

DEC
2020



Disruption and Harms in Online Gaming Framework

Building a Penalty and Reporting System

ADL Center for Technology & Society

In a world riddled with cyberhate, online harassment and misuses of technology, the Center for Technology & Society (CTS) serves as a resource to tech platforms and develops proactive solutions. Launched in 2017 and headquartered in Silicon Valley, CTS aims for global impacts and applications in an increasingly borderless space.

It is a force for innovation, producing cutting-edge research to enable online civility, protect vulnerable populations, support digital citizenship and engage youth. CTS builds on ADL's experience over more than a century building a world without hate and supplies the tools to make that a possibility both online and offline.

Fair Play Alliance

The Fair Play Alliance is a global coalition of gaming professionals and companies committed to developing quality games. We provide a forum for gaming professionals and companies to work together to develop and share best practices in encouraging healthy communities and awesome player interactions in online gaming.

We envision a world where games are free of harassment, discrimination, and abuse, and where players can express themselves through play.

Where to Learn More

Please visit our resource hub for more resources:

fairplayalliance.org/resources

For developers, by developers. The FPA is an industry-lead alliance here to help. Visit www.fairplayalliance.org if you would like to access any of our resources, or reach out to info@fairplayalliance.org for support from any of our resident experts in player dynamics or to learn more about how you can help.

**Disruption and Harms in
Online Gaming Framework**

Building a Penalty and Reporting System

Building a Penalty and Reporting System

Penalty and reporting (P&R) systems are challenging because the validity of a report depends on both the situation and the reporting player's mindset. It is difficult to determine whether the behavior was problematic, intentional or a misunderstanding. Reports can indicate a disagreement where neither party was misbehaving, but both felt that the other was inappropriate. If the behavior is problematic, assessing its severity and applying the right response is even more challenging. This guide looks at some of the critical aspects of building out a P&R system.

Stop! Before going further, it is useful to review the [Planning a Penalty & Reporting System](#) resource.

Note: Designing and building a P&R system can seem daunting for small and large studios. Thankfully, there are third-party and platform moderation options available, making it much easier to access an API rather than build and support all of these systems.

Introduction and Overview

There are several stages to any P&R system. Once a player files a report against another player, they must be notified the matter is under review. The notification lets players know that their reports are taken seriously and that your studio wants to maintain a healthy environment.

Upon receiving a report, studios should investigate what transpired and apply the appropriate penalty, if warranted. There should be a means to enforce that penalty for the appropriate duration (e.g., if it is a content access restriction, such as a cooldown, then that check must be enabled for the offending player). And there should be a path back to good standing. If an offense is severe, your studio may consider exiting a player from the community (often referred to as a permanent ban, or more colloquially as the "ban hammer"). If a player continues to offend, you might institute an escalation ladder of progressively severe restrictions that include removing the player.





Diagram 1. Penalty/reporting lifecycle.

Player Reporting

By providing an avenue for reporting, you help control where and how players reach out. Doing so allows you to better understand players' experiences and offer support, including collecting metrics on your players' satisfaction and your ability to resolve their concerns.

Estimates for the number of reports submitted by players vary, but are typically in the range of 5-10% of your active player base. Approximately 1 to 2% are actionable, and 0.1% are serious infractions. Note that numbers outside of this range may indicate problems (see under- and over-reporting below).

There are a few critical considerations for helping players report effectively:



Discoverability. Are the avenues for reporting discoverable and accessible when players need them? For example, if a player is forced to email player support rather than use an in-game option to report, they may forget or not want to bother. This can impact your ability to assess what is truly affecting players. When reporting options are easily available, relevant players are more likely to use them and there is little evidence of overuse (though take care to reduce the chance of misclicks!).



Think carefully about the form of reports. If you are designing a reporting interface, think carefully about how you ask players for information. If you provide a freeform text box, will you get the information you need at scale? Do you have enough staff to read each report thoroughly? And how do you define “thorough”? Allow players to highlight problematic content when reporting, particularly in text logs.

Disproportionately affected. Are you able to understand the impact on vulnerable groups, such as children or people of color, based on reports submitted? Allow players to report if they feel they have been a victim of identity-based or hate-based harassment.

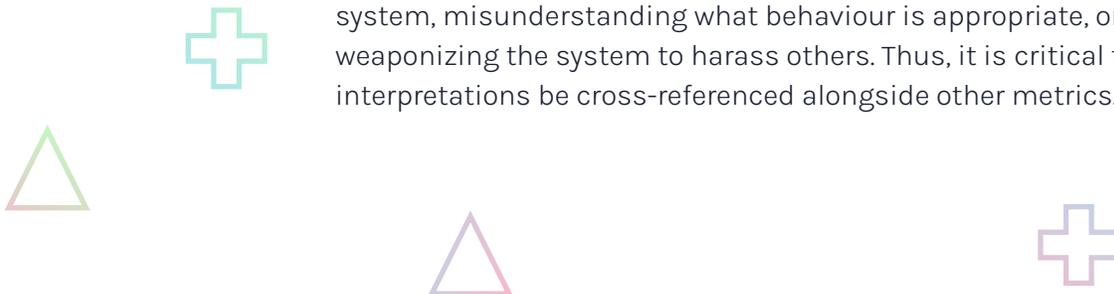
Note: Players will fit their complaints into the language you select and may choose not to report if unable to find what they need. Additionally, players may feel unwelcome or conclude that such behavior is acceptable in the absence of an appropriate reporting category.

Note: It is good practice to provide links to resources for players who may need mental health support and outline a path to assess players’ risk for self-harm. Consider the excellent work of organizations like [Take This](https://www.takethis.org), a nonprofit that promotes mental health in games: www.takethis.org.

Language matters. Reporting categories represent the language of what is deemed acceptable within a game or game-related space. Thus, the reporting categories you choose and their descriptions serve a crucial role in supporting players and gathering information about your community’s health. When choosing language, check its consistency with the Code of Conduct and through all stages of the P&R system, internally and externally.

Triage. Some reports may be timely, such as threats to personal safety, including mental well-being. Ensure you have the means to triage your ticket queue to the best of your team’s ability, and if appropriate, reassure players that you have received their report. It is best to partner with organizations, such as regional crisis centers, that have experience supporting threats of self-harm. Studios should develop a playbook to identify players in crisis and direct them to appropriate helplines. Similarly, knowing how and when to contact law enforcement is essential.

Under- and overreporting. Reporting rates alone do little to provide information about the health or behavior of your community. These rates can be influenced by a failure to engage with the reporting system, misunderstanding what behaviour is appropriate, or players weaponizing the system to harass others. Thus, it is critical that interpretations be cross-referenced alongside other metrics.



Underreporting is when a player fails to report an incident. It can indicate a lack of trust in the system, inconvenient timing, disagreement over whether the behavior was inappropriate, an improvement in conditions or the inability to find the right button.

Overreporting is when a player reports too often. It can indicate a misunderstanding between players on game expectations, or, more seriously, a significant behavior problem within your community. Overreporting, similar to the problem of underreporting, can be caused by an interface failure, such as a button that is too easily clicked.

Do not be disheartened. Building healthy communities is a journey, and one done in partnership with players. See players' reporting patterns not only as a call to action for your team, but a way to give a voice to players and support their well-being, as well as insight into how you can improve your game experience.

Assessment

When determining your needs for assessment, there are several aspects to consider.

Automation. Automated systems for assessment and moderation typically leverage an API that accepts chat or other game systems as input. They then determine if an infraction has occurred or provide a score indicating a measure of severity. Machine learning is an excellent tool for exposing trends at scale and mapping those trends to positive or negative outcomes. On the other hand, machine learning is not suited for more nuanced cases or specific issues, such as detecting problematic terms. It may be unable to provide data in the right format (such as gameplay information) to make training a system feasible. In those cases, rule-based, human-reviewed or fully manual systems could be more effective, and typically a combination is best.

CAUTION: A system trained in the same language but for a different region may not only be less effective, but may encode potentially harmful biases. It can also fail to capture inappropriate comments that use the same spelling as innocuous ones in the other region.

Manual review. Some degree of manual review or intervention is unavoidable. Players may appeal the decisions of automated systems. Because the systems are imperfect, there will always be an expected number of errors that you will need to walk back from, and have a policy for doing so (see the discussion on tolerance for false positives and false negatives on page 11). However, fully manual systems scale poorly and are likely untenable for any audience above several thousand players.

Storage needs. If data is to be reviewed, it has to be stored somewhere. Depending on the scale of your operation, and the type of data, this can get expensive quickly. Reducing this footprint, as well as ensuring you have data sunsetting procedures and privacy measures in place (including the right to be forgotten), is crucial. Note that you may have legal or government requirements for the long-term storage of evidence of actions that you take against players, such as when banning access to a purchased digital item. If you provide feedback to actioned players that includes logs or other information, be mindful of the access needs and turnaround times (and similarly understand the turnaround times for the assessment, too).

Interpretation. Whether you have manual or automatic assessment, problems interpreting the rules and spirit of your Code of Conduct will persist. Automated systems typically require explicit codification to be trained. However, they can expose harder decisions for human review and be consistent to a fault by failing to consider any extenuating circumstances. Manual systems can better interpret the spirit of the rules, but can be inconsistent and laborious.

Questions to ask. When thinking about how to assess reported behavior, keep the following questions in mind:

- What is the nature of the conduct? How was it identified? Who did it target, and why? These queries will help you track important patterns of abuse.
- What is the severity of the harm to the involved players? The answer will help you in assessing an appropriate response.
- What is the history of the transgressor? To help you determine if you should escalate your response.
- How badly is the community harmed? What example is this setting? Understand the larger forces driving community patterns and why you see these types of behaviors in the first place.
- How confident can you be in your answer to any of the above questions? How can you increase that confidence? How do you protect against overconfidence?
- What is your tolerance for false positives? False negatives? What is the cost of being wrong? The answers will inform your systems' accuracy requirements and how you manage communication with your player base.

Penalties & Feedback

Designing effective penalty and feedback systems is worthy of a separate guide (coming soon), but here are some takeaways.

What makes a good penalty? A penalty serves as a deterrent and means for expressing that conduct was inappropriate, and that the perpetrator may face more serious consequences. Penalties reinforce rules for the broader community and support the Code of Conduct.

Deciding on what penalties to enforce can be overwhelming. What to consider:

- Express your penalties in clear, consistent terms. Providing easy to understand feedback to players is perhaps the most important and overlooked aspect of penalties. Players may neither realize their conduct is unacceptable nor have a model for better behavior. First warnings with feedback and access to resources (such as developing greater resilience) can decrease recidivism.
- Teaching players to be more collaborative and empathetic helps them become stronger contributors. An added benefit is that it decreases the likelihood they will leave the community and carry their negative attitudes elsewhere. The safety and well-being of the player base are paramount, so exercise caution when giving feedback and second chances to players versus removing them.
- Consider logistics. Will you be able to enforce the penalty? Do you have the means to create the necessary infrastructure, and is this work on your roadmap? What information will you require to apply this penalty? Is there additional training for the company you will need to provide?
- Avoid excessive punishment. Do you have enough variance among your penalties, or is banning the only hammer? If so, you may lump together more minor offenses with serious infractions in a way that seems unfair to players and reduce the credibility of your system.

Penalties reinforce rules for the broader community and support the Code of Conduct.

A Note on Permanent Bans: AKA “The Ban Hammer”

The average player does not aim to ruin the playing experience, but is a product of the gaming environment. Banning a player reduces their attachment to the game or sense of responsibility for their actions. It leads players to create new accounts on free-to-play titles, removing a player’s feeling of ownership because they do not have a consistent account or identity. Thus, a player no longer feels the need to protect their account or worry about social consequences. Consider a lighter penalty with feedback, and explore why these behavioral patterns emerge.

If banning is still the right choice, determine if this is a permanent ban. Plan the logistics of upholding these bans and how to monitor them. Decide if you will need to enforce an IP or machine-ID ban for serial offenders and document this carefully.

Questions to answer:

- Will future staff have the appropriate context?
- Is a banned account deactivated or destroyed? Can a player get an account back? Are there any conditions under which you would consider revoking a ban?
- Are you able to walk back from a mistake?
- Will you have a policy for future games?
- Will the username eventually be released, and under what circumstances?

Metrics & Measurement

As your studio develops metrics, keep the following in mind. Also, review the section on Metrics and Assessment: Getting to a Methodology in the Disruption and Harms in Online Gaming Framework.

Plan ahead. Make sure your tools and systems are designed to allow your studio to measure the metrics you want. Work with your design and development team to guarantee that a comprehensive measurement plan is in place as early as possible. When you know what you want the game to look like (see Assessing the Behavior Landscape), you can concentrate on setting milestones toward achieving your goals. If you encounter blockers, such as tech limitations, planning ahead will give you time to find alternatives.

Diagnostics. Ensure that you have sufficient measures to understand that your system is working as intended; otherwise, it will be tough to assess the efficacy of your interventions. You will want to know if your false positives (applying or escalating penalties inappropriately) or false negatives (failing to apply or escalate a penalty) match your system operating expectations.

Efficacy and outcomes. A system's usefulness will depend on what you hope to accomplish for the community (What change are you trying to bring?) and understanding behavioral trends. You will want to review your false positives and false negatives to determine if they meet your expectations, and are within acceptable tolerance levels.

Note: A game's ecosystem requires continuous monitoring. Language evolves, governments change, and world events can spill over into games. One company caught its systems banning 200% above average when it did not detect a language shift in time.

False negatives. They create a perception of inconsistency and unfairness and permit unhealthy patterns to propagate and harden.

False positives. They affect player well-being and trust, as players are wrongly penalized; they teach players that the rules are inconsistent or do not matter, destabilizing a community and worsening behavior.



Reporting trends. Get a feel for reporting trends per region—who tends to report, when, and why? What is the typical report density throughout the week? You will see spikes corresponding with concurrent users (CCU), however, you may see peaks based on who is playing when, such as when kids are out of school.



Behavioral trends. Understand what types of behavior you see per region, how they change over time and whether your measures need improvement. P&R systems allow you to see the kinds of behavior that worries players and determine if there is a mismatch between your goals for the community and what players report. Regional conditions can change rapidly, set off by world or local events, a contentious company call or changes to the game itself. A daily review of key telemetry, such as report rates, penalties issued or support tickets, to monitor outliers with a weekly or bimonthly review of overall trends across all metrics is good practice.

P&R systems allow you to see the kinds of behavior that worries players and determine if there is a mismatch between your goals for the community and what players report.



 fairplayalliance.org/

 info@fairplayalliance.org

 [@fairplaya](https://twitter.com/fairplaya)



 adl.org

 [Anti-Defamation League](https://www.facebook.com/ADL)

 [@ADL](https://twitter.com/ADL)

 [@adl_national](https://www.instagram.com/adl_national)

