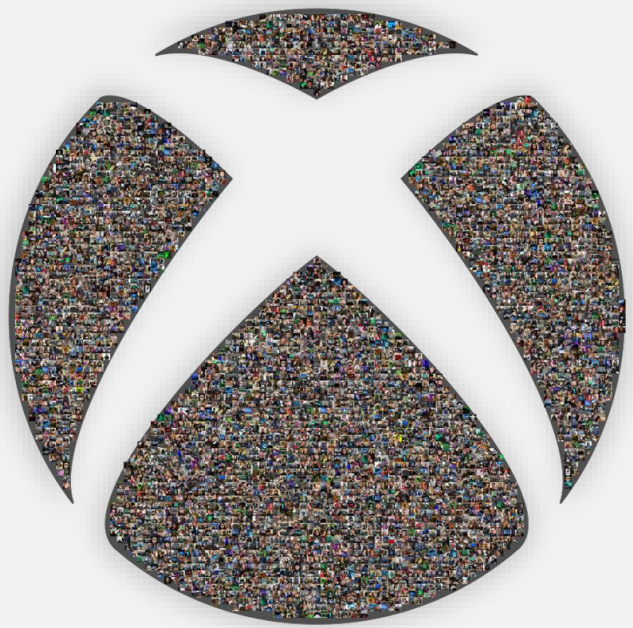
OUR VISION

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.

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 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 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

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)

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.

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
- Enforcement Strike System FAQ

SHARING OUR SAFETY DATA

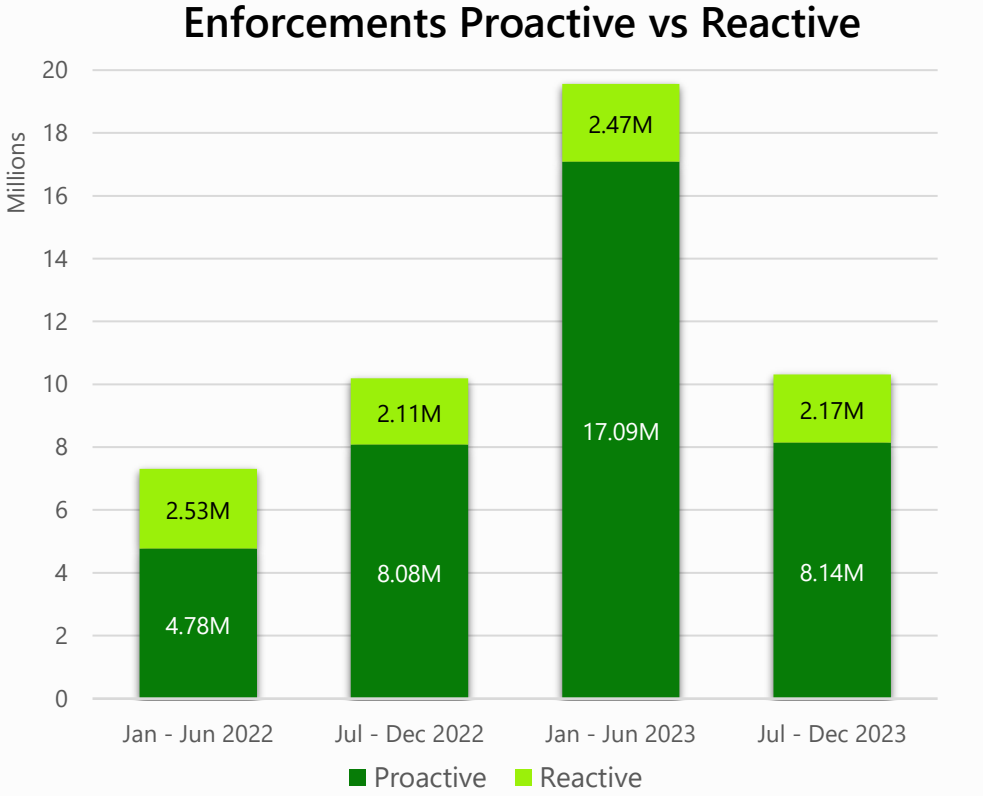
The data that we'll be sharing below covers the time period between **Jan 1 – Jun 30, 2023** and was collected in accordance with Microsoft's commitment to privacy.

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 reporting period, we saw the first decline from the previous reporting period for proactive enforcements from 17.09M to 8.14M.

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 and this remains a primary focus. Our proactive methodologies are always evolving to respond to dynamic challenges.

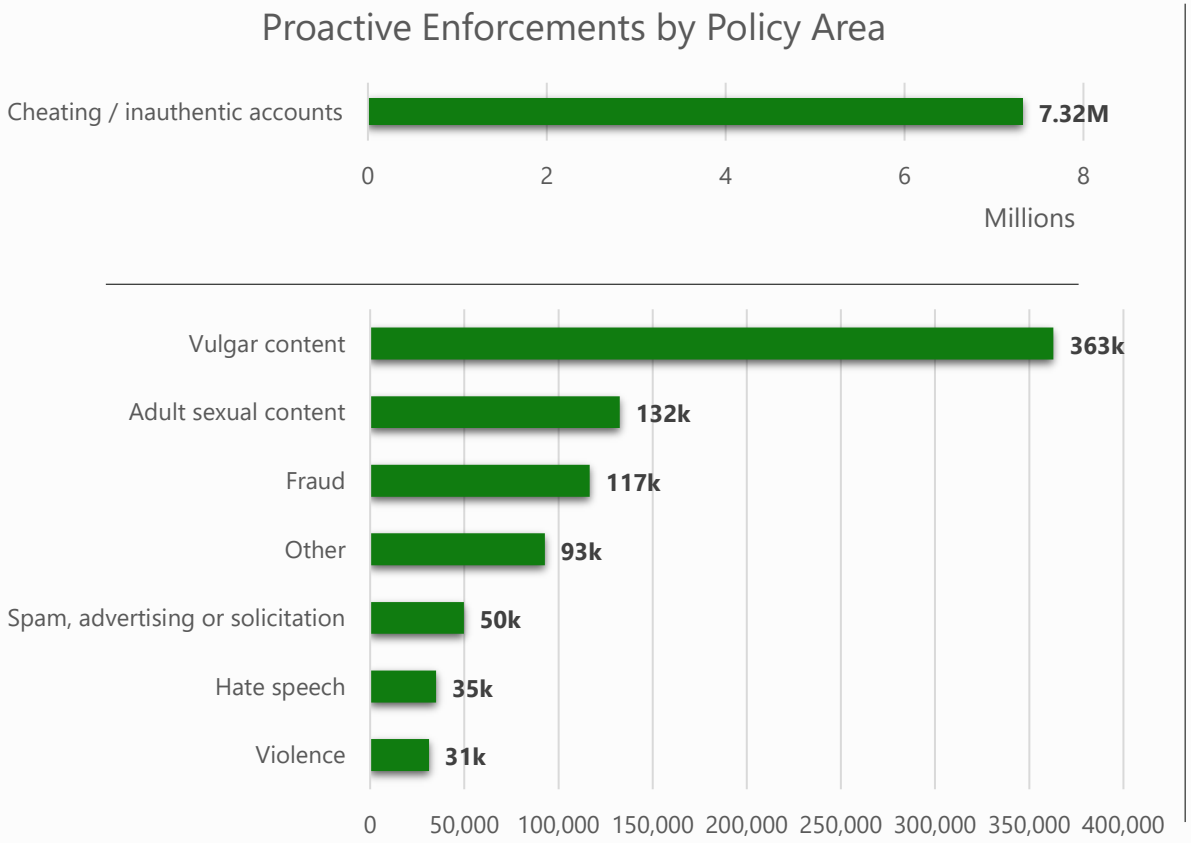
Below is a breakdown of proactive vs. reactive enforcements over time:



Approximately 79% of the enforcements issued in the second half of 2023 were a result of proactive detection.

SHARING OUR SAFETY DATA

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



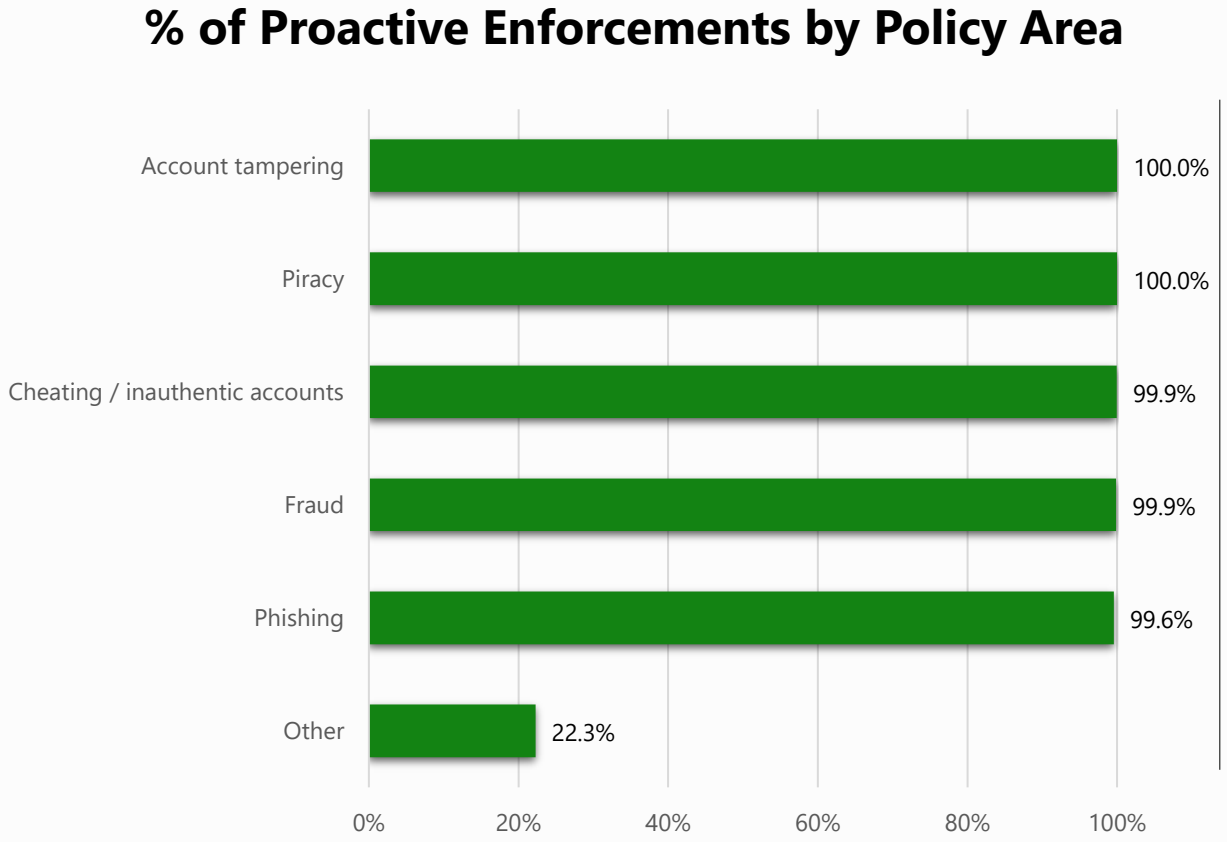
Beyond our focus on stopping inauthentic accounts as soon as they're created, the other areas that see high numbers of proactive enforcements include vulgar content, adult sexual content, fraud, spam, hate speech, and violence. The Other category includes smaller volume areas such as profanity, piracy,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

We saw a 70% increase in fraud proactive enforcements from new investigative processes targeting multiple types of fraud such as payment instrument fraud or account takeover. Updates to policy definitions saw much of the proactive enforcements for harassment or bullying shift into other categories including hate speech and violence which increased 33% and 38%.

Data shown above covers the time period of July-December 2023

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:



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.

Data shown above covers the time period of July-December 2023

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.

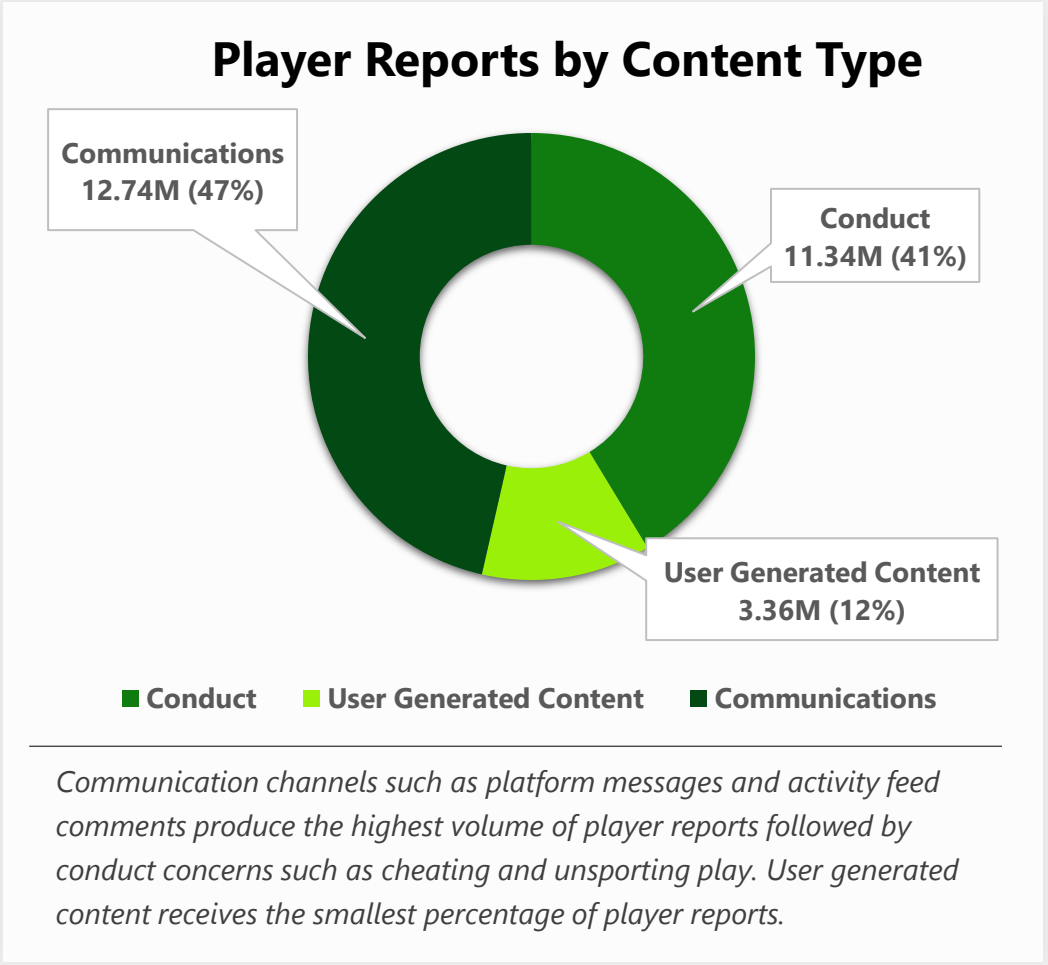
We classify player reported content in three main categories:

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:



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

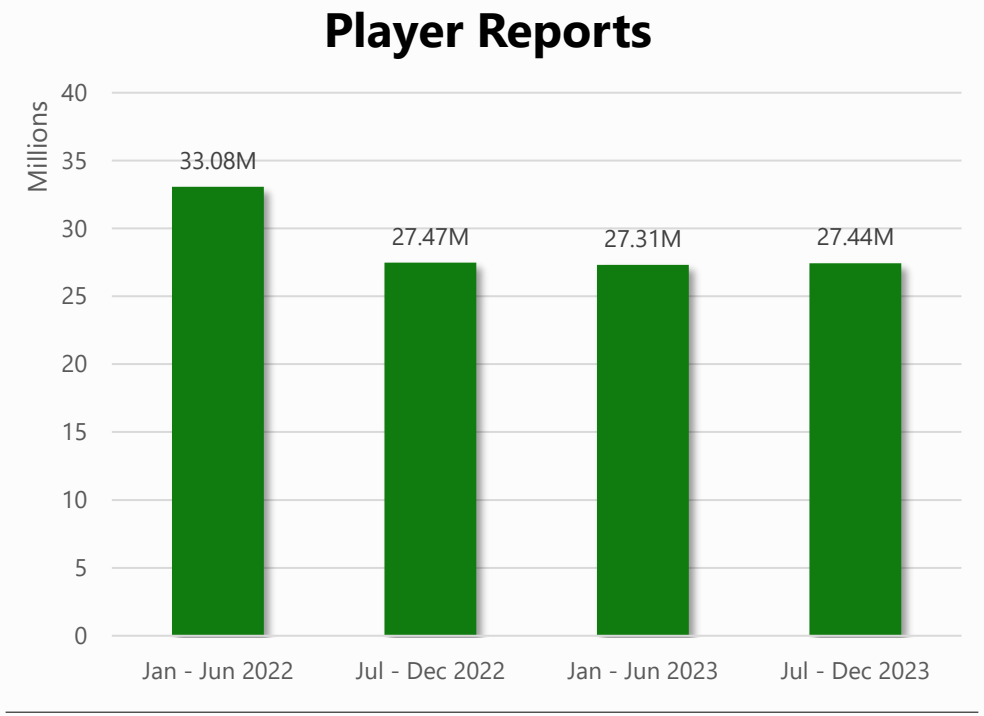
Data shown above covers the time period of July-December 2023

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:



The rate at which players file reports on the Xbox platform leveled in the most recent reporting period.

Enforcements Data

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

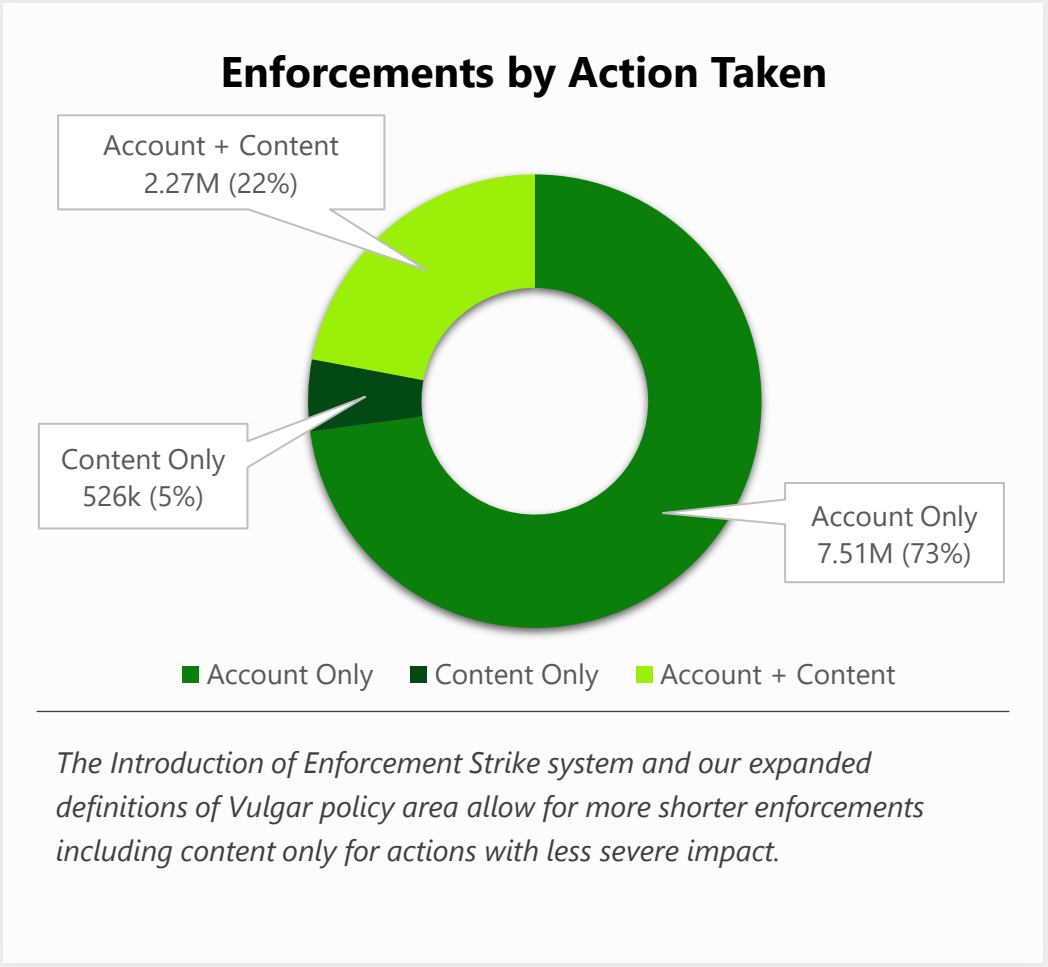
01 The content is removed (Content-Only Enforcement)

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 half of 2023:



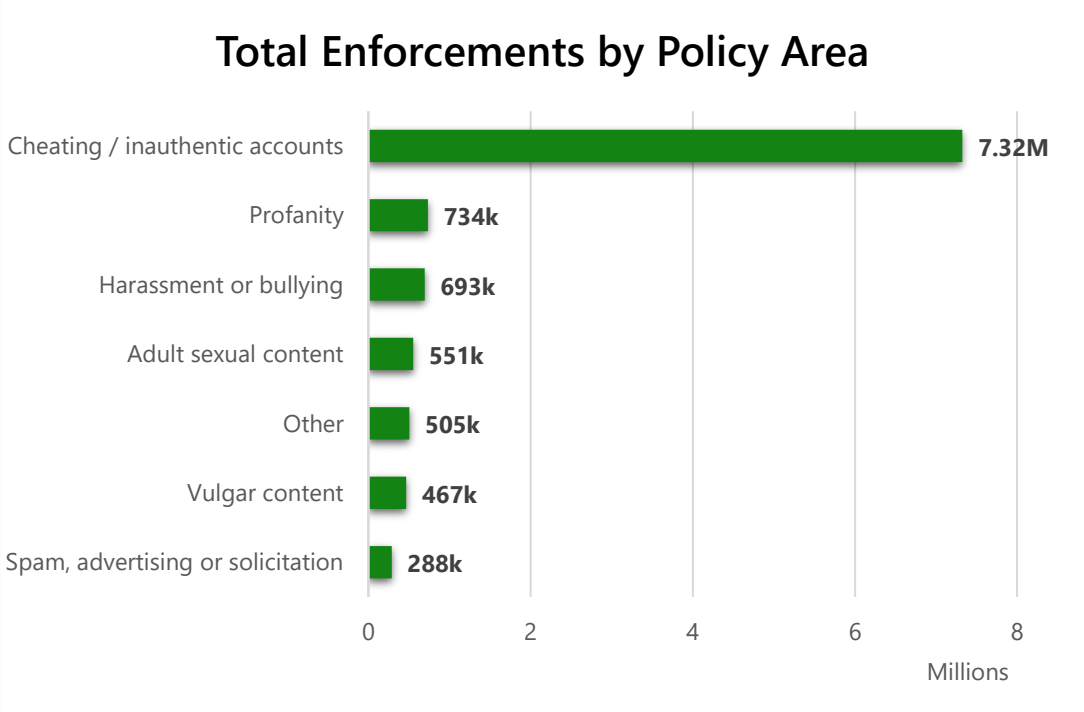
The Introduction of Enforcement Strike system and our expanded definitions of Vulgar policy area allow for more shorter enforcements including content only for actions with less severe impact.

Data shown above covers the time period of July-December 2023

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

Data shown above covers the time period of July-December 2023

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, **674** 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 **2,076** Crisis Text Line messages to players.

Appeals (Case Review) Data

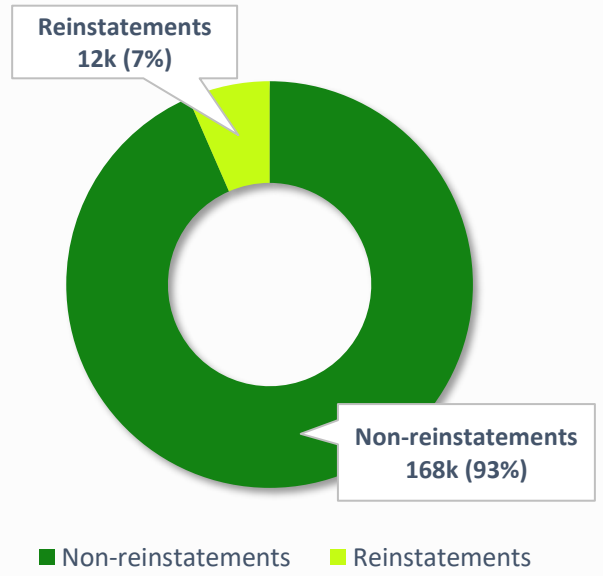
When a player receives an enforcement beyond a certain length of time, they can dispute or ask for clarification through an appeal, otherwise known as a case review.

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 **179k** appeal cases, down **36%** from the previous period. The reinstatement rate increased to 7% 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:

Appeals (Case Review) Volume & Reinstatement %



We handled over 179k appeals (case reviews) during this last period, with a reinstatement rate of 7%. 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 July-December 2023

Toxicity Prevented

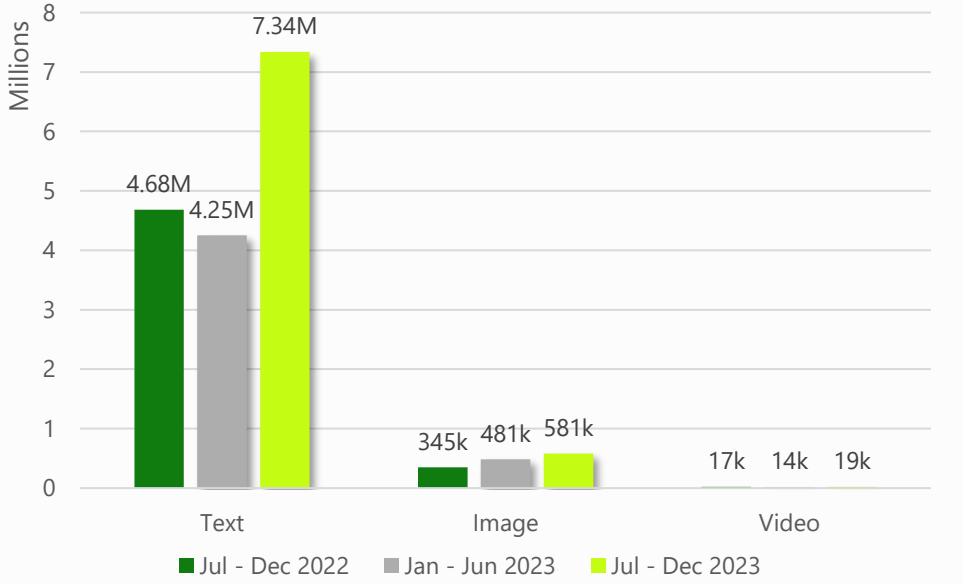
As we focus on creating safer experiences for our players, measuring the impact of the safety technologies we bring to bear becomes an important consideration. Keeping track of the amount of content that we prevent from entering or proliferating on our platform is one of the better indicators of the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

Our increased efforts in proactive moderation led to a **73% increase** in inappropriate text prevention.

Additionally, the continued **increase of 21%** in preventing inappropriate imagery is the result of investments in advanced technology such as Turing Bletchley v3 foundation model.

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



The amount of text content that was prevented was up 73%, imagery was up 21%, and video was up 36% period over period.

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

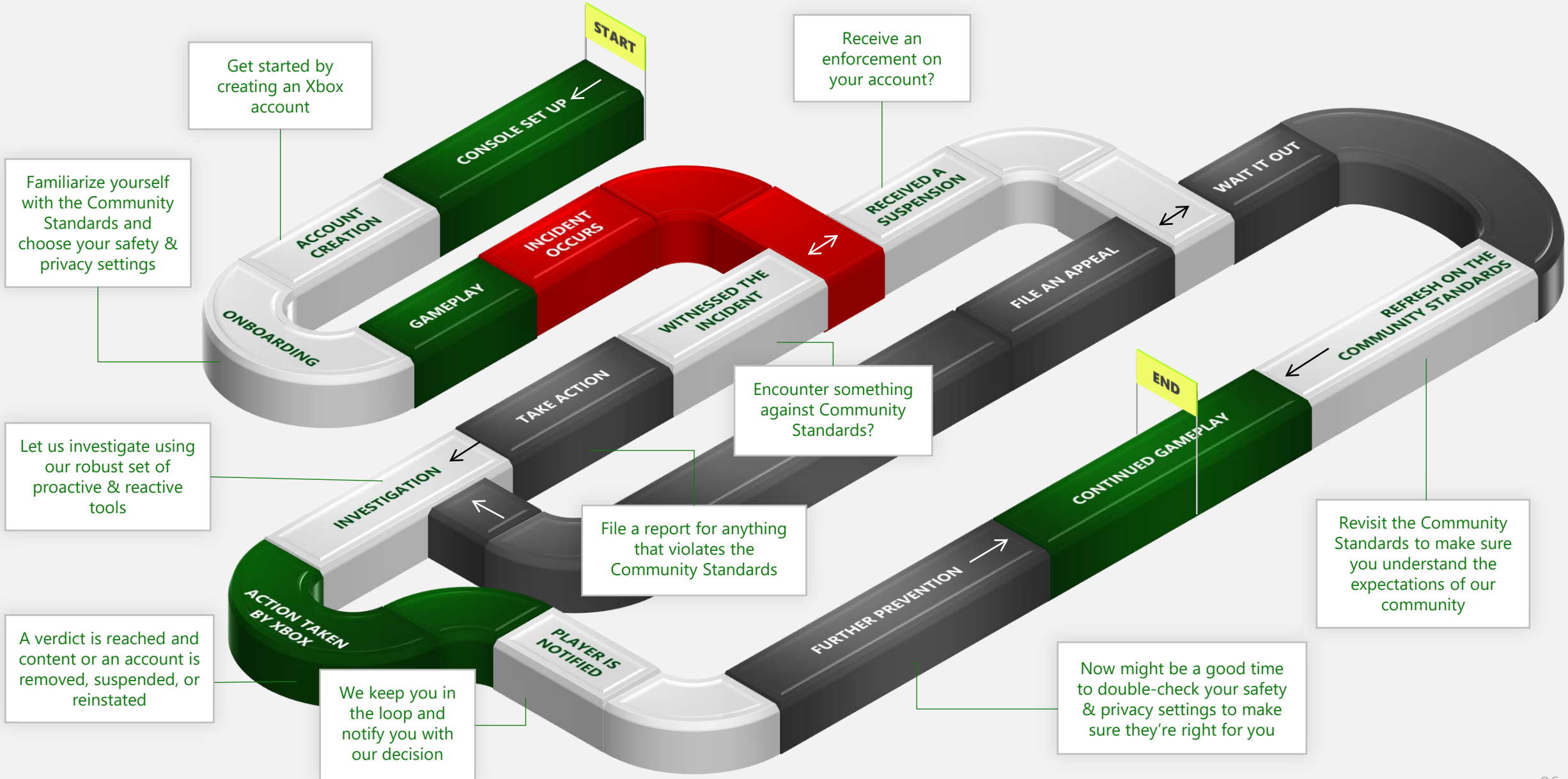
Reactive Enforcement – When we action on inappropriate content or conduct via a player bringing it to our attention

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

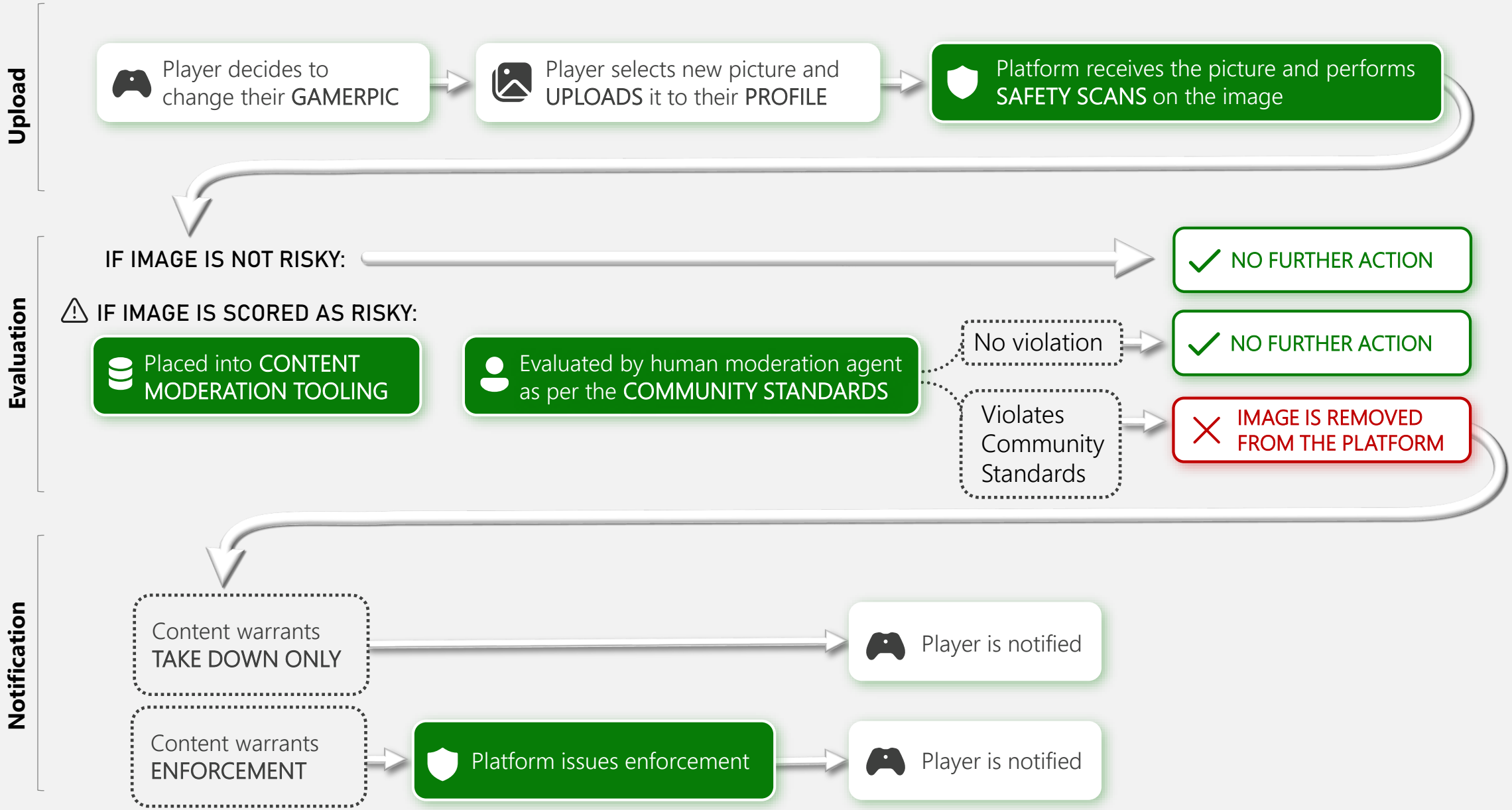
TVEC – Terrorist and Violent Extremist Content

APPENDIX

Player Journey Infographic



PLAYER IMAGE UPLOAD INFOGRAPHIC

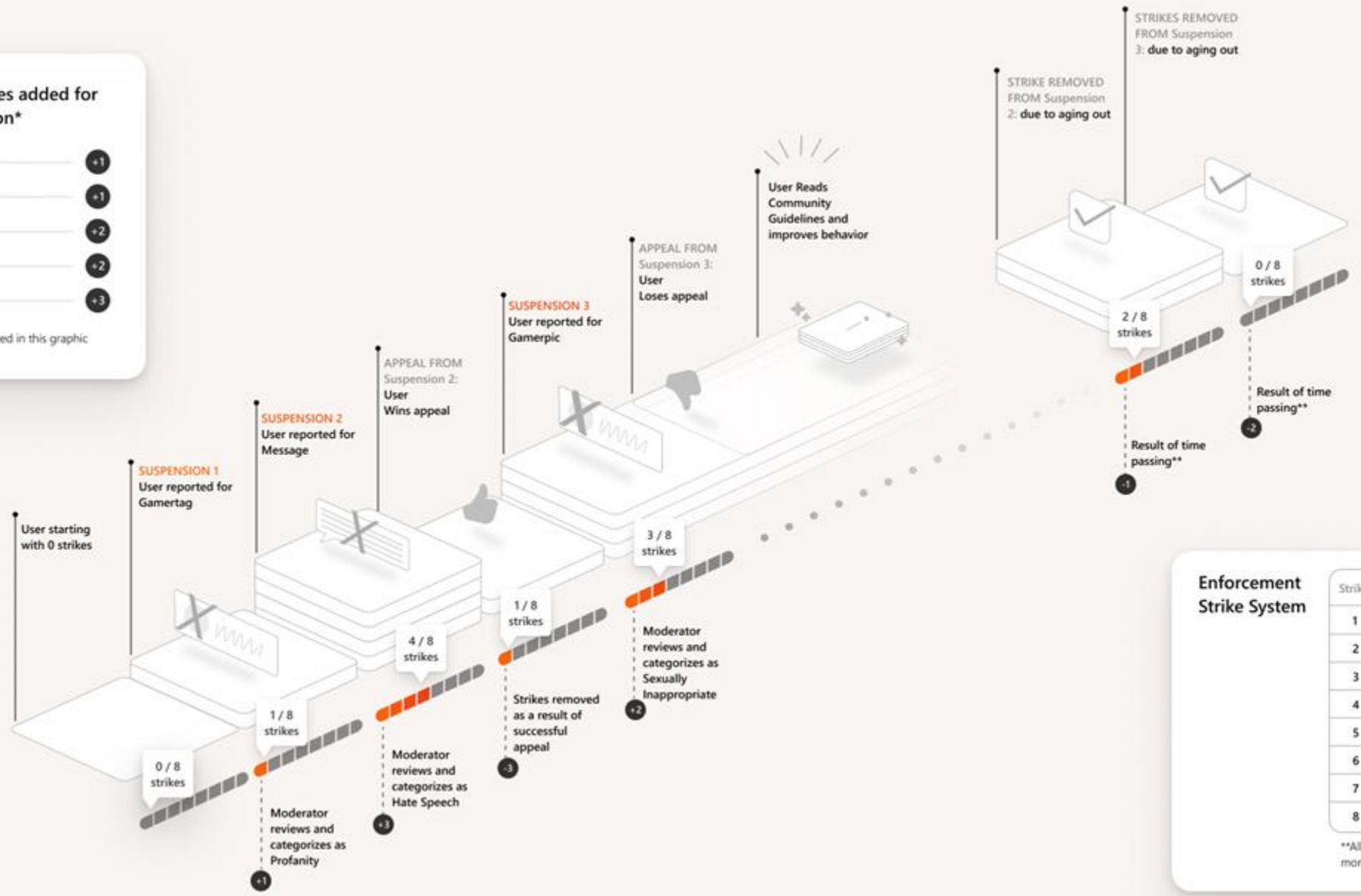


ENFORCEMENT STRIKE SYSTEM | USER JOURNEY INFOGRAPHIC

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