Disruption and Harms in Online Gaming Resource: Penalty and Reporting Systems

Sections

1. Planning a Penalty and Reporting System
2. Building a Penalty and Reporting System

PLANNING A PENALTY AND REPORTING SYSTEM

Penalty-and-reporting (P&R) systems are complex and require careful planning in their design, implementation, launch, and support as a live service. Knowing what to look out for ensures there are no surprises and helps a studio make informed decisions on introducing a new P&R system for existing titles.

Unlike many aspects of making games, P&R systems require a higher degree of due diligence. Failure to do so compromises the efficacy of your endeavors and the health of your player base. Even something as straightforward as a ban for severe misconduct carries many consequences, in some cases increasing the risk that a banned player will create a new account and harass new, vulnerable players.

This guide reviews what to think about when creating your planning and reporting system.
Conduct through design and careful consideration of how you nudge players to behave. We will cover both aspects in greater detail in future resources.

**Introduction to Penalty and Reporting Systems**

P&R systems help players understand expectations of conduct, communicate values beyond the studio, and establish trust with players.

P&R systems have three components:

**Policy.** Rules of conduct that define what you expect from players. Your policies should set guidelines for how the staff manages the rules and how they behave in-game.

**Procedure.** Formal courses of action to manage policy infractions, launch new behavior-related features, and communicate about transgressions to those affected. These procedures must be developed amid regional considerations, such as legal and government expectations, and informed by social norms that are culturally sensitive and developmentally appropriate.

**Product.** The features and systems you integrate to enforce policy and support studio procedures. They comprise reporting systems, tools to assess conduct and enact penalties, metrics for measurement and tracking, and systems for providing feedback to players.

*CAUTION: In the absence of a P&R system, players may attempt to take matters into their own hands, reaching out directly to channels that are ill-prepared to support them. If you do not have the means to assess their reports, players harmed in your game will have no recourse, and you will not be able to support them.*
Product Planning

Studios should consider the following process to align on a P&R plan no later than pre-production.

CAUTION: A common misunderstanding among developers is that they can address systems or policies later, once the game is live or after it is successful. By then, the problem is too large, and the dominant patterns, such as game culture, are already entrenched, so changing community behavior will be challenging. If you don't take steps early, the decisions around what is acceptable will be out of your control.

Step 1: Assess your game. Read the FPA’s guide to assessing a game's behavior landscape. Take time to consider what you want to prevent—do you want to thwart individuals or bots from creating accounts? If so, require some level of verification, such as a verified email address, before players can activate an account.

Step 2: Generate a player code of conduct. A Code of Conduct defines what is acceptable behavior and what good citizenry looks like in your play space. It captures the spirit of intended play. Establishing a robust Code of Conduct is not the final step before launch, but a foundational element in designing games with healthier environments. It's important to think of your company vision and values, behavior landscape, and game analysis to prioritize the types of behavior named in your Code of Conduct. The more specific the Code of Conduct, the more effective it is.

Frame the Code of Conduct using the language of the game and its players to make it relatable. It should be clear and accessible, but it doesn't need to be dry—the goal is for players to abide by the Code to the best of their ability. Include examples. Consider ways to support it throughout the game experience, from when players first learn the game to those who become your biggest advocates. Look for ways to lift up the examples of great conduct in-game and beyond.
Note: Plan to write several drafts of your Code of Conduct, first aligning on the core values and then refining the language. Treat your Code of Conduct like any feature—be thoughtful about presentation and accessibility so it will resonate with players.

**Step 3:** List problematic behaviors. Using the Code of Conduct as a foundation, spend time to identify the types of disruptive or potentially disruptive behaviors you may see (or are seeing in the case of a live game). Rank these by severity. It's not likely you'll address all issues, so it is essential to prioritize.

Note: Sometimes, you cannot address conduct using penalties but through other means. Systems for encouragement, peer review, or instilling a sense of shared responsibility so players speak out against those not upholding the game’s spirit can work, too.

**Step 4:** Determine feasibility of detecting disruptive behavior. Before you take action against any behavior, first confirm it happened. For each disruptive behavior you list from step 3, assess how feasible it will be to detect it during launch. Discuss whether you will build any assessment systems or integrate third-party tools, implement human review, or combine both. Also, factor in any logistical, regional or legal requirements, such as compliance with the General Data Protection Regulation (GDPR) compliance or the cost of storage if you need to retain evidence for a report of problematic conduct. In the case of a new game, this step should be completed no later than pre-production to prioritize any supporting work.

Note: There are several excellent third-party assessment and moderation tools available. If you plan on using such a service, be sure to engage early and ensure the cost and features are appropriate for your needs.

**Step 5:** Craft a thorough launch plan. Whether you do an alpha launch or a full launch, how you introduce your expectations and values will determine success. Communicate clearly about your values and expectations from the moment you
are in front of players and have the means to uphold them when necessary. Work closely with player support to handle incoming tickets. Be aware that your launch's scale impacts your ability to accommodate behavior issues. Without any guidance, the average gaming community will default to disruptive behavior and dominant norms.

**Note:** *Do not fall into the trap of thinking you can deal with behavior later or that early audiences will somehow behave better. This is rarely true and if you’re wrong, it can set a dangerous precedent that snowballs before you can regain control.*

**Studio-Operating Procedures**
On top of coming up with your policy, procedures, and any features or systems, you will also need to have strategies to support your staff and players. This section offers advice on keeping your studio running smoothly.

**Player Engagement**
Unlike other aspects of player support, behavior-related tickets can be challenging to address. Players may disagree with how you assess their behavior and feel that you are not doing enough as a company. They may ask questions that are hard to answer. When high-profile cases emerge, make statements that reflect your policy and values, not about the player, to set healthier norms.

The following are some action items to set up players and your player-support team for success.

**Empower the Team**
Studios juggle creating a rapport with players, empowering their teams and operating consistently across their player bases. Unfortunately, there's no formula for what the right balance looks like, but you should regularly calibrate as a team.
Extensive one-on-one support is tough to scale effectively; you spend too much time on individual players versus helping the larger player population. You run the risk of fatiguing your front-line support team. Suppose you only provide a rigid response form. In that case, players will likely feel that you are not listening, and you miss an opportunity to learn more about their experiences. Your team may conclude they lack agency to help players. The right balance is a judgment call and worth revisiting as player and studio patterns develop.

It helps to establish protocols around the Code of Conduct to equip those engaging directly with players. Consider the following:

- Easy access to the player Code of Conduct
- Access to necessary tools and services, such as changing a player's name or temporarily suspending an account pending investigation
- A playbook for handling certain types of tickets and a path to get a second opinion when the next steps are unclear, or when the tools above are absent.

**Create a Playbook**

A playbook can bring consistency to the support process while empowering your staff to help players. Any such process should be built on the values and goals that you have as a company, and echo the spirit of your Code of Conduct.

Defining steps in the path to escalation. Have a concrete process for dealing with potentially criminal behavior or threats of self-harm. What are the hotlines for the regions in which you operate? Can staff get help quickly when they need it? Is there a senior leader to whom they can reach out for guidance in an emergency? What is the process for contacting law enforcement?

Handling ambiguity. Decisions made in ambiguous situations become rules. Therefore, it is important to review emergency cases carefully. Be sure to communicate about your decisions and codify them.
**Don’t Underestimate the Toll**

Scale. If your game attracts a large audience, there will be a spike in the number of behavior-related tickets. Do you feel equipped to handle the increased workload given the complexity of handling behavior-related issues? Passing judgment on behavior is more taxing than almost any other support request yet lumped together as part of day-to-day support.

Managing emotional fatigue and trauma. The constant exposure to negativity when dealing with behavior complaints and appeals can take its toll on your staff. Staff might see disturbing content or have troubling interactions with players. Even a single occurrence can be damaging. Do staff have appropriate access to counseling? Are you able to rotate roles to avoid prolonged exposure? Do you support paid time off? Are you aware of signs of desensitization?

**Consistency**

Responsibility and accountability. Assign staff accountable for penalty systems. Be clear on who can issue, modify, or revoke penalties. Changes made to systems or policies relating to penalties must go through appropriate channels to ensure that penalties are applied consistently, fairly, and in compliance with all regional and legal requirements. Avoid the confusion of many small decisions made independently by staff by keeping the necessary people informed and implementing a cohesive, systematic approach to changes.

Remove inconsistency. Enforce your Code of Conduct or Terms of Service, regardless of a player’s status. Keep clear track of penalties issued or modified and why.

Note: Careful documentation may be legally required in some regions, especially those with restrictions on paid content. Be aware that team members with good intentions might wish to apply penalties as they encounter disruptive behavior. We recommend against this because it can cause inconsistency and reflect a power imbalance in favor of the developer. Instead, have effective reporting
systems in place and prompt players to report the incident directly. If a player approaches you to penalize another player, encourage them to use the reporting system to ensure that any complaint goes through proper channels.

**Legal and Regional Concerns**

Take explicit steps to understand the rules and customs in each region you operate. The country hosting your servers determines your legal obligations. Note that any regional customs might generate different expectations of acceptable behavior. For instance, in a country where shared ownership is the norm among family members, you may find enforcing a single user per account challenging. Understanding what is important to players in a region, and why you want to enforce your rules, can help meet everyone's needs. Be humble and accommodating when learning about new cultures and be aware of mismatches around values that prove difficult to resolve (a robust Code of Conduct helps add clarity).

**Other Things to Note**

It's not easy. We're all in this together and no one expects perfection. Many of the situations you will encounter do not have a right or wrong solution. To the best of your ability, share your values and goals, stay humble, and commit to working with players to follow your rules.

Manage expectations carefully. Zero tolerance is not the same as 100% enforcement; finding and addressing behavior is hard. Not everything fits neatly into categories of “good” or “bad.” Be prepared to face ambiguity, and manage audience expectations. Don't be afraid to call shots, but do avoid problems that you cannot realistically solve. Remember, you're laying down the foundations to support players for as long as the game is live.

Convey positivity. If you take a negative tone with players, you risk the community seeing and expecting more transgressive conduct. Model good behavior and reaffirm your expectations and values. Celebrate examples of healthy behavior,
and do not highlight problematic players or influencers who contradict your values no matter their popularity. Help players understand there is a path to improvement.

**Note:** Don’t be afraid to admit mistakes as long as they are presented with an authentic, actionable plan for improvement. Always be consistent with your Code of Conduct.

**Where to Learn More**

If you're ready to build a P&R system, consider checking out [Building a Penalty and Reporting System](https://fairplayalliance.org/resources/).

Please visit our website for more resources: [https://fairplayalliance.org/resources/](https://fairplayalliance.org/resources/)

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

1. [https://gdpr-info.eu/](https://gdpr-info.eu/)

**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.

**Introduction & 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.

![Placeholder diagram. Penalty/reporting lifecycle](image)

Upon receiving a report, studios must investigate what transpired and apply the appropriate penalty, if warranted. There must 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 must 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 that identifies progressively severe
restrictions that include removing the player.

**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.

**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 likely 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, vs. 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 where and when they are 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.

**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.

*Note:* Players will fit their complaints into that language and 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.

**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 organisations, such as regional crisis centres, 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.

*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 organisations like Take This, a nonprofit that promotes mental health in games: [https://www.takethis.org](https://www.takethis.org).
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 behavior is appropriate, or players weaponizing the system to harass others. Thus, it is critical that interpretations need to 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.

Don’t 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 above). 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:

Questions to ask. More generally, when thinking about how to assess reported behavior, keep the following items 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 and 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 the perpetrator faces serious consequences. These penalties reinforce rules to the broader community and support for 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 player 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.

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.

**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.

**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.

**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, destabilising 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’ll 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.

Where to Learn More
Please visit our resource hub for more resources:
https://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.