Combating toxicity in multiplayer gameplay with behavioral science

Contents

The rise of toxic gaming 1

Laying the foundation 3

Shifting the mindset 4

Building a better community with behavioral science 6

Conclusion 8

Authors 8

Endnotes 9

The rise of toxic gaming

A season of change in the gaming industry

On March 11, 2020, after more than 118,000 global cases, the World Health Organization declared COVID-19 a global pandemic.[1] Government lockdowns restricted the movement of people and goods to slow the spread, forcing populations to remain largely indoors. With these restrictions, many industries floundered. However, not all struggled to find purchase. Valued at $159.3 billion in 2020 and $180.3 billion in 2021, the gaming industry experienced sustained—and even accelerated—growth.[2]

A thriving gaming ecosystem

As social distancing practices took a toll on the general population, many turned to games as a social outlet. Veteran and rookie gamers alike found opportunities for engagement and social connection through chat features and online play capabilities. In the United States and United Kingdom, more than 50% of gamers surveyed agreed video games helped them stay connected to others.[3] And that connection came in many forms. Of US adults ages 18–45 who played online multiplayer games, 86% reported being able to help other players, and 83% reported making new friends and finding a sense of belonging.

Video games offered an outlet for self-expression, immersion, connection, and relaxation. And amid the mental and emotional stress of lockdowns and social distancing requirements, the pandemic seemed to underscore the value of socialization in digital worlds.

Growing virtual dangers

But virtual worlds can also attract the downside of physical life. As the gaming industry grows, so, too, can instances of bullying and harassment. As of February 2024, 83 million of the 110 million US online multiplayer gamers had experienced hate and/or harassment in their online games in the past six months.[4] Throughout 2023, 76% of adults experienced harassment in their online multiplayer games. Of these adult gamers, 67% were called offensive names, 47% reported discrimination, 38% reported being physically threatened, and an astonishing 29% reported that online video game harassment evolved into stalking.[5]

Toxic gaming also affects children. Heading into 2024, 74% of young gamers between the ages of 10 and 17 reported instances of in-game harassment in the form of trolling, being called offensive names, being bullied across multiple gaming sessions, being personally embarrassed or discriminated against, or being excluded from chats.[6]

Virtual spaces can also become overtly toxic to specific social groups, with women, racially and ethnically diverse groups, and other marginalized groups often being targeted. For example, in 2023, 48% of all US female gamers, 26% of LGBTQIA+ players, and 50% of African American players experienced harassment based on gender and/or sexual orientation, with 21% of disabled gamers reporting of being targeted because of their disability.[7]

Beyond the emotional and psychological damage such experiences can have on players, toxic gameplay can tarnish titles, provoke regulators, and threaten revenues.

Owning the problem

But where does the responsibility for gamer behavior truly lie? Some argue that the onus for controlling toxic gameplay lies with the game creator, while others believe governments have a role to play: 59% of adult gamers surveyed expressed a desire that some degree of legislation be passed to increase transparency around how game publishers and developers are handling instances of in-game toxicity.

In response to this push, many publishers have already taken steps to address toxic gameplay. For example, we have seen publishers institute features allowing players to set filters on in-game messaging or auto-mute features to tamp down on toxic behavior. One major publisher, for example, has custom filters that players may select to limit or flag the type of communications received. Examples of filter levels include Friendly, Medium, Mature, and Unfiltered.[8] Another publisher has also implemented its own tactics to combat toxicity in its first-person shooter (FPS) title through auto-muting features for players who are repeat offenders for explicit in-game voice commentary. In addition, they implemented an AI system that reviews in-game chats and distributes warnings or bans immediately after a message is sent. During its first 45 days of activity, this system recorded more than 20,000 bans in the title.[9] Some developers have also become more transparent in communicating the outcomes of combating disruptive behavior online, with one major FPS releasing anti-toxicity reports[10] to the public. These outline their efforts to enforce positive gameplay through in-game filters to catch offensive usernames, filters on in-game messaging that bans users for inappropriate messages and even clan tags with toxic meaning behind them. In some instances, anti-toxicity reports have showcased a developer banning more than 350,000 accounts for racist or toxic behavior.[11] While efforts like these are important, they may only breach the tip of the toxic iceberg.

Envisioning a solution

Today, the approach to addressing toxic gameplay is largely reactive and based on punishment.[12] Players who violate community guidelines, harass others, or otherwise participate in poor social conduct often receive temporary restrictions. These methods of punishment are critical to removing a toxic player from the gaming ecosystem immediately, and in the short term, they can be effective. However, these are often brief, with full gameplay capabilities restored after a defined period and only the most grievous of behaviors resulting in permanent bans. This likely does little to create long-term behavioral shifts as players either wait until their accounts are restored or learn how to find alternatives (like creating new accounts once banned). The ability to hide from punishment behind crafty gameplay and account management may cause the punishment model to lack "stickiness" as behavioral habits may not change, and players could default back to their toxic behaviors.

To help combat toxic gaming, long-term behavioral shifts should occur via new habit formation. To help achieve this, any potential solution should consider supplementing the current punishment schema of today with behavioral science to craft a system that considers prioritizing three key tenets:

1. Habitual journeys over one-time interactions
2. Proactivity over reactivity
3. Positive reinforcement over punishment

In other words, the gaming community should shift the behavioral norm via a system that nudges player—morally and ethically—in the right direction prior to the exhibition of any toxic behavior while rewarding them for doing so, encouraging the formation of new, positive behaviors and habits that "stick."

Laying the foundation

Understanding the issue

For a system to detect toxic behavior, data on toxic behavior should be gathered to train the system algorithm. This may require:

- A common definition of toxic gameplay
- A reliable method for identifying toxic behavior

Defining toxicity as disruption

While seemingly straightforward, coming to a common definition of "toxic gameplay" can be challenging. This is because "toxic" is an ambiguous term that represents a nuanced social concept. What is considered "toxic" by one may not be considered toxic by another. Social factors like culture, language, and situational context can change the meaning of the words we say and the things we do, potentially altering the perception of toxicity at both the peer group level and individual level.

The Fair Play Alliance (FPA)[13] has acknowledged this challenge and has determined that toxic gameplay may be better stated as "disruptive behaviors." Unlike the term "toxic," "disruptive behaviors" can be concretely defined as "anything deemed unacceptable by a player of the game company." Replacing the nuanced concept of toxicity can allow game makers to more reliably identify instances of behavior that may, in fact, be toxic.

Identifying disruptive behaviors

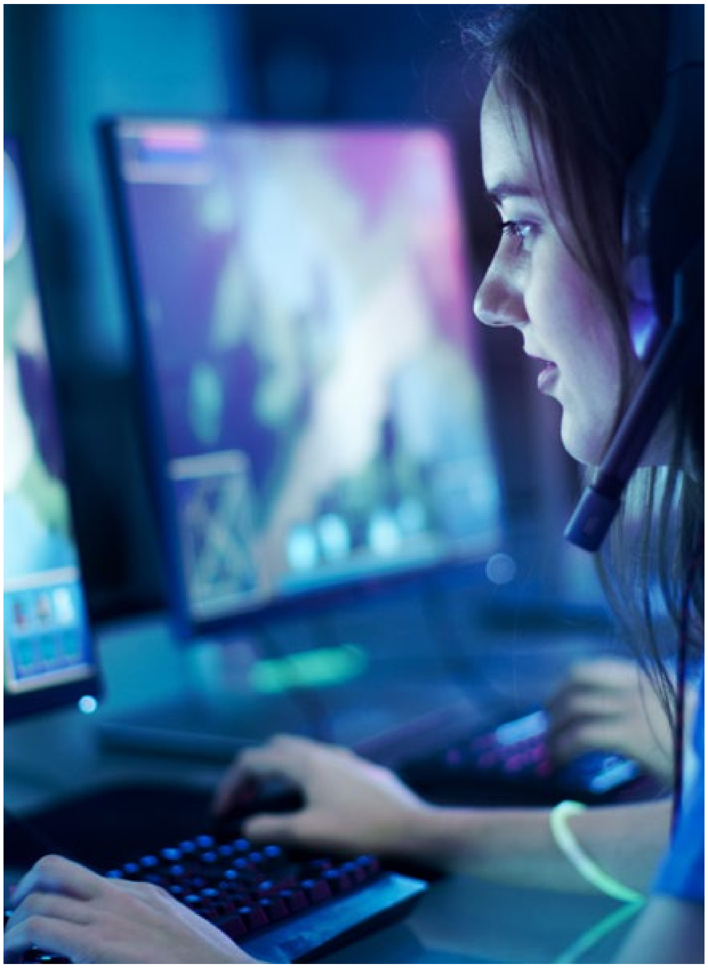
To help determine if a behavior qualifies as toxic, the FPA developed a four-element framework to assess disruption.[14]

Expression addresses the form in which the disruption took place. For example, was the behavior unintended? If so, it would not be considered toxic. However, if the behavior manifests in the form of cheating, intimidation, etc., then this would be considered toxic.

Delivery channel is about where the behavior occurred (game lobby, console chat, etc.). These disruptive behaviors can leak into game-adjacent settings, like social media as well.

Impact seeks to understand who was harmed and the radius of the impact of that harm.

Root cause seeks to understand why. This is a complex behavioral assessment, as there are many factors that can contribute to why and how toxicity becomes normalized.



Positioning for long-term success

Gathering reliable and unambiguous data can help publishers to gain a clearer picture of the toxic gameplay taking place. From there, it can become possible to enhance behavior-detection algorithms responsible for generating the insights on which codes of conduct and gameplay mechanics are based.

However, to truly be successful in the long term, games and policies should be designed and written with the intention not only to punish but also to shift the social norm through (1) new habits, (2) proactivity, and (3) positive reinforcement.

Shifting the mindset

Losing out on the long term

Today, the consequences for toxic gameplay may simply be too short-lived to break bad habits. When accounts are restored and matchmaking freezes end, players often return to their harmful play style as old habits take over. Breaking these bad habits is a first step in encouraging long-term behavioral shifts.

The importance of habits

Habits are defined as being born from smaller behaviors repeated over time and strengthened as they are rewarded.[15] These habits can be the building blocks for entire behaviors that become ingrained in our procedural memories. Sometimes we intend to collect our habits; occasionally we do not. Regardless of intent, however, these habits often dictate our routines. For toxic gamers, acting in a habitually disruptive manner may be a default state. These disruptive habits should be broken and positive habits built in their place.

But breaking habits isn't always easy. Quite often it isn't the new behavior that's most difficult to adhere to—it's avoiding the old behaviors. Studies show that, on average, it can take around 66 days to break a habit.[16]

Ultimately, building habits that "stick" could require both proactive behavioral sensing to help players avoid "indulging" their bad habits, as well as strategically written and enforced codes of conduct rooted in positive reinforcement that rewards the player for demonstrating the desired behavior.

Breaking bad habits through proactivity

A first step in breaking a bad habit is to avoid partaking in the habit itself. Assisting someone to avoid indulging their bad habit is known as pre-correction.[17] Widely cited in educational curriculums and parenting styles, this practice prompts good behavior and entails giving consistent reminders of the behavioral expectations prior to the activity under which there is a history, or high risk, of failure.[18] For example, a teacher may ask "which side of the hall do we use to walk to lunch?" prior to the bell to avoid a failure mode wherein students run wildly through the halls.

In the gaming ecosystem, prompting appropriate behavior through pre-correction is intended to remind those at risk of failure (e.g., players who have previously demonstrated toxic gameplay) of the desired norms before their behavior can deteriorate into the realm of "disruptive." For example, dialogue boxes that pop up during matchmaking reminding players of expected behavior. Or a visual flag that appears prior to each match to remind players how many times they were reported within the past week.

If developers and publishers continue to rely on systems that wait to act until after behavior deteriorates and becomes disruptive, then pre-correction cannot take place, and players will likely continue to indulge their bad habits, potentially making them even harder to break.

The pitfalls of punishment

To break habits, proactivity via pre-correction may not be enough. The punishment-focused models of today should be supplemented with reinforcement. Specifically, positive reinforcement. It is this paradigm shift that requires a more thorough understanding of behavioral science—one born from operant conditioning theory (OCT).[19]

OCT is a human behavior and learning model that explains how people respond differently between punishment and reinforcement (also known as "reward"). Under this theory, punishment is designed to decrease a behavior, whereas reinforcement is designed to increase a behavior. In addition, both punishment and reinforcement can leverage positive or negative impetus.

A positive schema is defined by the addition of an element, whereas a negative schema is defined by the removal of an element. Therefore, there are four types of operant conditioning: positive punishment, negative punishment, positive reinforcement, and negative reinforcement.

When considering the approach to stopping toxic behavior, the initial instinct may be to resort primarily to punishment, considering the definition of punishment is to decrease a behavior from occurring. However, relying solely on punishment to try to shift the collective mindset on toxicity in gaming could be an error. While it inevitably may be necessary to continue leveraging punishments to address the more heinous social infractions, attempting to rely on a widespread punishment model alone may cause players to cease participating in online gaming completely out of fear or agitation. This "stick" approach would have the potential to put live services at risk or perhaps even cannibalize anti-harassment and bullying efforts by leading the rule-abiding players to abandon the server altogether, leaving only the toxic players who are unconcerned with the punishments they may face.

Building new habits through positive reinforcement

To avoid the pitfalls of the punishment approach, gaming companies should supplement with reinforcement. But should that reinforcement method be positive or negative?

Remember that, in this context, positive does not mean good, nor does negative mean bad. Rather, the term "positive" indicates the addition of an element, whereas "negative" indicates the removal of an element.

For example, a parent can reward their child with a new toy for behaving—positively reinforcing the development of good mannerisms. On the other hand, the toy can also be taken away if the child begins to misbehave. Set in the context of gaming, positive reinforcement models can take many forms such as in-game rewards, higher-quality matchmaking, and more.

Taking a positive reinforcement approach over a negative reinforcement approach has been demonstrated to create more immediate, long-lasting effects as we seek to stimulate the reward pathway in the brain.[20]

Scientifically known as the mesolimbic dopamine system (MDS),[21] the reward pathway is our biological programming to survive—survival being the ultimate reward. It is responsible for our innate urge to seek food, water, shelter, and warmth. When this pathway is stimulated, the brain floods with dopamine, which makes us feel good. And each time the reward pathway is stimulated it's strengthened, reinforcing our desire to do that activity again.

The MDS is also connected to areas of the brain that control long-term memory, and as our behavior is rewarded and the MDS is activated in response, that behavior is encoded into the long-term portion of the brain.

Some publishers have already started to identify this critical difference between reward and punishment. One developer of a highly popular multiplayer online battle arena (MOBA), for example, has acknowledged that only 5% of the MOBA's population are consistently disruptive, with 86% registering as occasional offenders who lash out infrequently during bad games.[22] As a result, the developer's behavioral systems team has decided to shift away from focusing on punishing the 5%, and instead focus on rewarding the rest of the normally well-behaved players. For example, players who reach a certain level in its rewards system will be rewarded with a free skin in celebration.

To help combat toxic gaming in a manner that encourages long-term and eager behavioral modifications, the industry should consider shifting from a punishment-focused model to a positive reinforcement model.

Building a better community with behavioral science

Laying out the infrastructure

Considering the global scale of the gaming industry and the social interactions therein, adapting game mechanics and codes of conduct to adhere to the tenets of proactivity and positive reinforcement may not be easy and could rely upon viewing the act of breaking habits and changing behavior as a science.

Using the COM-B model

The COM-B model, developed in 2011 by Susan Michie, Maartje van Stralen, and Robert West, is a framework to consider when thinking about how to design and implement new approaches to anti-toxic efforts and codes of conduct.

The model takes into consideration that human behavior is complex and comprises many influencing factors. However, this model has simplified these factors to a manageable degree. It has defined the long-term capacity for behavior change (B) as a factor of an individual's capability to change (C), opportunity to change (O), and motivation to change (M).[23] In other words, a change in behavior (B) = C + O + M. The criteria of all three pieces—capability, opportunity, and motivation—should be present and met before human behavior can change and persist.

Capability

Within the COM-B model, a person's *capability* refers to their individual psychological and physical ability to participate in the change that is taking place. This can be a challenging component of the model as, within the context of gaming, we are asking the question: "Does this player understand what it means to be positive in the way we want them to be?"

When viewed through a gaming lens, this is where enhanced data collection becomes important. Capabilities such as natural language processing (NLP) and machine learning (ML) combined with other data science strategies can support an algorithm in detecting patterns in an individual's behavior. In turn, these patterns can inform publishers on how to better design games and equip players with the capability they need to truly understand what it means to be positive.

Opportunity

Opportunity refers to the external factors that enable a behavior, making it possible. In other words, is there a chance to demonstrate positive behavior within the game itself? For example, in some online games there are opportunities built into the game mechanics that enable players to demonstrate positivity and receive rewards in return. This has been done using mechanics like positive emotes (e.g., salutes, handshakes, hugs, dances) and pre-scripted quick chats (e.g., "Good game!"). In these games, there are usually systems in place that count these positive interactions and assign a "score" based on how many positive interactions a player has. Often called "karma" balances, "reputation" levels, or "honor" systems, there are usually opportunities for the system to reduce point values if a player exhibits negative play styles or toxic behaviors. In this way, the system's score is tied directly to the player's behavior. It is not only the opportunity to demonstrate good behavior that could matter

but also removing opportunities to demonstrate disruptive behavior. This can be done in a myriad of ways, such as moving text that skews negative to the bottom of a chat wheel, forcing players to have to scroll further to send angry messages to opponents. Ultimately, the social moments within the games should technically be built in a way that encourages players to seize easy opportunities to be kind to others and rewards them for doing so while, at the same time, limiting the opportunities to demonstrate disruptive behaviors.

Motivation

Motivation refers to both the conscious and unconscious psychological processes that inspire behaviors. This is due to motivation addressing the concept of "want" and turning certain behaviors from something they need to do, to something they want to do.

It is also important to note that not everyone will react the same way to behavioral interventions. In other words, not everyone is motivated to change their behavior in the same way.

A study found that behavioral interventions can be designed around identified phenotypes that have distinct behavioral traits.[24] Targeting behavioral interventions tailored to these phenotypes are seen to produce a response and yield better long-term results.

But current efforts to combat toxicity in gaming typically fail to consider the array of phenotypes that may be present in the gaming ecosystem, instead offering a "one-size-fits-all" approach to punishing toxic players. Gamers often are not endowed with sufficient, tailored motivation to turn the "need" to be kind into a "want" to be kind, feeding back into the concepts of proactivity (or helping players to avoid indulging in disruptive habits) and reinforcement (closing the reward feedback loop in the brain, encouraging repeated positive behaviors).

Pulling it all together

Ultimately, creating long-lasting changes in human behavior may not be easy. But finding ways to encourage behavior that "sticks" is likely important to enabling the spread of positivity through a population. Fleeting moments of kindness may not be enough to shift the norm and expectation of the gaming ecosystem to demand positivity.

However, by working to ensure that players can understand expectations, demonstrate desired behavior, and are provided sufficient motivation for retaining it, good habits could slowly and surely begin to shift the social norms.

Conclusion

Ultimately, humans tend to be creatures of habit. Many of us have them. Some good, some bad. Some we've developed consciously, and others we've developed in ways unknown to us. Regardless of how we've obtained our habits, they dictate our decision-making and behavior.

In online gaming today, approximately 75% of US players report harassment, bullying, and other toxic behaviors perpetuated by other players.[25] These disruptive behaviors often create uncomfortable spaces. To help tackle the issue of toxic gameplay, those with disruptive habits should be encouraged to change those habits. This could be done through (1) proactive sensing and (2) positive reinforcement.

Using proactive sensing and intervention, game developers and publishers can help habitually disruptive players avoid opportunities to indulge in toxic gameplay habits until new, healthy habits are formed in their place.

But like a disease, even breakthrough moments of bad behavior can infect a healthy population if left unchecked. That is why frequent, positive reinforcement of desired behaviors can be critical to seeing sustainable shifts to the social norm.

But redesigning codes of conduct and designing technical aspects of games to support proactive sensing and positive reinforcement is a large undertaking and begs the question: "What should we consider when making these changes?" One answer? Behavioral science. Using the COM-B model, publishers and developers can gain a better sense of what it truly takes to shift not only habits but also personal norms. By using tools like NLP and ML, by designing the technological moments in games to make opportunities to be kind easy and accessible, and by offering sufficient motivation to retain the changed behavior over time, the gaming industry could see massive strides in the fight against toxic gameplay.

Authors

Greg Szwartz
Managing Director
Deloitte Consulting LLP
gszwartz@deloitte.com

Richard Goldsmith
Senior Manager
Deloitte Consulting LLP
rgoldsmith@deloitte.com

Vincent Attonito
Manager
Deloitte Consulting LLP
vattonito@deloitte.com

Abigail Miller
Senior Consultant
Deloitte Consulting LLP
abmiller@deloitte.com

Endnotes

1. US Centers for Disease Control and Prevention (CDC), CDC Museum COVID-19 Timeline page, timeline posting dated March 11, 2020, accessed October 2024.
2. Andres Lahiguera, "Media & Entertainment: Video Games sector," International Trade Administration, January 25, 2021.
3. Kevin Westcott et al., "2022 Digital Media Trends, 16th Edition: Toward the Metaverse," *Deloitte Insights,* 2022.
4. Anti-Defamation League (ADL), "*Hate is no game: Hate and harassment in online games 2023*," February 6, 2024.
5. Ibid.
6. Ibid..
7. Ibid.
8. Kurt Perry, "Valve's new CS:Go auto-mute system to tackle toxic players," *PC Invasion*, February 7, 2024.
9. Ibid.
10. Nathan Grayson, "*Call of Duty* has banned over 350,000 players for racism and toxicity in the past year," *Kotaku,*, May 26, 2021.
11. Ibid.
12. Riot Games Behavioral Systems Team, "Behavioral Systems: April 2021," for *League of Leagues*, Riot Games, April 22, 2021.
13. Thriving in Games Group (TIGG, previously the Fair Play Alliance), TIGG , homepage, accessed October 2024.
14. Fair Play Alliance and the ADL's Center for Technology and Society, Disruption and Harms in Gaming Framework, December 2020.
15. Wendy Wood and Dennis Rünger, "Psychology of habit", *Annual Review of Psychology 67*, (2016): pp. 289–314.
16. Phillippa Lally et al., How are habits formed: Modelling habit formation in the real world,"*European Journal of Social Psychology* 40, no. 6 (2010): pp. 998-1009.
17. Terry Jackson, "Intervention Guide: Precorrection," ibestt Project, University of Washington, 2017.
18. University of Louisville, "*Pre-correction/Prompting – Behavior*," accessed October 2024.
19. Matt McMillen, "*Operant conditioning: What it is and how it works*," WebMD, last updated December 27, 2023.
20. Kelly J. Bouxsein, Henry S. Roane, and Tara Harper, "Evaluating the separate and combined effects of positive and negative reinforcement on task compliance," *Journal of Applied Behavior Analysis* 44, no. 1 (2011): pp. 175–79.
21. ScienceDirect, "Mesolimbic pathway," accessed October 2024.
22. Sabrina Ahn, "Behavioral system changes in *League of Legend*," *EarlyGame*, March 12, 2024.
23. Dan Pilat and Sekoul Krastev, "The COM-B model for behavior change," The Decision Lab, accessed October 2024.
24. Xisui Shirley Chen et al., "Association between behavioral phenotypes and response to a physical activity intervention using gamification and social incentives: Secondary analysis of the STEP UP randomized clinical trial," *PLoS ONE* 15, no. 10 (2020): e0239288.
25. ADL, "*Hate is no game: Hate and harassment in online games 2023*," February 6, 2024.

This article contains general information only and Deloitte is not, by means of this article, rendering accounting, business, financial, investment, legal, tax, or other professional advice or services. This article is not a substitute for such professional advice or services, nor should it be used as a basis for any decision or action that may affect your business. Before making any decision or taking any action that may affect your business, you should consult a qualified professional advisor.

Deloitte shall not be responsible for any loss sustained by any person who relies on this article.

About Deloitte

Deloitte refers to one or more of Deloitte Touche Tohmatsu Limited, a UK private company limited by guarantee ("DTTL"), its network of member firms, and their related entities. DTTL and each of its member firms are legally separate and independent entities. DTTL (also referred to as "Deloitte Global") does not provide services to clients. In the United States, Deloitte refers to one or more of the US member firms of DTTL, their related entities that operate using the "Deloitte" name in the United States and their respective affiliates. Certain services may not be available to attest clients under the rules and regulations of public accounting Please see www.deloitte.com/about to learn more about our global network of member firms.

8191054