

# Trust, Play, and Platforms: Sharing Lessons for Safer Digital Spaces

---

SAMANTHA BRADSHAW AND DEAN JACKSON



 NYU | STERN

Center for Business  
and Human Rights

May 2026

# Table of Contents

1. Why Trust & Safety in games is serious business for digital rights .....	1
2. Governing online games and social media platforms .....	2
3. Governance as part of the design process .....	5
4. Trust & Safety governance in transition .....	7
5. Conclusion and lessons learned .....	11
6. Considering digital lives holistically .....	13

## Authors

Samantha Bradshaw is the director of the Center for Security, Innovation & New Technology (CSINT) at American University (AU) and an Assistant Professor at AU's School of International Service.

Dean Jackson is a non-resident fellow at American University's Center for Security, Innovation, and New Technology.

Mariana Olaizola Rosenblat, policy advisor on technology and law at the NYU Stern Center for Business and Human Rights; Nicole Gaouette, the Center's consulting editor; and Dhanaraj Thakur, Director, Emerging Technologies Initiative, Multiracial Democracy Project, George Washington University Law School, provided editorial assistance for this report.

This report was commissioned by the NYU Stern Center for Business and Human Rights in preparation for the 2025 conference on "Mainstreaming Trust & Safety in Online Gaming," of which American University's Center for Security, Innovation, & New Technology and the Center for Democracy & Technology were co-organizers.

# Executive Summary

---

This paper provides an overview of current Trust & Safety practices in online gaming, examining how these platforms govern user behavior through a mix of rules, moderation systems, and design choices. It highlights key similarities and differences between online gaming and social media platforms, including the challenges of moderating real-time interactions in games and the use of concepts such as disruptive behavior and community-driven enforcement. The paper also outlines broader trends affecting both industries, including increased reliance on third-party vendors, the growing role of generative AI, and expanding regulatory scrutiny. It concludes by identifying ongoing challenges related to transparency, coordination across platforms, and resource constraints.

## 1. Why Trust & Safety in games is serious business for digital rights

In 2004, the moderators of *Star Wars Galaxies* [had a problem](#). A glitch in the massive online multiplayer game allowed players to counterfeit the in-game currency, called “credits,” in large quantities. Because currency circulated between players in the game’s economy, many players possessed counterfeit credits even though they were not responsible for making them. Unable to tell the cheaters apart from rule-following players, the moderators took the draconian step of banning all players whose character possessed counterfeit currency. The backlash led to swift rebellion. Friends of the banned players gathered in the game’s spaceports—highly visible areas where they spammed publicly viewable protest messages, disrupting play.

The moderators responded by teleporting all the protesting players’ characters into the cold, black void of space.

The antagonism between moderators and the game’s player base grew stronger with time. After another incident, moderators summoned virtual stormtroopers to prevent protesters from re-spawning in-game. Outraged, players [left in droves](#). The game died. Trust & Safety failures killed *Star Wars Galaxies*.

Online and especially multiplayer online video games have always been spaces where trust, safety and content moderation challenges unfold. However, they are often only explored through distinct lenses, such as child safety, online radicalization, and online harassment. This is surprising given that estimates of [global gaming industry](#) revenues are larger than those for professional [sports](#), [music](#), and [film](#). Nearly 160 million American adults across age brackets and genders [play games](#) for at least an hour a week.

Video games are a mainstream cultural force, but they are not a mainstream topic among policy analysts focused on digital rights or online Trust & Safety. This may be partly because the immersive and entertaining nature of video games can obscure the governance decisions that shape them. Within these ecosystems, decisions about moderation, monetization, design aspects like algorithmic recommendation, and community management carry enormous implications for online trust and safety and human rights. But the gamified context can make these decisions less visible—and less political—than comparable debates about social media content moderation.

This strikes us as an oversight, not because games are an industry to be harshly scrutinized, but because the experiences of game professionals and players could be a valuable source of insights for governing the internet writ large. Likewise, the gaming industry is affected by larger trends in technology and regulation, but is underrepresented in relevant discussions. This should change.

Like social media, online games are digital platforms where people interact. While less likely to be a source of news or information, their cultural—and thus political—impact is [undeniable](#). It is also bidirectional: politics influence games just as games influence politics, as former US President Joe Biden’s 2020 presidential campaign demonstrated when it [established](#) an outpost in the game “Animal Crossing.” (Nintendo, the game’s publisher, later [asked](#) businesses and organizations to “refrain from bringing politics into the game.”) Gaming companies, like social media companies, grapple daily with [user-generated content](#) and the implications of [generative AI](#).

If every online game is a different platform, then they vastly outnumber social media networks—and there is significant variety between games. Each is something like its own laboratory for Trust & Safety problems and solutions. We are not the first to make this observation: to give one example, in 2023 the Atlantic Council’s Task Force for A Trustworthy Future Web [found](#) that “the gaming industry offers unique potential for insights and innovation... there are lessons to be learned from the industry’s successful and less successful approaches to content moderation, trust and safety, and product design.” Resources like the [Digital Thriving Playbook](#) show that the games industry has developed its own thought leadership, even without significant crossover from the digital rights community. Still, the two communities can and should learn from one another.

Regulators and advocacy organizations can also learn by comparing the different universes of online platforms. For instance, a recent presentation at the International Conference on Machine Learning [explored](#) how policy discussions about generative AI could be informed by past debates over social media regulation, avoiding the impulse to start from scratch. In a similar spirit, we suggest that greater consideration of the games industry by policymakers and digital rights and safety advocates would benefit all stakeholders.

This paper is intended to help digital civil society researchers and advocates bridge understanding of social media and online gaming—two spaces of enormous cultural, economic, and political significance and burgeoning regulatory interest. While not comprehensive in scope, if it is successful, this report will serve as a catalyst for productive conversations between scholars, industry practitioners, and policy analysts across both areas.

## 2. Governing online games and social media platforms

From the humble origins of coin-operated arcades to today’s massive online networks, video games have transformed into a worldwide cultural phenomenon. As Pete Etchells argues in *The Guardian*, video games are no longer niche pastimes for a small number of people, but rather [the defining entertainment media of our time](#). Like social media, online games are dynamic social spaces that connect millions of people from all around the globe. On both gaming and social media platforms, people shape their online identities, interact with communities, and participate in shared cultural experiences. These platforms are also deeply shaped by technological systems—algorithms, digital advertising, and social graph functions all influence how people interact. Common monetization models include subscriptions to surveillance-driven advertising, yet these commercial mechanisms are often masked by the playful and social elements that make the platforms so engaging.

Although there are important differences between social media and online games, the boundaries are blurring as these two spaces increasingly share attributes. Online games often rely on matchmaking algorithms to pair players with evenly matched opponents, while social media platforms use recommendation systems to curate content. Yet gaming platforms now employ recommendation algorithms as well—for example, Roblox’s ‘Recommended for You’ feature suggests experiences based on users’

past activities and preferences. Likewise, while most social media platforms monetize through digital advertising, some have adopted subscription-based models, such as Twitter (now X) or Meta's paid verification services.

As the features of online games and social media increasingly converge, so too do their models of governance. Both rely on a mix of formal rules (such as terms of service and privacy policies), algorithmic enforcement, and community-driven norms. Human moderators step in where automated systems fall short, especially in cases requiring contextual judgment. These mechanisms can also come together when players livestream gameplay on a social media platform, or where Trust & Safety teams from gaming communities and social media platforms coordinate to track cross-platform harms. However, social media platforms and online games also diverge in important ways. While a comprehensive treatment of every governance mechanism across platforms is impossible here, a high-level overview of major governance categories reveals important similarities and differences. Below, we highlight some of the most significant points of similarity and divergence as well as trends and debates concerning Trust & Safety in social media and online games.

## Rules and enforcement

Both social media and gaming platforms ground their governance frameworks in terms of service agreements and privacy policies, which function as legal contracts that set the broad parameters of platform use. Alongside these, they typically publish community guidelines or player codes of conduct that translate these rules into more accessible language for users, spelling out what kind of behaviors and content are considered acceptable. While these documents often lay out the rules for platform use, their enforcement has had a [long and complex history](#). On both social media and gaming platforms, the enforcement of these terms relies on a combination of [automated](#) and [human](#) moderation practices. This involves machine-learning classifiers and keyword filters that flag potentially harmful or prohibited content for further review. Human moderators, employed directly by the company or sometimes contracted through third-party vendors, review ambiguous or severe cases. While this approach is intended to balance scalability with nuance, it often comes with [important trade-offs](#) that have implications for how users appeal content moderation decisions, or how platforms are held accountable when decisions go poorly. For example, automated systems can misinterpret context, as seen when [Meta's automated filters repeatedly remove posts about breast cancer awareness campaigns](#). They are also much less robust when it comes to [detecting hate speech in low-resource languages](#), or could unfairly moderate against [certain communities due to data bias](#). In gaming, similar systems can wrongly penalize new players who are simply learning and losing frequently, mistaking them for players who are deliberately sabotaging the game. Additionally, hate speech and content that can promote political violence, even though they are often prohibited by terms of service and community guidelines, [can still be missed by automated systems in games](#).

## Harmful content and “toxicity” vs. disruptive behavior

The language used to describe undesired activity on a platform has changed with time, in different ways across the gaming and social media industries. Phrases like “[violative content](#)” focus on individual instances of user expression that transgress against written (and sometimes [quite detailed](#)) community guidelines. A high degree of specificity lets T&S professionals feel justified in their decisions and the legitimacy of their roles, but introduces the problem of [borderline content](#) which violates the spirit, but not letter, of a policy. “[Problematic content](#)” is a broader category that accounts for borderline content but introduces some vagueness, and neither term reflects the reality that creating and sharing content is a form of behavior—and behavioral patterns can harm Trust & Safety in ways that go beyond individual instances of content. Bullying or harassment, for instance, often entails posting negative comments or sending negative messages consistently over time.

Sometimes, these challenges are grouped together under the term “[toxicity](#).” But in the gaming industry, some professionals warn that “toxicity” is a [broad and subjective term](#). Instead, they prefer to focus on “disruptive behavior,” which one paper [defines](#) as “conduct that mars a player's experience or a

community’s well-being,” or “conduct that does not align with the norms that a player and the community have set.” The paper goes on to say that many game developers prefer this term to the oft-used “harmful content,” because games often feature forms of disruption like unsportsmanlike conduct (e.g. rage-quitting) or unexpected playstyles that do not necessarily contribute to serious mental, emotional, or physical harm. Ultimately, moderating disruptive behavior is also a subjective process: there is no escaping normative judgment when designing and governing an online platform. However, the focus on behavior and its effects on others and the platform usefully reframes the conversation away from individual acts of expression and toward the creation and enforcement of communal norms.

## The scale and speed of moderation

While the underlying mechanics of moderation—combining automated detection and human review—are similar across industries, social media and gaming companies confront distinct challenges of scale and speed. Social media platforms must moderate vast quantities of user-generated content: posts, images, and videos that can spread virally within minutes. Speed is paramount—if harmful content is not quickly identified or removed it may reach thousands or millions of people. Online games, by contrast, grapple with real-time interactions that can immediately affect player safety and the quality of play. Issues such as harassment, cheating, or exploiting mechanics often unfold in the moment, leaving little time for reactive moderation. A post on Facebook may “go viral” in hours or days, but in a multiplayer match, a single disruptive player can sour the experience for dozens of others instantly. Games also present unique moderation challenges, such as policing in-game voice chat at scale—something social media platforms face less often, with the notable exception of livestreaming services.

One way of describing these differences is “synchronicity”: do individuals experience the rules violation all at the same moment, or separately over time? Since player interaction in games is synchronous much more often than on social media, it may be that gaming companies are more likely to moderate player accounts while social media companies are more likely to take action on content. Consider for instance, that Microsoft’s transparency report for the first half of 2024 [shows](#) many more instances of enforcement against accounts, compared with Meta’s quarterly reports which are largely [focused](#) on content.

There are exceptions; for example, [livestreaming](#) is an area of synchronous activity where games and social media intersect. Livestreaming platforms have experimented with additional methods for detecting potentially violative behavior, such as streams initiated by brand-new accounts that are [watched](#) by a high number of VPN users.

Private messaging systems are another area of overlap. Such systems are commonplace in both games and social media and create spaces where harassment and abuse can occur in real time but out of public view. As platforms innovate, these challenges compound, requiring governance models that balance speed, accuracy, and fairness while preserving the social and entertainment value that draws people in the first place.

## Communities making and enforcing the rules

As online governance has evolved, platforms of all types are increasingly concerned with the legitimacy of their moderation efforts. Users who believe (rightly or wrongly) that moderators are arbitrary, unfair, or overly strict [may leave](#) for [another service](#). In some instances, [entire communities](#) may do so. The effects on a platform’s sustainability and health can vary from a sigh of relief as bad actors depart to existential risk from user churn. This challenge is compounded by the scale of today’s large platforms, which struggle to moderate communities and sub-communities in a nuanced, sensitive fashion. Scale is inversely related to nuance because automated tools are typically blunt, and it is usually impossible to put human eyes on every problem for a meaningful amount of time.

Community moderation is a common way of squaring both circles. Reddit and [Wikipedia](#) are both examples of platforms with significant community moderation; in online games, EVE Online offers

examples of community moderation through both sophisticated self-moderation within player groups and the “[Council of Stellar Management](#),” a player advocacy body elected directly by players themselves. Approaches like these merit more research: for example, in a 2025 paper, Kiene, Hwang, and Teblunthuis et al. [argue](#) that academic literature has spent insufficient time examining the design and adoption of rules on community-moderated platforms.

Reddit provides other interesting models. First, it is divided into subreddits with volunteer community moderators who set and enforce their own rules. Reddit sets a [policy floor](#) (no illegal content; no misleading impersonation of real entities; no revealing another user’s private information, etc.); subreddit moderators are free to layer additional policies on top of that floor and to enforce those policies, including through the use of third-party tools. If a community repeatedly violates Reddit’s policy expectations, the moderators offer a point of intervention. Reddit may be able to better communicate its expectations to community moderators, to provide them with support to improve outcomes, or—rarely, in the case of moderators who act in bad faith and are antagonistic to Reddit’s rules—to quarantine or ban subreddits entirely.

Reddit also maintains a reputation system through upvotes, downvotes, and “karma,” which is essentially a ranking of a user’s behavior and community contributions. In gaming, systems like League of Legends’ “Honor” system may play a similar role. In a paper for the mental-health nonprofit Take This, Elizabeth Kilmer and Rachel Kowert [suggest](#) more research is needed to evaluate the impact of such systems. A 2020 paper by Johanna Brewer, Morgan Romine, and TL Taylor for the “Designing Interactive Systems” conference [assessed](#) an intervention for video game livestreamers who receive a digital (and physical) badge confirming they adhere to a voluntary code of conduct encouraging respectful behavior. They found that this relatively simple social intervention generated significant user interest (more than 370,000 gamers participated) and showed promise in reducing disruptive behavior.

### 3. Governance as part of the design process

Platform rules and moderation take place after a service is designed (though policies are often drafted alongside the design process and updated after launch). Other governance features are even more inherent to the design process. The types of settings available to users and the incentives provided to them by the range of actions available are also part of platform governance. They have direct implications for the health, safety, and accessibility of a service.

#### User settings, middleware, and safety-by-default

The settings available to users can allow them to directly participate in the governance of their own experience on a service. Steam, for example, [allows](#) users to turn filters for profanity and slurs on or off and to add to their own personal list of filtered words (Steam says the ability to opt into slurs is intended to allow marginalized groups to reclaim those words). A platform’s default settings can be an expression of its own stance on safety and other issues; for example, in 2024, Instagram [changed](#) the default settings for users who indicated they are under 17 years of age so they received fewer notifications at night, were less likely to see graphic content, and were unable to receive messages from strangers. In its transparency report for the first half of 2024, Microsoft’s Xbox [touted](#) the range of settings available to both users and concerned parents. Elsewhere, proposals for “[middleware](#),” or third-party applications that change the user experience, make similar promises to increase user agency and platform safety.

On the one hand, customizable settings allow individuals greater agency, potentially leading to greater satisfaction with a service overall. On the other hand, too much emphasis on user settings as a solution to disruptive and harmful behavior can shift responsibility from corporate actors to parents and individuals with fewer resources and less know-how—especially if the most protective settings are not made the default. Consider, for example, that while social media platforms often offer users options related to data privacy, [very few users ever access those settings](#). This is not because they are not [concerned](#) about privacy; more likely they do not know such options exist or how to access them.

## Pro-social design as an alternative to reactive moderation and dark patterns

The classic, most basic conception of content moderation is reactive. A user does something harmful or disruptive. Moderators notice or are somehow alerted to it. Maybe it violates written rules, or maybe there are no codified policies governing behavior and moderators have discretion to act. Either way, they may take corrective action—like content removal or an account ban. This process has grown more complex over time and across platforms: for instance, large social media companies reduce reliance on user reports by deploying predictive AI classifiers to detect violative content or behavior.

Sometimes companies describe this form of moderation as proactive, but in reality it still takes place after undesired user behavior. “Pro-social design” asks developers in both the games and social media industries to [consider](#) approaches that encourage good behavior over disruptive behavior (a similar concept, [safety-by-design](#), seeks to anticipate and prevent online harms).

Behavior—good and bad—flows from design choices that create cues and incentives for users. A 2025 [paper](#) by researchers from the [Prosocial Design Network](#) notes that this field has “no canonical definition,” but offers a working one:

*“Prosocial Design is the set of design patterns, features and processes which foster healthy interactions between individuals and which create the conditions for those interactions to thrive by ensuring individuals’ safety, wellbeing and dignity.”*

Some stakeholders feel strongly that pro-social design is an important alternative to reactive moderation. For instance, a [report](#) by the Thriving in Games Group (previously the Fair Play Alliance) and the Anti-Defamation League states bluntly that “increasing attempt[s] to moderate and control behaviour is not a sustainable path.”

Both the gaming and social media industries provide examples of pro-social design; so do the fields of academics who study them. A 2022 study, for example, [found](#) that when players perceive themselves as mutually dependent on others, they are more likely to exhibit pro-social behavior. In a talk for the 2020 Game Developers Conference, Kimberly Voll and Weszt Hart [gave examples](#) of how games can promote sportsmanship over conflict, such as thoughtfully assigning roles to players by skill level or creating non-zero-sum systems for loot sharing. In the Digital Thriving Playbook, Hart offers other examples of ways in which games [encourage](#) players to treat each other with respect by creating options to assist, give gifts to, or congratulate others.

The social media space has its own equivalents. The Pro-Social Design Network has [collected](#) forty-five evidence-backed approaches, such as introducing a comment “[cool down](#)” period so that users leave more thoughtful comments (currently in use on Discord and some Facebook groups) or [nudging](#) users to move conversations into small groups instead of public threads (based on an experiment on Nextdoor). The Neely Center at the University of Southern California created a “[design code](#)” for social media offering nine consensus suggestions intended to “enable greater explicit user control, protect children through better defaults, improve incentives, and prevent small groups of users from manipulating and harming others.” Pro-social design features like supportive emoji reactions or comment cooldowns can discourage phenomena that are common in games and social media like [disindividuation](#) (the lack of empathy that can arise when interacting with strangers across a screen) and [emotional dysregulation](#) (failure to control ordinary frustration, which can lead to disruptive, even aggressive, behavior).

Incorporating pro-social design choices into the “[loop](#)” of user (or player) action can reduce disruptive behavior, increase satisfaction, and improve user retention rates. In gaming, a loop is a recurring sequence of actions performed by a player. Loops are often designed to trigger the brain’s dopamine-driven reward-seeking cycle, which underlies both gaming and social media use. Pro-social design asks developers across industries to consider how that cycle can be activated in ways that are non-coercive and which incentivize normatively positive behaviors over disruptive ones. It is both related and opposed to “[dark patterns](#),” the manipulative use of design to encourage compulsive use or overconsumption. While both rely on the ability of product design to engage the brain’s reward

circuitry, pro-social design emphasizes the agency and dignity of individuals while promoting respect, thoughtfulness, and other norms. In this way, it also challenges the idea that the problems with games and social media are wholly the fault of end users, whose behavior is shaped by at-scale incentives emerging from design choices.

Recognizing this, some [legislators](#) and policy [analysts](#) have expressed interest in governing technology platforms at the design level. While these efforts have primarily focused on social media, there are at least two reasons games might merit greater attention. First, the relative diversity of the gaming industry means there are many more laboratories in which to test interventions and approaches. And second, players may be more likely to leave a game, for which there are many more alternatives than there are for online social networks. It is possible that the risk of player churn might incentivize game companies to consider pro-social design more closely than companies in similar industries do.

## Multistakeholder approaches

The tech industry also refines the governance of online platforms through relationships with peer companies, nonprofits, researchers, and government officials through multistakeholder mechanisms that provide various benefits to Trust & Safety. [Nonprofit research institutions](#) have, for example, convened [cross-sectoral working groups](#) to explore shared challenges and areas of consensus and possible progress among stakeholders. Some multistakeholder groups like the [Working Group on Gaming and Regulation](#), bring together members from industry, academia, civil society and regulators to explore regulatory gaps, challenges and opportunities for online games. Others, like the [Robust Open Online Safety Tools](#) (or “ROOST,” as the partnership is commonly known), are developing open source trust and safety tooling for a variety of online platforms across both the gaming and social media industries.

Because Trust & Safety can span multiple issue areas, many of the multistakeholder governance efforts are topic-specific, such as combatting terrorism and far right extremism. In 2014, during the height of the mob harassment movement known as [GamerGate](#), Twitter [partnered](#) with a nonprofit—Women, Action, & the Media—to collect instances of gendered harassment on the platform. During the 2020 US elections, government agencies, nonprofits, and tech companies [met regularly](#) to discuss risks related to foreign interference, voter suppression, and political violence. The [Christchurch Call](#) led to the expansion of the [Global Internet Forum to Counter Terrorism](#), which provides a shared database of extremist content to enhance Trust & Safety responses. Research organizations such as the [Thriving in Games Group](#), the [Extremism Gaming Research Network](#) (EGRN), and the [Global Network on Extremism and Technology](#) (GNET) are working to develop [common analytical frameworks](#), pursuing solution-oriented studies, and [advocating for transparent access to platform data](#) to better understand and mitigate these complex issues.

Despite all of these efforts, formalized structures and externally mandated multistakeholder approaches remain limited—particularly in the gaming space, where scholars Kowert and Kilmer [argue](#) that “aside from the efforts of a few collaborative groups, progress in this space remains siloed within individual studios or behind third-party paywalls, restricting innovation as an industry.” As a result, much of this work is fragmented and voluntary. This is different from social media platforms, which have seen more explicit and formalized multistakeholder efforts, often driven by legislative and human rights frameworks to address gender-and sexuality-based harassment. For example, under the European Union’s Digital Services Act (DSA), most gaming platforms are not required to conduct assessments for systemic risks, including gender-based violence, or to provide data access for researchers to study these phenomena. These top-down regulatory approaches have indirectly fostered multi-stakeholder collaboration.

## 4. Trust & Safety governance in transition

The field of Trust & Safety is currently undergoing many changes. Large, in-house teams dedicated to platform safety are increasingly being fragmented across smaller teams and external vendors. Economic pressures—real or perceived, including the notion that generative AI can and should replace human moderation—have led executives to frame Trust & Safety as a cost center, rather than a long-term in-

vestment, resulting in layoffs and leaner teams across both social media and gaming industries. Generative AI is amplifying these challenges by creating novel types of harmful or deceptive content that existing safeguards struggle to manage. Meanwhile, regulators are playing closer attention to digital platforms, creating new compliance obligations and industry uncertainty about requirements and expectations. Trends like these demonstrate how Trust & Safety, while more important than ever, is changing in ways that are not uniformly positive and may, in fact, weaken the profession significantly.

### Small teams and the struggle for buy-in

If one school of thought holds that Trust & Safety is a business-essential investment in user experience, an opposing school views it primarily as a cost center. In social media, the latter view is currently ascendant. Technology firms are in a multi-year cycle of [layoffs](#) as executives scale back COVID-era spending, reduce employee headcount, and replace some job functions such as [mid-level software engineering](#) or [human rights risk assessments](#) with artificial intelligence tools.

A 2025 article by Rachel Elizabeth Moran, Joseph Schafer, Mert Bayar, and Kate Starbird, provocatively titled “[The End of Trust and Safety](#),” captures this dynamic through twenty interviews with T&S professionals currently or formerly employed in the social media industry. “Foundationally, it’s rooted in the long-term growth and viability of the platform. I believe T&S work is not a cost center. It is a profit enabler,” said one. However, another respondent cautioned that “the return on investment isn’t totally obvious.” A third T&S professional was blunter: “If I’m not in software engineering, developing a new pair of goggles or a headset, or coding to improve, I’m a cost. I’m not an investment.”

In gaming, by contrast, Trust & Safety teams started smaller. Necessarily so, because gaming companies have smaller overall headcounts, with staff often spread across many games, not all of which are online multiplayer experiences. After subtracting the number of game-company employees assigned to online games as opposed to single-player experiences, the numbers dwindle further. Those numbers never experienced growth powered by political inquiries around foreign election interference and other social media scandals. Despite that, they [may be growing now](#) as regulatory requirements affecting game companies and social media alike come online around the world.

As with social media, though, the question remains whether gaming executives will reach a consensus view that growing T&S teams are either a cost center or a safeguard for long-term value.

### Team safety and mental health are shared challenges

Another challenging aspect of T&S work described in “The End of Trust and Safety” is the mental health burden carried by workers who spend significant amounts of time confronting troubling forms of content such as graphic violence or sexual imagery. This challenge is well [documented](#) in social media contexts, where scholars note the high risk of psychological distress from exposure to hate speech, harassment, and troubling imagery, including child sexual abuse material (CSAM).

Similar risks exist in gaming. The Digital Thriving Playbook [mentions](#) several of the same content problems T&S workers face, in addition to challenges like managing abuse from “[volatile](#)” fandoms, which sometimes attack both T&S workers and game developers on social media. The Playbook also lays out specific jobs in gaming beyond content moderation that are still relevant to Trust & Safety. They include community managers, player support specialists, researchers, producers, T&S policy specialists, and “developer advocates” who engage directly with players to receive feedback and promote the product.

In both industries, professionals [report](#) facing personal [harassment](#) for their role as content moderators despite the backstage nature of their jobs. The feeling that their role is expendable and that they are not heard or valued by managers focused on profit maximization also damages morale in both industries.

### Vendors’ growing role deserves more attention

One consequence of tech sector layoffs has been the rise of third-party vendors that offer T&S as a service. “The End of Trust and Safety” notes that several of the interviewed professionals “chose to leave T&S positions for other opportunities, with most moving away from roles at major platforms to

instead take positions at vendor and BPO (Business Process Outsourcing) companies or to become self-employed consultants.”

Some of these companies provide services in other industries—[Kount](#), for instance, offers fraud protection for a wide variety of commercial client types including online gaming companies. Others, like [safetykit](#), advertise artificial intelligence tools for “manual review, onboarding, and investigations,” while [Clarifai](#) and [Unitary](#) promise dramatic improvements in content moderation efficiency using automated tools. Companies like [Hive Moderation](#) and [SightEngine](#) apply LLMs to challenges like identifying violative imagery, text, video, and music. [Modulate](#) similarly uses voice intelligence technology to proactively detect and mitigate toxic behavior in voice chat within online games. Many of these vendors service both social media and gaming companies; they include large and well-known cybersecurity vendors like [ActiveFence](#). Some companies, like [Cinder](#), offer their own full-stack T&S platforms while supporting open-source toolsets like [ROOST](#). Meanwhile, Microsoft created its own tool, [Community Sift](#), which it provides to other companies via “value-added resellers” like [Keywords Studios](#).

The growing role of vendors in and across the gaming and social media industries is both under analyzed and evolving rapidly. There are many open questions, for example:

- ***Will advertised AI tools deliver?*** The use of machine learning to classify and detect content that violates corporate policies is widespread and not new, but vendors may put AI tools in the hands of many smaller companies. Some of these tools may offer new capabilities; others may oversell their effectiveness.
- ***Can vendors advocate for T&S?*** As Trust & Safety professionals within large social media platforms warn that they are losing institutional clout and leverage, can third-party vendors effectively advocate for their clients to adopt better policies? Externally, how will their growing voice influence the broader industry?
- ***Will vendors cross-pollinate approaches and policy?*** In the [words](#) of a former Meta employee, “if you let [contractors] make the calls for you, you’ve outsourced policy.” How will the sharing of tools and personnel across companies and entire sectors affect the development of new approaches and capabilities? And how does the outsourcing of content moderation affect platform responsibility and accountability when failures happen?

## Generative AI poses new moderation challenges

While innovations in generative AI have created new opportunities for content creation and expression, they have also introduced complex moderation challenges for Trust & Safety in digital spaces. From the spread of AI-generated slop flooding [recommendation algorithms and newsfeeds](#) to fraudulent and [even harmful AI-authored books](#) inundating online marketplaces like Amazon, digital platforms now have to contend with increasing amounts of deceptive content. The scale and sophistication of this content has placed new strains on both the human and automated moderation systems set-up to limit harm—systems that are already under pressure from company layoffs and shifts in platform priorities.

In online games, these moderation challenges can take on additional dimensions as developers experiment with generative AI for gameplay and community interaction. Gaming companies are already experimenting with generative [AI-powered non-player characters \(NPCs\)](#) that can converse dynamically with players. Others are building generative AI engines for dynamic content creation that allow players to create custom levels, maps, or puzzles. For example, Roblox recently announced [a new generative AI tool](#) for creating dynamic environmental content. These applications create promising opportunities for the future of play, but can also introduce additional moderation considerations.

We already know that generative AI applications can lead to harm. There have been numerous reports of chatbots [giving dangerous advice to teens about self-harm and eating disorders](#), recommending [violent or extremist views](#), or being used to teach the next generation of cybercriminals [how to create malware](#). There are also security vulnerabilities that are relevant, including [prompt injection](#) that could allow players to bypass safety guardrails or extract sensitive data. In online games, similar risks could manifest, and given the interactive and immersive nature of games, these harms could be amplified.

Generative-AI moderation challenges are already difficult to address because they are highly context-dependent, and the deceptive quality of AI can be high. Studies have shown that [AI-generated content can be surprisingly persuasive](#), often mimicking human reasoning, style, and emotional cues that make detection more difficult. At the same time, generative AI content is unpredictable in at least three basic ways. First, gaming companies might build user experiences on top of models that have been trained by a third party, which might have training data or model specifications that could cause unexpected harm. Second, the instructions that companies provide to AI models (e.g. “do not advocate for violence”) may not work as intended. Third, users may circumvent both those [instructions](#) and [post-generation predictive safety tools](#) in ways that AI classifiers cannot detect when testing for policy compliance.

At the same time, existing regulatory frameworks are currently insufficient and struggle to keep pace with the rapid development and deployment of generative AI across all sectors of society, leaving platforms and developers to manage risks in real time.

## Rising regulatory scrutiny affects games too

Regulators around the world are paying greater attention to the digital sector, and while social media companies have received the lion’s share of scrutiny, games are increasingly caught in the same regulatory net. This regulatory convergence illustrates a key point of this paper: gaming and social media platforms face similar governance challenges and are subject to increasingly similar oversight mechanisms.

In Europe, the Digital Services Act requires online platforms to [submit](#) Trust & Safety transparency reports regularly. This includes some of the large gaming companies which act as online platforms, and reports for [Microsoft](#), [Roblox](#), [Nintendo](#), [Ubisoft](#), [Square Enix](#), and other companies are readily available online. These reports include basic statistics about types of content reported by users and moderated by platforms. They are shorter than the risk assessments required for [very large online platforms](#)—those with more than forty-five million users per month. Nineteen such platforms operate in the European Union. (None are gaming companies, but Facebook, YouTube, TikTok, Instagram, LinkedIn, Snapchat, Pinterest, and X are all included.)

A few companies also publish longer transparency reports in the United States. [Microsoft](#)’s Xbox has published transparency reports every six months since 2022. [Sony](#) released its first transparency report in 2025, and [Roblox](#) began releasing quarterly reports in 2023 after determining it qualified as a social media platform under [California Bill 587](#).

Observers are conflicted about the ultimate impact of rising compliance requirements for social media, for reasons that also apply to the gaming industry. In “The End of Trust and Safety” report, one interview participant said “I actually do think [the trend toward regulation] does set a... bare minimum, and companies will adhere to that, and especially in times when there’s cost cutting, it’s like... you can’t cut this thing now because it’s required by law.” This suggests regulatory requirements could protect T&S investments in gaming companies during economic downturns.

However, others share the concerns of legal scholar and former associate general counsel for Google, Daphne Keller, who [warns](#) about the “rise of the compliant speech platform” under which what is regulated gets measured, and “what gets measured gets managed,” leaving more complicated challenges by the wayside. For gaming companies, this could mean focusing on easily quantifiable metrics while neglecting nuanced community management approaches that have proven effective in gaming contexts.

Public interest research often relies on data from corporate platforms. This has been a [challenge](#) in the social media space, where companies have [raised prices](#) for data access and [deprecated](#) analytic tools on which researchers once relied. The European Union’s Digital Services Act created a process by which researchers may apply to access data from “very large online platforms” as defined by the law, but to date, no gaming platform has been designated as such.

Today, European and state-level legislation in the United States, along with the [UK Online Safety Act](#), impacts both social media and gaming companies. [Australia](#) has similarly expanded some regulatory mechanisms from social media to games. With the focus of regulation shifting to issues like child safety, future rules and legislation are more likely to impact games than in the past.

In the summer of 2025, the UK Online Safety Act required websites and apps to implement “highly effective” age verification to prevent minors from accessing certain websites or applications, including pornography and social media. Half of US states have also [passed](#) laws mandating some form of age verification for access to pornographic materials. Parents are increasingly concerned about the impact that exposure to self-harm or sexually explicit content can have on young users, as well as the [effects of long-term social media use on children’s brain development](#).

But many of the age-verification methods introduce [serious privacy and security risks](#), and fail to prevent children from accessing restricted content [while also excluding adults](#) who should have lawful access. In a study of 45 parents and teenagers conducted by the Center for Democracy and Technology, [most participants agreed](#) that age verification systems are overly invasive and risky and preferred parental consent models that allow families to make context-based decisions together. While there [might be some promising privacy-preserving approaches on the horizon](#), age verification is going to remain a contentious point of regulation for both social media platforms and online games, underscoring the need for [inclusive and accurate frameworks that are protective of all users’ rights](#).

## 5. Conclusion and lessons learned

As a large industry at the forefront of consumer tech, few questions related to games are ever answered definitively. This paper aimed to introduce conversations about online games and how those spaces are governed, especially when compared to social media platforms.

In September 2025, the Center for Democracy & Technology, the NYU Stern Center for Business and Human Rights, and American University’s Center for Security, Innovation, and New Technology held a symposium on “[Mainstreaming Trust and Safety in Online Games](#).” Here, we draw on expert reflections during that symposium and offer some important takeaways for how player behavior, industry practices, and regulatory scrutiny continue to evolve in tandem. Because these frontiers are in constant flux, we conclude with a set of practical observations and emerging areas for future discussion.

### 1. Play is a human right

The first is that, as the anthropologist T.L. Taylor reminds us, [play is a human right](#) specifically articulated in the UN Convention on the Rights of the Child. Sometimes considered a “forgotten right,” UNICEF considers it alongside the rights to culture and leisure, which are also enjoyed by adults. The right to play requires a thoughtful mix of independence from adult (or moderator) intervention, but not the absence of oversight or safety precautions. This is a difficult and subjective line to walk and requires careful deliberation between industry, regulators, parents, children, and player communities of all ages.

### 2. Rehabilitation is preferable to punitive responses

For both games and social media, industry professionals and researchers broadly seem to prefer rehabilitative approaches to moderation. A ban from a platform, like removal from a playground, is something of a last resort. If, as professor Lindsay D. Grace writes, play often incorporates forms of [rebellion](#), then punitive approaches can backfire by drawing users into adversarial relationships with norms and moderators. On the other hand, [nudging](#) approaches that remind users of [pro-social norms](#) (preferably those agreed upon and formed with participation from players) can improve content moderation outcomes without punitive action. These are not trivial observations. Consider that after Twitter [banned](#) 70,000 accounts after the January 6 insurrection, misinformation on the platform declined—but many of those individuals [relocated](#) to Gab, where they became more active and their content more toxic.

### 3. Generative AI is changing the industry

While online games have had a few high-profile features that use GAI, the industry has been slower to adopt the technology compared to social media platforms. This is in contrast to the games industry's historic role at the forefront of technological innovation. Over the past decade, the emergence of the graphics processing unit as a valuable economic resource with applications beyond games and video production (e.g. AI development and cryptocurrency mining) has transformed gaming from a driver of technological change to a more ordinary entertainment industry.

On the other hand, GAI has applications for Trust & Safety that transcend the divide between games and social media. Third-party T&S vendors are [building and deploying](#) bespoke tools that use GAI to detect and analyze policy violations with greater speed and accuracy than previous predictive approaches reliant on AI classifiers. This has ramifications for both large platforms (which struggle with the sheer scale of content on their service) and small platforms (which are often less well-resourced).

### 4. Weighing risks to user wellbeing and discrete vs. general regulation

In 2018, the World Health Organization released the eleventh edition of the International Classification of Diseases; it listed [gaming disorder](#) as a new condition, defined as “a pattern of gaming behavior... characterized by impaired control over gaming, increasing priority given to gaming over other activities to the extent that gaming takes precedence over other interests and daily activities, and continuation or escalation of gaming despite the occurrence of negative consequences.” The decision was [controversial](#); not all medical professionals agreed, and the Diagnostic and Statistical Manual of Mental Disorders (commonly called the DSM) did not include gaming addiction in its fifth edition.

Games—like social media—often make use of a dopamine feedback loop to drive user engagement. This is not (always) crass manipulation; in game design, a “[loop](#)” is a repeated set of actions players take in pursuit of objectives, and they are core to what makes most games fun. However, loops can drive compulsive behavior, leading to behavioral and financial risks for players. In this way, online games are akin to pre-digital games of chance, arcade games, and even [pinball machines](#). Today, other online services like [prediction markets](#) and [sports-betting apps](#) provide similar thrills, with similar risks. When these industries are considered holistically, the question for public health professionals and regulators becomes whether games (or social media) are harmful in unique ways that deserve discrete forms of regulation, or whether a broader set of safeguards are needed.

### 5. Age verification is not a panacea

Age verification solutions are often put forward as mechanisms to protect children online. While knowing a user's age can support some safety objectives, it is often treated as a cure-all for a [diverse set of risks](#) that are not, in fact, solved by age gating alone. Focusing on child safety as principally an “access control” problem can obscure many of the structural factors that shape harmful experiences online and risks producing policy interventions that are narrow and ultimately ineffective. More meaningful protection will instead require harm-specific regulations that target systems, design choices, and accountability mechanisms, with voluntary industry commitments serving as a complement—not a substitute—for enforceable standards.

## 6. Increasing transparency, cross-platform sharing, and cross-industry collaboration

Increasing transparency and cross-platform collaboration are essential to addressing today's trust and safety challenges. Many of the most pressing harms do not occur on a single platform, but instead move fluidly across online games, streaming services, social media platforms, and communication apps, making it difficult to assess risk in isolation. Two problems stand out.

First, gaming companies lag social media platforms in providing meaningful transparency about harms on their services. Regulators have pushed social platforms to share data with researchers, and gaming companies would similarly benefit from stronger reporting practices and partnerships with external experts. [The Working Group on Gaming and Regulation](#), for example, is drafting transparency guidelines recommending data disclosure of user reports by harm category, enforcement actions, and moderation capacity, including team size and structure.

Second, existing safety efforts remain siloed by sector, despite the cross-platform (and cross-industry) nature of abuse. Stronger inter-industry collaboration—between games, social media, livestreaming, and messaging services—is necessary to enable information-sharing, align safety standards, and disrupt harmful behavior that exploits gaps between platforms. As governments increasingly treat games and social media as part of the same regulatory landscape, the gaming industry's innovations in community management and Trust & Safety become relevant to policymakers governing all digital platforms. Regulators would benefit from understanding the unique approaches gaming companies have developed, particularly around community-driven moderation and harm prevention.

## 6. Considering digital lives holistically

This paper aimed to open a discussion among watchers of social media and digital regulation about the relevance of online games. As such, it has offered few concrete recommendations. By far its most important suggestion is that the governance of online games provides important insights for other spaces and, as such, merits more attention from decision-makers and researchers. When the study of the two is segregated—especially if such separation is informed by implicit or explicit belief that one is serious and the other frivolous—opportunities are lost.

Human lives are compartmentalized as individuals show different faces to commerce, politics, community, and family. But these divisions are artificial, and policy research has long recognized the entangled nature of the spaces where we govern, work, shop, and live. Why not where we play?

NYU Stern Center for Business and Human Rights  
Leonard N. Stern School of Business  
44 West 4th Street, Suite 800  
New York, NY 10012  
+1 212-998-0261  
bhr@stern.nyu.edu  
bhr.stern.nyu.edu

© 2026 NYU Stern Center for Business and Human Rights  
All rights reserved. This work is licensed under the Creative  
Commons Attribution-NonCommercial 4.0 International  
License. To view a copy of the license, visit:  
<http://creativecommons.org/licenses/by-nc/4.0/>.



Center for Business  
and Human Rights