For Honor, for Toxicity: Detecting Toxic Behavior through Gameplay

Alessandro Canossa
acan@kglakademi.dk
The Royal Danish Academy
Copenhagen, Denmark

Dmitry Salimov
dmitry.salimov@ubisoft.com
Ubisoft Blue Byte
Düsseldorf, Germany

Ahmad Azadvar
ahmad.azadvar@massive.se
Massive Entertainment - A Ubisoft Studio
Malmo, Sweden

Casper Harteveld
c.harteveld@northeastern.edu
Northeastern University
Boston, Massachusetts, United States

Georgios N. Yannakakis
georgios.yannakakis@um.edu.mt
University of Malta
Msida, Malta

ABSTRACT

Is it possible to detect toxicity in games just by observing in-game behavior? If so, what are the behavioral factors that will help machine learning to discover the unknown relationship between gameplay and toxic behavior? In this initial study, we examine whether it is possible to predict toxicity in the MOBA game *For Honor* by observing in-game behavior for players that have been labeled as toxic (i.e. players that have been sanctioned by Ubisoft community managers). We test our hypothesis of detecting toxicity through gameplay with a dataset of almost 1,800 sanctioned players, and comparing these sanctioned players with unsanctioned players. Sanctioned players are defined by their toxic action type (offensive behavior vs. unfair advantage) and degree of severity (warned vs. banned). Our findings, based on supervised learning with random forests, suggest that it is not only possible to behaviorally distinguish sanctioned from unsanctioned players based on selected features of gameplay; it is also possible to predict both the sanction severity (warned vs. banned) and the sanction type (offensive behavior vs. unfair advantage). In particular, all random forest models predict toxicity, its severity, and type, with an accuracy of at least 82%, on average, on unseen players. This research shows that observing in-game behavior can support the work of community managers in moderating and possibly containing the burden of toxic behavior.

CCS CONCEPTS

• **Human-centered computing** → **Human computer interaction (HCI)**; *HCI design and evaluation methods*; User studies.

KEYWORDS

toxicity, video games, labeled dataset, random forest, machine learning

ACM Reference Format:
Alessandro Canossa, Dmitry Salimov, Ahmad Azadvar, Casper Harteveld, and Georgios N. Yannakakis. 2021. For Honor, for Toxicity: Detecting Toxic Behavior through Gameplay. *Proc. ACM Hum.-Comput. Interact.* 5, CHI PLAY, Article 253 (September 2021), 20 pages. https://doi.org/10.1145/3474680

Authors' addresses: Alessandro Canossa, acan@kglakademi.dk, The Royal Danish Academy, Philip de Langes Allé 10, Copenhagen, Denmark; Dmitry Salimov, dmitry.salimov@ubisoft.com, Ubisoft Blue Byte, Düsseldorf, Germany; Ahmad Azadvar, ahmad.azadvar@massive.se, Massive Entertainment - A Ubisoft Studio, Malmo, Sweden; Casper Harteveld, c.harteveld@northeastern.edu, Northeastern University, 360 Huntington Ave, Boston, Massachusetts, United States; Georgios N. Yannakakis, georgios.yannakakis@um.edu.mt, University of Malta, 20 Triq l-Esperanto, Msida, Malta.

2573-0142/2021/9-ART253 $15.00
https://doi.org/10.1145/3474680

1 INTRODUCTION

Toxic behavior has been identified as a persistent issue in online games, especially online competitive games such as *Overwatch* [17], *League of Legends (LoL)* [24], *Defense of the Ancients 2 (DotA 2)* [14] and other multiplayer online battle arenas (MOBAs) (e.g., [3, 9, 15, 37, 39, 44]). According to a recent study by the Anti Defamation League (ADL), examples of toxicity range from emotional abuse to blaming others for losses, offensive words, derogatory appellatives, unsportsmanlike behaviors, and selfish conduct [5]. The same study reveals that of the 733 US gamers aged 18–45 who played online multiplayer games, 81% reported some form of harassment related to their race/ethnicity, religion, ability, gender, or sexual orientation in the previous six months, with a stunning 68% of online multiplayer gamers experiencing more severe abuse, including physical threats, stalking, and sustained harassment. This toxicity has an impact on the game experience as well as the well-being of players: the study reports that 64% of gamers feel harassment is shaping their gaming experiences, and such that players perform less, avoid and stop playing certain games, become less social and feel isolated, and, most concerning, have depressive or suicidal thoughts. It is thus not a surprise that toxicity has an impact on the retention and Life Time Value of games [16, 66].

Game publishers, platform owners, online voice-chat applications, and even the police and national intelligence and security services are aware of these issues and are working to confront them, but it is far from a trivial matter as freedom of speech issues, technical difficulties, and a lack of chargeable offenses on the legal side make toxic elements a challenge to extinguish [67]. Recently Sparrow et al. [3] examined the ethics of fighting toxicity from an industry perspective, based on a reflexive thematic analysis of 21 in-depth interviews with games industry professionals. One of the main themes emerging is that "the goal of completely eradicating toxicity is unfeasible and unreasonable". This effort further highlights that there is an important need for supporting community managers because (1) notions of 'right' and 'wrong' player behavior are sometimes unclear and (2) industry professionals can be unprepared and unsupported in making governing decisions.

In the Washington Post [67], Carlos Figueiredo, one of the founders of the Fair Play Alliance, which is a coalition of game studios and companies formed to combat toxicity [20], declared that "*identifying a mechanism to combat toxic elements would be a rising tide that would benefit everyone in the video-game industry.*" The current main "*mechanism*" is that of community peer-reporting. However, players do not tend to report offenses: fewer than half of respondents of the ADL study said they reported toxicity using in-game tools [5]. This happens for a number of reasons, including the effort required in the reporting process, reports not being effective or taken seriously, or toxicity being a normalized part of the play experience. Regarding normalization, a recent study by Beres et al. [9] finds that players who do not report rationalize the toxic behaviors (e.g., banter, typical of games) or absolve themselves of responsibility ('not my circus'). Sparrow et al. [68] point out that normalization may be a result of players having a lack in faith in reporting systems. Interestingly, players appropriate reporting systems for instrumental purposes other than what they are intended for [40], which may propagate a lack of faith in such systems. While efforts can be taken to improve community peer-reporting, given the issues associated with it, it is important to consider other strategies as well to help community managers and, ultimately, combat toxicity.

The "*mechanism*" we explore in this paper is automatically detecting toxic behavior through in-game behaviors, studied in the context of the game *For Honor (FH)* [75].

1.1 Definition of Toxicity

There is no standard definition of toxic behavior. In fact, as noted by Kou [39], researchers disagree upon the definition and scope of toxic behavior, and use many related and/or overlapping concepts like deviant behavior [66], griefing [21, 60], cyberbullying [44], trolling [32], anti-social behavior [45], prejudice [12], etc. However, in the context of games it is generally defined as behavior that *intentionally* disturbs other player's experience and well-being. The definition of toxic behavior in each game can vary, but, among others, it includes "flaming", acting nosy, cheating, and illegal behaviors [9, 10, 44]. The ADL defines it as "*disruptive behavior*" such as personally embarrassing another online player, calling offensive names, threatening with physical violence, harassing for a sustained period of time, stalking, sexually harassing, discriminating against by a stranger, or doxing [5]. The Fair Play Alliance even advocates the use of the term 'disruptive behavior' instead of toxic behavior as to them the key problem is "that players feel their games are disrupted by other players too often" and "Sometimes those disruptions come about innocently...sometimes they're completely our fault as developers because of game design or mismatched players, and sometimes it's a player behaving inappropriately or abusively" [20].

In an effort to more systematically understand and capture toxic behavior in games, both Kou [39] and Kowert [42] propose taxonomies of toxicity. Based on analyzing analyzing players' online expressions about toxic behavior in the '/r/leagueoflegends' subreddit, Kou [39] suggests that toxicity should be seen as "a situated sequence of player emotions or actions at either the individual or the collective level that put teamwork at a disadvantage" and presents five primary types of toxic behavior (e.g., communicative aggression, hostage holding), as well as five contextual factors that could lead to toxic behavior (e.g., in-team conflicts, perceived loss). Kowert [42] also recognizes the need for a situated understanding of toxicity: behaviors that are considered toxic in one situation might not be considered toxic in another. As a result, Kowert suggests the broad heading of 'dark participation' for all deviant behavior that takes place online and that any outcome of these behaviors that cause harm to another's health or wellbeing are then considered toxic behaviors. This dark participation, consisting of 17 actions derived from literature and social media suggestions, is categorized according to performance type (verbal or behavioral) and impact type (transient or strategic).

It is precisely the situated nature of toxic behaviors that renders automatized efforts of detecting toxicity so difficult. Because of this, in our study we rely on human-curated set of labels provided by Ubisoft community managers for the game *FH*.

1.1.1 Definition of Toxicity in For Honor. Most important in terms of how we approached and defined toxicity, for the purpose of this study we divided the player population between unsanctioned and sanctioned players. Sanctioned here means that a player has been reported to the community managers by other players and they had been confirmed guilty of a breach of the game's Code of Conduct and received a sanction. Sanctioned players are further subdivided in four groups according to the gravity of the offense (warned or banned) and according to the type of offense (offensive behavior or unfair advantage). These four categories constitute the sanction matrix (described in detail in Section 3.2.1) and represent the target labels that this study is trying to predict.

1.2 Why For Honor

According to Kwak et al. [44], the most prominent features that make certain online multiplayer games outlets for toxic behavior are: (1) *competitiveness*: games where players compete with others, victory is paramount, and it feels like the game is not fun if it is not won; (2) *anonymity*: as players use nicknames and most likely will not meet directly, they feel free to say anything or act like there are no consequences; (3) *counterfactual thinking*: a psychological phenomenon to imagine possible alternatives to what actually happened, which in online multiplayer games means that players tend to blame others for unwanted events; and (4) *negative social culture*: as players spend time in communities where there is no empathy and having fun watching other people suffer is normalized, it is a matter of time to adopt anti-social behaviors. MOBA games are "*notoriously toxic*" [50], most likely because they have all these aforementioned prominent features: they are highly competitive, depend on team-based efforts of anonymous players, and have developed a negative social culture over time. This is why such games have been of focal interest in studying toxicity in games [53]. As for the game *FH* specifically, it has been known for its toxicity with some players even saying that "*this game has the most toxic player base I've ever encountered in a video game*" (Havoc1003 on Gamespot)[1] or, similarly:

> *The game has the most toxic community I've ever seen like everyone just starts insulting you in chat if you beat them despite the fact they're just light spamming or if someone beats you they just keep calling you trash and insulting you out of nowhere. Or people start ganging up in brawls for no reason.* –KAAMG1 on r/forhonor[2]

or

> *I was told once that he hoped id catch aids and malaria.. hoped id get a death sentence but spend some time in jail to be raped and beaten and that when i finally decide to kill myself that id have hitler and the devil gang banging my body in the depths of hell. All because i through him off the ledge for 2v1 me.*–TrowserShnake on r/forhonor[3]

Similar to other MOBAs, the prevalent toxicity in *FH* offers the opportunity to study this phenomenon in this particular game. We were able to do this by working with a unique dataset that includes the full behavioral data of nearly 1,800 sanctioned players and comparing their behavior with unsanctioned players. The opportunity we considered here is to study if we can behaviorally distinguish these sanctioned players from unsanctioned players.

1.3 How the Game Industry Deals with Toxicity

As for "*identifying a mechanism to combat toxic elements*" [67], the industry has been focused for some time on devising strategies to curb toxicity in games [11, 71, 76]. The gravity of the problem, and the recognition that not a single company can address this alone, led to the establishment of the Fair Play Alliance in 2017, which as of today includes nearly 200 gaming companies, including Blizzard, Riot, and Ubisoft [20]. Some of these largest gaming companies have started creating systems to combat the unpleasantness in gaming communities, for example, players that are reported for racism or profanity can be temporarily muted by other players. Most of the time these measures are insufficient as players can evade them easily by omitting letters or adding numbers or special characters [80] when typing offensive messages or selecting user names. Players can also report toxic players using in-game menus or the reporting options offered by Xbox Live and PlayStation Network. These reports can lead to banning abusive players. For example, Blizzard recently banned more than 18,000 Overwatch accounts for toxic behavior [67]. Riot Games has been studying and trying to reduce toxicity for many years by adding pro-social in-game tips to encourage positive interactions, as well as implementing a system in which players are rewarded with in-game goods for sportsmanship and virtuous behavior. That brought down verbal abuse by about 6% and offensive language by 11% [49]. Blizzard's Jeff Kaplan (*Overwatch* Lead Designer) reported a decrease of more than 25% in both players being abusive and matches containing abuse after adding features that encourage positive comments and allowing players to create filters for their online matchmaking [36]. Ubisoft, where toxicity management is a priority, started a task force with the end goal to track negative player behavior, manage players that behave poorly, and implement features that will encourage players to improve their behavior such as chat improvements, and team kill tracking [2]. Overall, the various efforts from the industry seem to be concerned with implementing rapid and solid reporting systems and engineering pro-social behaviors. Stimulating pro-social behaviors is an effort that is promoted and suggested by the academic community as well [23, 35].

There have been calls, however, both in academia and in industry, to leverage machine learning [81] to help detect toxicity as this might be "*a rising tide*" [67]. For example, in examining toxicity among esports players Türkay et al. [74] concluded that:

[1] https://gamefaqs.gamespot.com/boards/168620-for-honor/75998673
[2] https://www.reddit.com/r/forhonor/comments/gsqoj0/toxicity_in_for_honor/
[3] https://www.reddit.com/r/forhonor/comments/64xboe/how_toxic_do_you_think_the_for_honor_community_is/

Game companies may also need to investigate systems to combat toxicity that do not rely on player reporting. With the rise of machine learning, perhaps in the future, game companies will not need active reporting to target toxicity.

The limited efforts thus far have focused on Natural Language Processing (NLP) and text data, and thus verbal actions [26, 54, 55, 69, 72]. These efforts are similar to how toxicity issues are addressed in other media, such as YouTube [58] and Twitter [25]. However, games are different from such media as users do not only demonstrate toxicity through verbal actions but also through behavioral actions [42]. While we include chat actions in our approach (i.e., not the content, just the behavioral act of chatting operationalized as the 'number of messages', see Section 4.2.5, Table 2, and Appendix), our work in detecting toxic behavior is focused on such behavioral actions, as a form of player modeling [81]. To our knowledge, this is the first study that attempts to detect toxicity through gameplay data *only* on a *large scale* using *machine learning*. Much to our surprise, the only work that comes close is from 2014 by Blackburn and Kwak [10], which uses in-game performance, initial user reports, and linguistic analysis of chat data to predict—using a supervised learning approach—crowdsourced decisions in *LoL*'s "*Tribunal*", a peer review system introduced in 2011 and abandoned in 2014. Still closely related is the work by Shen et al. [65] who leveraged a large-scale behavioral dataset from *World of Tanks* [79] to study individual and team-level predictors of toxicity using statistical models. Thus, predicting toxic behavior from gameplay data, especially from such data only, is still very much in its infancy.

Note that we do not advocate for the complete automatization of toxicity detection; we merely propose to supplement the manual efforts of the community managers in terms of proactive flagging of toxic players. We recommend that a final human verification is necessary to close the loop and impose a sanction.

1.4 Current Work

This study aims to establish a framework for prediction of toxic behavior in online games by (1) appropriately sampling players for comparison, (2) comparing multiple methods of categorization and feature selection, and then (3) using machine learning, in particular random forests (RF) and support vector machines (SVM) with the purpose of predicting toxicity for a large sample of data from the game *For Honor (FH)*, based on labels derived from the sanction matrix (see Section 3.2.1). We compared behavioral aspects of gameplay such as match performance, chat actions, and playtime patterns to predict not only the binary outcome of being sanctioned by community managers for involvement with toxic behavior but also the type of toxic behavior and severity of sanctions imposed on players. In other words, our aim was not only to determine if we can behaviorally distinguish 'sanctioned players' vs. 'unsanctioned players' but also the degree to which it is possible to identify in-game behaviors for players that have been labelled as toxic. Accordingly, we evaluated three hypotheses:

H1: Toxic players are behaviorally distinguishable from other players (defined by 5 types of metrics: activity modes, disengagement and AFK, movement modifiers, match performance, and chat actions, see Section 4.2 for a detailed explanation);

H2: Examining in-game behaviors, it is possible to distinguish between different levels of severity of toxic behaviors (defined as warned vs. banned, see Section 3.2.1 for details on the sanction matrix);

H3: Looking at in-game behaviors, it is possible to distinguish between different types of toxic behavior (defined as unfair advantage vs. offensive behavior, see Section 3.2.1 for details on the sanction matrix).

As discussed in this section, a few toxic players can negatively affect the player experience and psychological well-being for a large amount of players, resulting in a corrupted player experience, mental health issues, and in considerable revenue loss as many victims of toxic behavior tend to leave the game. Identifying toxic players is still very much an open problem left often to peer reporting and community managers. The findings of our work on toxicity detection through gameplay support the above-mentioned hypotheses, providing a fertile ground for further research. Specifically, we contribute (1) a scalable and generalizable method for detecting toxicity based on in-game behaviors, and (2) demonstrate this method in the context of the MOBA game *FH* with high accuracy results. Importantly, this research shows for the first time that by looking at in-game behaviors it is feasible to label toxic players, which can support the difficult work of human community managers in keeping a safe, healthy environment [3]. Even if human community managers are still involved in the sanctioning procedure, this research would provide a number of advantages: a faster response time for community managers; a wider reach in terms of the number of problematic players examined; more objective red flags for confirming or detecting potentially problematic players; and, finally, a parallel process that does not solely depend on players reporting offending individuals.

2 RELATED WORK

Toxicity in games has been a considerable problem for years. Therefore, the academic community has researched this topic thoroughly, from very different domains ranging from social studies to computer science. Although the present paper does not claim to offer an exhaustive overview of the field, it is necessary to briefly outline the extent of existing research efforts. Existing research can be grouped in several distinct areas: (1) studies on victims of toxicity, (2) studies on toxic players, and (3) studies on toxic games.

2.1 Studies on Victims of Toxicity

This kind of research, by far the most prolific domain, is focused at identifying common socio-demographic traits of the players that are most frequently victimized by anti-social behavior in online games as well as assessing the impact of the harassment. The ADL report [5] shows how toxicity is not restricted to but strongly tied to gender, race/ethnicity, and other player demographics.

Much of the recent work is centered on esports, which is a fast-growing area of research within games in general but also particularly relevant for toxicity due the extreme competitive nature of esports. Türkay et al. [74] investigated how collegiate esports players define,

experience, and deal with toxicity and use various coping mechanisms, including the fact that players often rationalize such toxicity as a normal part of gaming. Hayday et al. [30] explored current experiences of identity and esport community membership focusing on the ideological grounding, current practices, and tensions present within the communities. Madden et al. [48] focused on exploring gender biases in esports by interviewing 19 self-identified female and male professional gamers and event organizers, and find that gender biases in esports are a consequence of stereotypical gender roles in gaming tout-court (e.g., girls do not like violence, boys are competitive by nature) and that female gamers are looking for role-models and ways to grow in confidence.

As for gender, this has been another focal point of interest. For example, Kuznekoff and Rose [43] set out to determine how gamers' reactions to male voices differ from reactions to female voices and they found that "*the female voice received three times as many negative comments as the male voice or no voice. In addition, the female voice received more queries and more messages from other gamers than the male voice or no voice*". Then, McLean and Griffiths [51] explored female experiences of social support while playing online video games and they suggest that "*a lack of social support and harassment frequently led to female gamers playing alone, playing anonymously, and moving groups regularly. The female gamers reported experiencing anxiety and loneliness due to this lack of social support, and for many, this was mirrored in their experiences of social support outside of gaming*". This research proves that toxicity is more harmful to women, not only with respect to psychological well-being but also because of certain coping mechanisms, such as not using voice chat or hiding their gender. The previous study was confirmed by Eriksson and Bergström [18]. The authors, besides confirming that women are more affected than men, showed how toxicity puts women at a disadvantage within the game itself when trying to achieve higher ranks, compared to men. An additional insight comes from Fox and Tang [22]: the authors showed that harassment in general predicts women's withdrawal from online games. Despite this withdrawal, Cote [15] documented how women have actually built successful coping strategies (such as described above) but does call for a need for a cultural change to change the status of women as "outsiders" as these strategies have their limitations.

Much less work has focused on LGBTQIA players and players of color, which are also disproportionately affected. Through a survey of massively multiplayer online games (MMOGs) players, Ballard and Welch [6] find that male and heterosexual players engage in cyberbullying more than female and LGBT players, two groups of players who report to be the victim of toxicity more than male and heterosexual players. Gray [28, 29], on the other hand, has documented the experiences of women of color in Xbox Live, finding that they seek to build their own groups similar to women in esports in general [48]. Gray [27] also documented the experiences of black men in Xbox Live who are labeled deviant based on the stigma of their physical identity. Ortiz [59] finds that men of color in general cope with everyday racism in online gaming through a process of desensitization.

Cross-cultural analyses on toxicity are also limited. The few existing studies [63, 64] confirm the situated nature of toxicity suggested by Kou [39] and Kowert [42] (see Section 1.1), demonstrating that there is a culturally contextual aspect to toxicity. The authors also express that dictionary-based techniques are insufficient to detect, understand, and moderate toxicity due to cross-cultural differences in language use and norms, thus paving the way to consider techniques based on machine learning such as the one proposed here.

In our work, we do not focus on the victims of toxicity, but rather on detecting the toxic players, which is the next distinct area we discuss. However, as we describe in Section 6.3, we advocate for including the victims of toxicity in helping to further define and mitigate toxicity.

2.2 Studies on Toxic Players

Another area where academic research has focused on is studying the perpetrators, which involves trying to identify the socio-demographic markers as well as profiling and predicting their behavior both in-game and in physical life. Work in this area can be distinguished according to two approaches: (1) surveying players, which includes polling about experiences and behaviors but also exposing them to experimental manipulations; or (2) examining behavioral data, where the emphasis is much on chat data and thus *verbal actions* [42].

As for surveying players, following the work by Fox and Tang [22], Tang et al. [70] investigated the individual difference predictors for sexism in online gaming and find that hostile sexism, Social Dominance Orientation, sadism, Machiavellianism, and gamer identification predict self-reported sexual harassment in online video games. Lemercier-Dugarin et al. [45] also considered individual difference predictors but focused on aggressive behaviors broadly and considered video game habits, impulsivity, empathy, emotion reactivity, and motivations to play. Similar to other work in games [22, 32] and toxicity in general [8], they find that younger age, being male, and spending a lot of time playing per week, increased the likelihood of resorting to toxicity. Additionally, being highly achieving, having high emotional reactivity, and being high in two dimensions of impulsivity (negative urgency and sensation seeking) increased the likelihood of toxic behavior too. As normalization of toxicity is likely a key part of the problem in online games [9], Cary et al. [12] went beyond individual differences and looked at the interplay between individual differences and norms. They conform both are important for predicting if players engage in toxicity as well as report toxicity. The role of normative beliefs is experimentally tested (vignette of playing online multiplayer game vs. playing board game in a cafe) by Hilvert-Bruce and Neill [32] who find that harassment was perceived as more normal in online gaming contexts. Overall, this body of work highlights what type of individuals are more likely to engage in toxic behavior but also that contextual factors play a role.

As stated before, the other body of work, which focuses on examining behavioral data, has mostly focused on chat data. Such work includes the cross-cultural work mentioned earlier [63, 64]; the work by Ghosh [26] who analyzed Twitter and Reddit posts related to 13 popular games using NLP tools; and by Stoop et al. [69] who developed the framework HaRe (Harassment Recognizer), which is a machine-learning-based method to detect players that harass teammates or opponents in chat. Others focused on linking chat data to other behavioral data. For example, Märtens et al. [55] employed a novel natural language processing framework to detect profanity in chat-logs

and developed a method to classify toxic remarks, showing how toxicity is non-trivially linked to game success. This finding was confirmed by Neto et al. [57] using metrics. The study itself was expanded by Traas [73]: he found that toxic teams lose more matches if they were already losing and win less matches if they were already winning. Verschoor [77] considered the reverse relationship and showed how in-game events, such as the number of times that a player has died in the last minute, or the number of times that a player's team mates have died in the last minute, can predict toxicity in chat. As such, this body of work highlights the role of contextual factors and the impact of toxicity: game events can fuel toxicity, and toxicity itself influences game behavior.

In our work, we do not consider the content of the chat data. Shen et al. [65] ignore this too and similarly focus on whether individuals and teams have been reported as a measure of toxicity, but on a less granular level (i.e., toxic vs. non-toxic). Using a large-scale behavioral dataset from *World of Tanks* [79] they find that experienced and skillful players are more likely to commit toxic behaviors than newcomers, while losing teams and teams with high internal skill disparity among their members tend to breed toxicity. The most interesting finding is that toxicity is somewhat contagious: exposure in previous games has been shown to increase the likelihood that a player will commit toxic acts in future games. As mentioned in Section 1.3, even closer to our work is that of Blackburn and Kwak [10]. However, aside that they do consider the linguistic contents of chat data, they also included the initial report as part of their prediction model. In our case, we only consider the behavioral data.

Thus, considerable efforts have been made, through surveys or behavioral data, to either understand or detect toxic players. Our work fits into the latter, with the important distinctions that we (1) use gameplay data only, (2) do not intend to identify or describe the socio-demographic markers of toxic players, or (3) theoretically explain why players are toxic in our context. We simply try to predict if someone is a toxic player from their behavioral data. Additionally, the context of our work is on *For Honor*, which is a relatively less studied type of toxic game.

2.3 Studies on Toxic Games

The last line of research efforts is concerned with the affordances and circumstances that allow online games to breed toxicity, but also with the affordances and circumstances that help prevent or deal with it. In particular, Kordyaka et al. [37] set out to provide a clear theoretical explanation of toxic behavior in online games: they tested three different theoretical approaches (social cognitive theory, theory of planned behavior, and online disinhibition effect). They find that psychological (i.e., attitude and behavioral control), environmental (i.e., toxic behavior victimization), and technological (i.e., toxic disinhibition) constructs, as well as their interplay best explain toxic behavior. In other words, aside from individual differences, anonymity and behavior normalization, which includes being the recipient of toxic behaviors in the past, fuel toxicity. These outcomes confirm the contagion effect observed by Shen et al. [65]. While these theories are helpful, others (e.g., [12, 65]) use or refer to the theory of Social Identity model of Deindividuation Effects (SIDE) or moral disengagement [9] to explain toxic behavior (and normalization) in online games.

A very pertinent work in our further understanding of toxicity is by Sparrow et al. [3] who describe the perspectives from industry professionals, both designers and community managers. This work makes clear that considering ethics in the design process competes with functionality ("What button is it going to be on?"), and that certain ethical design decisions even come with risks to reputation, revenue, safety, and well-being. Complicating the matter further is that notions of 'right' or 'wrong' player behavior are sometimes unclear, making it difficult for industry professionals to make decisions, especially as they are put in "positions where they're having to monitor [and] guide thousands of people" and generally feel that they are unprepared and unsupported in making governing decisions. Overall, this work shows that from a design as well as community management perspective dealing with toxicity is complex. However, it also highlights opportunities for improvement, in particular for supporting community managers.

As for interventions to deal with toxicity, Kou and Nardie [41] suggest regulating anti-social behaviors is needed and point out the efforts of the game developer Riot with their "*Tribunal*", a peer review system that empowers players to judge misbehavior in two stages: in Stage 1 players submit a report and in Stage 2 players review the behavior and judge if the reviewed player should be pardoned or punished. While this system has been abandoned now, recently both Beres et al. [9] and Sparrow et al. [3] call for a reconsideration of this system as it is supportive of player agency in community moderation. However, peer reviewing in online games is also more complex and nuanced than seeing this system as simply being "slow" and "inaccurate". Kou and Gui [40] looked at the practices of community members to report (or flag) toxic behavior in *LoL*. They find that players (1) distrust the flagging system, (2) use the system beyond its intended use for toxicity, and (3) use it socially (e.g., team members discuss and "gang up" to flag another member).

As mentioned in Section 1.3, the game industry is considering fostering pro-social behaviors to address toxicity. Academic work suggest that this is a feasible route: people who play violent video games cooperatively engage in more pro-social and cooperative behaviors than those who play competitively [19], and play playing cooperatively appears to be associated with less aggressive behavior [33, 78]. A recent work by Johnson et al. [35] looked specifically at in-game helping behavior and suggest that fostering such behavior in games may help reduce in-game toxicity and improve well-being. Because it is just as important to foster such positive relationships in games, Frommel et al. [23] present a method to predict the quality of social interaction (defined as affiliation toward a partner). The work is based on audio, video, in-game, and self-report data from 23 dyads who played an online collaborative two-player game in an experimental setting. Similar to our work, the authors used random forest and support vector machine models. Together, all this work suggests that there is merit to not only utilize methods to detect toxicity but also to detect healthy player interactions (i.e., ones that foster empathy or emotional support).

Figure 1: For Honor (Ubisoft 2017): a typical fight between a samurai and knights.

While our work may complement efforts to regulate or mitigate anti-social behaviors described above, the present work only considers the possibility to detect toxicity from gameplay data. However, unique to this work is that this effort is pursued with the input from community managers and help of game designers, specifically in considering what game features should be considered in predicting toxic players.

3 FOR HONOR

For Honor (FH) [75] is an action multiplayer online battle arena (MOBA), first published in 2017. Like other popular MOBA games such as *LoL* [24] and *DotA 2* [14] the player takes on other players' heroes in a skill-based strategic combat arenas. The game takes place in a medieval fantasy setting, with a background story as follows: after a natural catastrophe happened, a warlord named Apollyon manipulated factions of warriors against one another in a fight over resources and territory, resulting in a war, with the aim to weed out the weak and create the strongest of men. In this game, players control a hero chosen from three factions, the Iron Legion, the Warborn, and the Dawn Empire, which represent knights, vikings, and samurai, respectively.[4] Each faction has four different classes: Vanguard (well-balanced offense and defense), Assassin (fast and efficient in duels, but less damage to multiple enemies), Heavies (slow in attack, but can withstand much damage), and Hybrid (combination of two of the three other classes). The gameplay blends together elements of fighting games, third-person action, and mass combat where the goal is to fight against the warlord Apollyon (Fig. 1). At the time of writing, over 20 million players have played the game. In the month of January 2021, there were on average 3,283 concurrent players online. Below we discuss in more detail the actual gameplay, as well as how behaviors in the game can be toxic and management thereof.

3.1 Gameplay

FH offers a third-person melee combat simulation with a variety of weapons and hero attributes. The game allows players to move freely about, sprint, climb, and even roll. Players can also lock onto single targets, entering a dueling mode that changes the control interface. In this mode, a player can choose to position their weapon in one of three sides (left, right, and up). If players have their weapon at the same angle as an oncoming attack, they will block the blow. Players may adjust their facing and stance maneuvering for a position of optimal advantage. To strike, a player must attack at one of two sides that an enemy is not defending against and choose between a light or heavy attack. A paper-rock-scissors of attacks, defenses, and guard breaks governs the engagement between combatants.

[4] In the Marching Fire expansion a fourth faction was added, the Wu Lin, which represent the ancient Chinese.

Figure 2: In-game reporting tool showing the kinds of red flags that can be generated by players about problematic individuals: offensive emblem, cheating, griefing.

3.2 Toxic Behavior in For Honor

Players are liable to receive a sanction if they breach any of the clauses in the game's Code of Conduct [1].The Code of Conduct was created based on insights from previous games as well as community managers and ranges from legally prohibited actions (e.g. hate speech, threats), naming and content policies (offensive or explicit names or images), advertisement, copyright infringement, misconduct (scamming and phishing), cheating (hacks and bots) and non-compliance with community managers and moderators. The Code of Conduct regulates in-game conduct, chat behavior, user names and any other actions that might affect in-game experience (hacking/modding). The categories of toxic behaviors described in the following sections are all based on the Code Of Conduct.

In *FH*, there is an elaborate system of identifying, prioritizing, and confirming various categories of infractions. The system relies on peer-reporting and community managers, which means players flagging each other for problematic behaviors that break the game's code of conduct and labeling the offense either as cheating, griefing or as an offensive emblem (Fig. 2). Once the community managers receive a player's report, a decision is made based on the report's severity: Minor first offenses are issued a warning (and possibly a request to change the offensive content), more serious offenses and repeated offenses may warrant temporary bans (of 1, 3, 5, 10, and 15-days), and it can escalate eventually to banning the account of the player permanently depending on the severity and frequency of the offense.

Not all reasons for sanctions are considered toxic, for example advertising or copyright infringements. For the purpose of this study we focused only on a subset of the ways in which the Code Of Conduct can be breached, that can be labelled as toxic:

1–*Offensive Language, Threats*: Posting or publishing any language or content that is hateful, racist, defamatory, ethnically or religiously offensive, obscene, vulgar, sexually explicit or inciting to the use of drugs.

2–*Harassment*: Harassing or bullying other players via verbal or written communications in and outside the game.

3–*Cheating / Modding / Hacking*: Running a modified or otherwise unauthorized version of the game client or a third party software providing an unfair advantage (wallhacks, aimhacks…) or causing detriment to other players' experience.

4–*Exploiting*: Triggering and using a bug or glitch, or bypassing established game rules by to gain a significant advantage or skip progression steps otherwise necessary.

The Code of Conduct also regulates the severity of the sanction administered to a player. Generally speaking a first offense would warrant a warning, a second offense would warrant a temporary ban from the game and lastly multiple offenses would result in a permanent ban. But particularly severe offences could warrant a temporary or permanent ban even if it is a first offence.

Table 1: Sanction Matrix: players can be sanctioned according to 4 labels dependent on severity (warned or banned) and type of toxic behavior (unfair advantage or offensive behavior), the examples in this table are taken from the official internal Ubisoft sanction matrix.

		Warned	Banned
Offensive Behavior	Definition	Offensive and harassing behavior (obscene, racist, hateful, vulgar, etc.) first offence	Particularly offensive and harassing behavior (obscene, racist, hateful, vulgar, etc.) or repeated offence
	Example	Usernames as Smokethejoint, GodRocks, Drinkcocacola	Usernames as Slut, Jesusisgay, Ariankiller
Unfair Advantage	Definition	Cheating or exploiting the game, first offence	Extreme cheating or exploiting the game, repeated offence
	Example	AFK farming without bots or cheat engine	Selling an account on Ebay or similar

As we have seen in section 1.1 offenses can be culturally relative, furthermore the Code of Conduct also leaves room for interpretation, therefore the process of defining, detecting, reporting and deciding what qualifies as sanctioned behavior may vary on a case by case basis. While we acknowledge that this manual process of identifying, reporting and confirming disruptive behavior is not free of bias, reviewing all the nuances of policies and procedures to label sanctioned players is beyond the scope of this study, therefore we take for granted the quality and trustworthiness of the sanctioning process in place at Ubisoft and accept the data to be a part of our study.

3.2.1 Sanction Matrix. For the purpose of this study we wanted to rely on the labelled dataset consisting of all the sanctions issued to players. It was necessary to focus only on sanctions issued for toxic reasons, so we focused on the 4 categories listed above: 1–Offensive Language, 2–Harassment, 3–Cheating and 4–Exploiting. These 4 categories have been grouped in 2 larger classes: *Offensive Behaviors* (categories 1 and 2) and *Unfair Advantage* (categories 3 and 4). In order to simplify the classes of the labelled dataset we also decided to group temporary bans and permanent bans in a single category, so we categorize the severity of a sanction either as a *warning* or as a *ban*.

Table 1 shows the Sanction Matrix that we established for this research, consisting of two types of offense and the two levels of severity; the table also includes some examples of infractions for each of the 4 profiles: warned offensive behavior, banned offensive behavior, warned unfair advantage and banned unfair advantage. The examples provided can give a sense of the types of infraction and severity levels. Players can be sanctioned according to four labels, according to severity (warned or banned) and type of toxic behavior (unfair advantage or offensive behavior). All the 1,793 players in our dataset have been labeled with these 4 sanction labels by community managers and that is what will allow us to map in-game behaviors with toxicity types and severity.

4 METHODOLOGY

In this section, we present the core aspects of our methodology for selecting the players to be considered for our analysis (Section 4.1), the method for selecting the features that would help us distinguish between toxic and non toxic players (Section 4.2 and Section 4.3), and finally the machine learning method (Section 4.4) used to learn to predict toxicity, its severity, and its type.

4.1 Player Selection

Only about 0.22% of all players of *FH* have ever been sanctioned. Of those players, the full set of behavioral data is available only for 1,793 players, specifically PC players. The number of sanctioned (toxic) players is very small compared to the number of unsanctioned players (see Figure 3) making our general population extremely unbalanced. Indicatively, out of over 20 million players in *FH* only approximately 45 thousand were ever labelled as sanctioned. Such an unbalanced dataset raises certain challenges with regards to machine learning and classification algorithms employed. In addition, not all of the behavioral game data is available for the whole player population due to tracking limitations, which further reduces the samples both for sanctioned and for unsanctioned players.

One simple way to cater for highly unbalanced data is to focus on the sanctioned player base, and sample a corresponding unsanctioned player base of equal size. While downsampling of the unsanctioned player class resolves the balancing issue, it raises questions about the representativeness of this class over the entire population of unsanctioned players. As our current focus is on examining the different ways we can identify toxic players, in this study we will adopt the downsampling approach while eliminating as much of its limitations. Given that our toxic player sample with extensive data available is very small (1,793 sanctioned players) compared to over 20 million players of the game, and the fact that a player could have shown disruptive behavior but not reported by other player and labeled as sanctioned by community managers, we put extra care when sampling from the unsanctioned player base. To test our initial hypothesis (H1: Toxic players are behaviorally distinguishable from other players), we need to select a control group of unsanctioned players with similar general behavioral patterns compared to the experiment group (sanctioned players), so that we can determine which specific aspects of behavior are unique to sanctioned players. Ideally, we would like to select unsanctioned players with no obvious gameplay behavior difference to sanctioned players. We can then leave it to machine learning to find specific features—beyond general gameplay characteristics—that will help us distinguish between toxic and non-toxic players.

To that end, we employ the k-nearest neighbors algorithm (k-NN) with $k = 1$ to find a closest match among the set of unsanctioned players for every sanctioned player. k-NN is a non-parametric classification algorithm; its input consists of the k closest training examples in

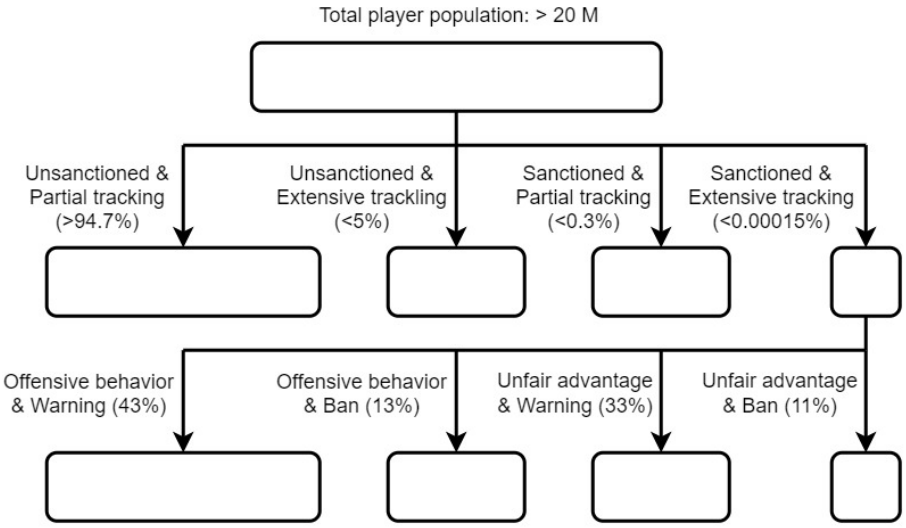

Figure 3: Sample distribution: from the total player population we have full tracking for less than 5% of the PC players; from the total population less than 0.3% of players were ever sanctioned. The intersection of these two sets returns 1,793 PC players.

the dataset whereas its output determines the class membership. In our case $k = 1$ which implies that the data object (the input of 1-NN) is assigned to the class of that single nearest neighbor. We base our 1-NN investigations on 6 general input features that we assume best serve to describe a *FH* player in general terms; these include the campaign progression rate, the total playtime, the ratio of PvP playtime relative to total playtime, the total hero level across all heroes, the KDA ratio (i.e., total kills and total assists over the total number of deaths), and the average rank (i.e., the average value of duel, kill, and objective rank). We chose these features as they provide accurate descriptions for important attributes such as playtime, ratio between time spent in single player versus multiplayer, basic performance descriptors and ratios between character versatility versus specialization. [5]

We noticed that 77% of the sanctioned activity occurs in 3 game modes: Dominion (DMN 33.65%), Duel (DL 21.89%) and Brawl (DDL 21.63%). Therefore we limit our study to those three game modes. In *FH* game modes are the rulesets that players chose to engage with the game. In *Duel*, players can battle other players or opt to battle an AI (game-controlled hero). In *Brawls*, players are joined in teams of two, fighting the opposing team. In *Dominion* mode two teams of four heroes must fight to control three zones on the map, one of which is the clashing point of both armies' of NPC fighters. Controlling the zones as well as defeating enemies grants points to the team. Once a team accumulates 1, 000 points, the opposing team is fighting for the last time, meaning that they will not respawn. After a team is defeated, points are counted to decide the winning team. We used these game modes to sample our player population: the majority of players engage with the 3 multiplayer game modes, where behaviors and attitudes are likely to have a tangible influence on the player experience of teammates and opponents alike.

4.2 Feature Extraction

Our feature extraction and selection method follows a hybrid process of involving a team of domain experts to select critical behavioral features (top-down) and relying on a statistical data analysis (bottom-up) for selecting additional features. For the top-down approach, a team of 8 experts that worked on the game was established with the purpose of flagging any game features that could be relevant for detecting toxicity. These experts' roles are: player experience manager, user research analyst, live coordinator, community manager, product owner, gameplay designer, technical director for community and security. The features that the team of domain experts selected bypassed the selection via frequency distribution comparison detailed in Section 4.3. The bottom-up approach consists of comparing the frequency distributions of the extracted features and select those whose difference between sanctioned and unsanctioned players was statistically significant. This process is explained in Section 4.3.

We extracted 36 features in total divided across 5 categories: *Activity Modes, Disengagement and AFK, Movement Modifiers, Match Performance,* and *Chat Actions*. For a comprehensive look at the 36 features, please refer to the table in the Appendix. The remainder of this section outlines each of the 5 aforementioned feature categories.

[5]The dataset utilized for this research is confidential and its sharing is regulated by GDPR, nevertheless the authors have requested Ubisoft legal department the authorization to share the anonymized dataset publicly.

4.2.1 Activity Modes. Activity modes define how players engage with the game. In a *custom game*, players can adjust detailed parameters of the match including number of players and AI-controlled heroes per team, rank, gear type, and severity of attacks, among others. *Practice mode* is a tutorial of basic movement and combat which is used to increase skill or remind oneself of the game rules and how it should be played. *Private matches* are a variation of custom matches that limits the matchmaking options to only include friends (1v1 game modes) and will be hosted by the group leader (as opposed to game servers). In *Ranked match* (added to the game in June 2017) players need to participate in qualifying matches then they are assigned to tiers and compete for a better placement. In *Tournament* (added to the game in June 2017) players need to participate in qualifying matches then assigned to tiers and to be crowned in their respective tier. Another approach to divide activities is the control of the opposing team which we categorized as *vs AI* or *vs Player*.

4.2.2 Disengagement and AFK. We also extracted a number of other behavioral measures to distinguish between the two samples related to abandoning the game and inactivity such as being away from keyboard (AFK), which we refer to as *Disengagement and AFK* features. These include the ratio of rounds where the player has exited the game manually and on their own volition, the ratio of rounds where at least one of the opponents has exited the game manually and on their own volition, and the ratio of rounds where the player was kicked by the game for being away from keyboard.

4.2.3 Match Performance. Another group of features we examined is labeled as *Match performance* and it consists of win rate and points acquired in Dominion matches. Note that we have previously used a measure of performance, KDA ratio (total kills + total assists / total deaths), as a selection criterion for our unsanctioned player population and by definition that measure could not be used for comparing samples.

4.2.4 Movement Modifiers. Another group of features pertain to movement modifiers within each modes of gameplay There are four types of movement style possible in *FH*, namely standing still, walking, running and sprinting which combined with the three types of game mode we selected for this study (duel, brawl and dominion) yields 12 options. We presumed that anomaly in average values of movement activity in specific game modes may be a sign of exploitative action.

4.2.5 Chat Actions. Finally, the last group of behavioral features aims at detecting eccentricity through in-game *chat* messages. We hypothesized that one of the channels of expression of toxic behavior will be manifested through the chat, yet we wanted to avoid taking the approach of text mining or Natural Language Processing because of privacy concerns and because there are already efforts in this direction. We focused instead on behavioral chat actions such as number of chat messages per minute, whether players message everyone or limit their messages to their team or group, and whether players take advantage of the quick chat feature to choose from predefined topics, such as help, courtesy, objective, strategic and navigational. We compared means of chat activity across all the distinct variables listed above.

4.3 Feature Selection

As explained earlier, out of the 36 features extracted, 4 were defined as critical features for the prediction of toxicity by the *FH* domain expert group, as outlined in Section 4.2. The 4 features selected by the domain expert group include all the disengagement and AFK features (abandon rate-self, abandon rate-opponent, and AFK rate) as well as the win rate. According to the *FH* domain experts the prediction of toxicity in that game depends on both the in-game performance (win rate) but also on the various expressions of the disengagement.

To select additional features for modeling toxicity in a bottom-up fashion we compare the frequency distributions of the remaining 32 extracted features and select those whose difference between sanctioned and unsanctioned players is statistically significant. The reader is referred to Figure 4 for the overview of the feature selection process we followed. The figure displays the frequency distributions of sanctioned vs. unsanctioned players across 9 of the 32 [6] different features extracted from the game. We perform Welch's t-tests on the frequency distributions across all 32 extracted features; the 9 features that appear to be significantly different between sanctioned and unsanctioned players are shown in Figure 4.

From the group of features "Activity Mode", the 2 features selected are *vs AI* and *vs Practice*, sanctioned players engage a lot more against AI-controlled team, while they engage considerably less in Practice, because they might already be proficient and do not need to learn the basics of combat. From the group of features "Movement Modifiers" the 2 features selected are*run rate* and *sprint rate* in Dominion game mode; sanctioned players consistently run more while unsanctioned players tend to sprint more. From the group of features "Chat Actions", the 4 features selected are messages per minute to everyone, messages per minute to own team, messages per minute to a specific group and strategic messages per minute; sanctioned players take advantage of the first three considerably more than unsanctioned players while unsanctioned players take more advantage of strategic messages. From the group of features "Match Performance" we selected the average personal points scored in Dominion game mode; unsanctioned players appear to score more than sanctioned players. Sanctioned players engage considerably less in practice, hinting that they have less experience with human players (also confirmed by their high rate of vs AI) and that may be why they run more and sprint less (the type of behavior you do against AI) and also win less than the average player. Maybe because of that frustration of losing against other players that they spam messages (but not the strategic ones because they are just frustrated).

The final set of the 13 features (4 supported by *FH* domain expert team and 9 selected) is presented in Table 2. Our analysis led to a small yet expressive feature set that captures various aspects of behavioral patterns that seem to be associated to toxicity behavior in the examined

[6] 36 features in total minus the 4 selected by the experts.

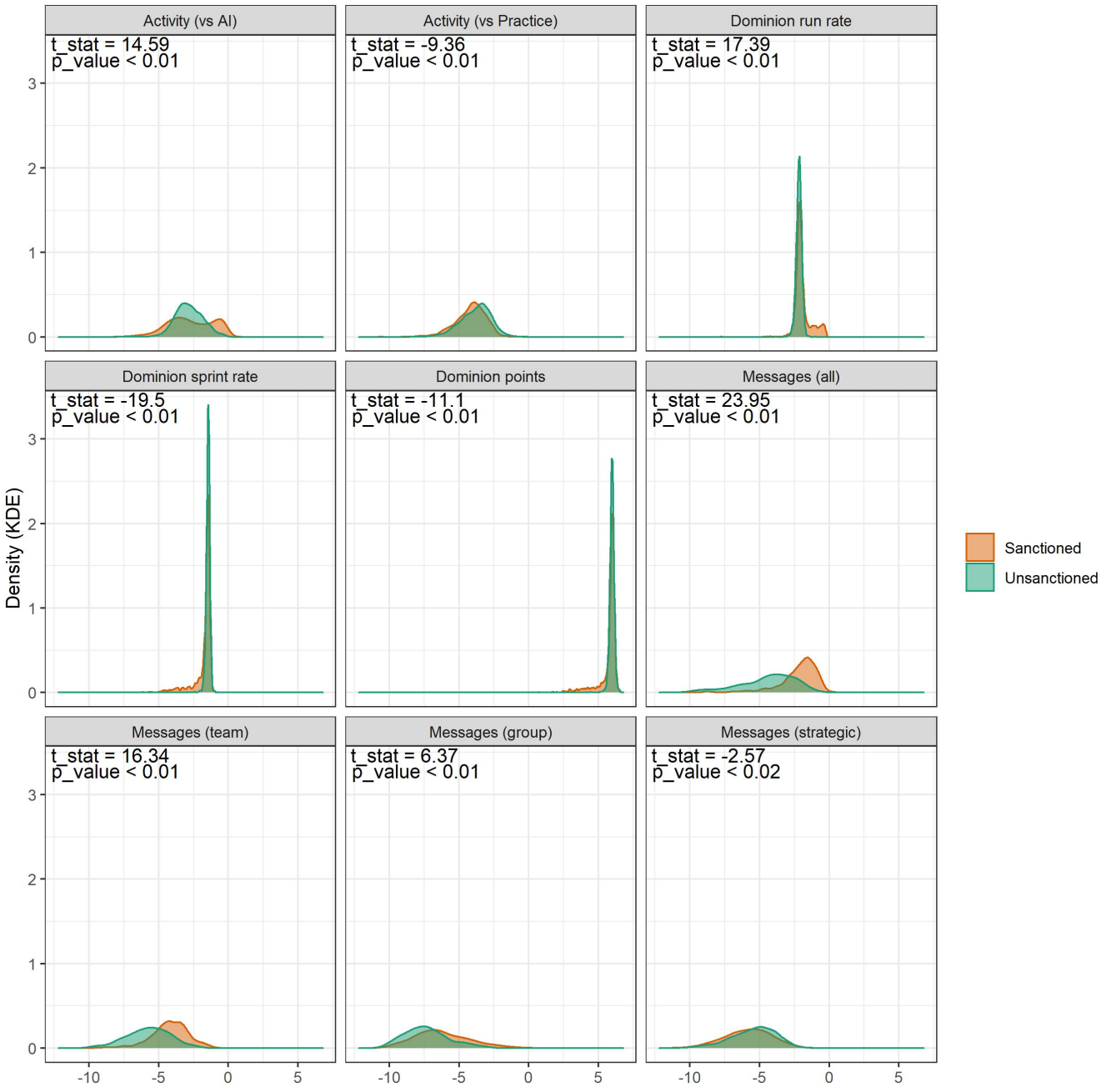

Figure 4: Feature selection method: frequency distributions of sanctioned vs. unsanctioned players across the variables that were selected based on Welch's t-test results.

game. It should be noted that the selected features are rather generic and thus might not be able to capture all possible situations in which a player can exhibit toxic behavior. We assume however that the set of features selected should be able to predict a large variety of such situations and thus form appropriate embeddings for classifying toxic players.

Table 2: The features derived from the feature selection approach followed in this study. The table provides information about the feature type and a corresponding short description. The four features appearing in bold are picked by a team of experts whereas the remaining features are selected though statistical analysis

Type	Feature Name	Description
Activity Modes	Activity (vs AI)	Ratio of "vs AI" matches over all matches
	Activity (vs Practice)	Ratio of "vs Practice" matches over all matches
Movement Modifiers	Dominion run rate	Average playtime spent running in Dominion game mode
	Dominion sprint rate	Average playtime spent sprinting in Dominion game mode
Chat Actions	Messages per minute (all)	'All' chat messages per minute of playtime
	Messages per minute (team)	'team' chat messages per minute of playtime
	Messages per minute (group)	'group' chat messages per minute of playtime
	Messages per minute (strategic)	'strategic' chat messages per minute of playtime
Disengagement and AFK	**Abandon rate (self)**	Ratio of rounds the player exited the game voluntarily
	Abandon rate (opponent)	Ratio of rounds an opponent exited the game voluntarily
	AFK rate	Ratio of rounds the player is labelled as AFK
Match Performance	**Win rate**	Average win rate
	Dominion personal points	Average personal points in Dominion game mode

4.4 Random Forests

In all experiments presented in this paper (see Section 5), we treat the problem of predicting toxicity, its severity, and its types as classification tasks. To construct accessible, expressive, and simple predictors of toxicity, in this initial study we employ *Random Forest Classifiers*. A Random Forest (RF) is an ensemble learning method, which operates by constructing a number of randomly initialised *decision trees* and uses the *mode* of their independent predictions as its output. Decision trees are simple learning algorithms, which operate through an acyclical network of nodes that split the decision process along smaller feature sets and model the prediction as a tree of decisions [46]. We select RFs as our supervised learning method in this study given their evident efficiency in modeling aspects of gameplay in the literature [52]. RFs also appear to be an appropriate method for classification given the moderate size of our dataset and the number of available extracted features. In this study we used the RF implementation available at the *randomForest* R library. [7] We initialise RFs with their default parameters and we set the number of trees to grow in the RF to 500. We used stratification [62] as a way to balance the data and thus name RF models that were trained on such data as stratified RF models. Stratification refers to stratified sampling [62]—an established technique used to normalize the performance of statistical models in unbalanced datasets (i.e. datasets with classes of substantially different sizes). For the stratified RF models we adjusted the number of trees to grow to $1,000$.

In addition to RFs it is important to note that we also employed support vector machines (SVMs) as an alternative baseline classification method. As SVMs achieved similar performance to RFs—and thus uninteresting findings—we will focus on experiments with RFs in this study.

5 DETECTING TOXICITY

In this section we present the core set of experiments run for the purposes of this paper. In particular, we examine to which degree we are able to detect sanctioned players as a function of their playing behavior (Section 5.1), whether we are able to model the severity of toxicity (Section 5.2), and finally model the type of toxic action performed (Section 5.3).

5.1 Modeling Toxicity (Sanctioned vs. Unsanctioned Players)

In our first attempt we view the task of modeling toxicity as a binary classification problem and we use random forests to distinguish between sanctioned and unsanctioned players in our dataset. The dataset size we examine initially is balanced and contains $3,586$ players; half of them ($1,793$) are labelled as sanctioned. We first investigate the predictive capacity of the 13 selected features as presented earlier. The results presented employ a 30% holdout validation method across 100 independent runs. We report the average accuracy and the 95% confidence interval values.

As seen in the confusion matrix of Table 3, the 13 features selected hold substantial predictive capacity as the toxicity models reach accuracy of 82%, on average with the 95% confidence interval across 100 runs lying between 80% and 83.7%. Comparatively, an SVM model trained the exact same data results in an average accuracy of 81.6% with the 95% confidence interval across 100 runs lying between 79.6% and 83.6%.

In an attempt to improve our testing accuracy, we do not merely rely on sampled players for whom we have 100% of the features available and expand the data sample by including players for whom we have at least 50% of the features available in the game data. This process results in an expanded data set of $24,513$ players. We process missing feature values via mean imputation and we overcome the resulting imbalance between the classes via stratification. As seen in Table 4, we can construct good predictive models even when some of the behavioral features are missing from our dataset. In particular, we reach high accuracy of 90.7%, on average, with the 95% confidence interval

[7] https://cran.r-project.org/web/packages/randomForest/

Table 3: Confusion matrix for the 2-class RF model based on sampled players across 100 runs. The dataset contains 3,586 players; the 30% testing portion reported here is 1,076 players. Rows are the actual sanctioned (S) vs. unsanctioned (U) players; columns are the predicted sanctioned (*S*) vs. unsanctioned (*U*) players. Unseen sanctioned and unsanctioned players are predicted correctly 77.2% and 87% of the time, respectively (in bold). The average accuracy of the 2-class RF model is 82%; i.e. the total number of correctly classified players over the total number of players.

	S	*U*	Total
S	415.5 (**77.2%**)	122.9 (22.8%)	538.4
U	69.9 (13%)	467.7 (**87%**)	537.6
Total	485.4	590.6	1,076

Table 4: Confusion matrix for the 2-class stratified RF model based on sampled players across 100 runs. The dataset contains 24,513 players; the 30% testing portion reported here is 7,354 players. Rows are the actual sanctioned (S) vs. unsanctioned (U) players; columns are the predicted sanctioned (*S*) vs. unsanctioned (*U*) players. Unseen sanctioned and unsanctioned players are predicted correctly 85.1% and 91.1% of the time, respectively (in bold). The average accuracy of the 2-class stratified RF model is 90.7%.

	S	*U*	Total
S	460.5 (**85.1%**)	80.4 (14.9%)	540.9
U	603 (8.9%)	6,210.1(**91.1%**)	6,813.1
Total	1,063.5	6,290.5	7,354

Table 5: Predicting Severity of Toxicity: Confusion matrix for the 3-class stratified RF model based on sampled players across 100 runs. The dataset contains 24,513 players; the 30% testing portion reported here is 7,354 players. Rows are the actual unsanctioned (U) vs. warned (W) vs. banned (B) players; columns are the predicted unsanctioned (*U*) vs. warned (*W*) vs. banned (*B*) players. Unseen unsanctioned, warned and banned players are predicted correctly 86.8%, 67.4% and 48.3% of the time, respectively (in bold). The average accuracy of the 3-class stratified RF model is 85%.

	U	*W*	*B*	Total
U	5,916.9(**86.8%**)	808.3(11.9%)	92(1.3%)	6,817.2
W	41.2(11%)	251.9(**67.4%**)	80.6(21.6%)	373.7
B	14.2(8.7%)	70.2(43%)	78.6(**48.3%**)	163
Total	5,972.3	1,130.4	251.2	7,354

across 100 runs lying between 90.0% and 91.4%. An SVM model trained on the exact same data results in somewhat lower average accuracy (88.1%) with its 95% confidence interval across 100 runs lying between 87.2% and 89.1%.

5.2 Modeling Severity of Toxicity

Given the very promising results we obtained in the binary (sanctioned vs. unsanctioned) experiments, our next step is to dive further into toxicity prediction and construct models that not only predict whether a player will be correctly labelled as toxic but also the *severity* level of the toxic behavior. In particular, sanctioned players either receive a *warning* (46,241 instances) or they are ultimately *banned* (28,904 instances) from the game. We test to which degree we can predict both toxic behavior and its severity by employing RF models that map between the game-related behavioral features and 3 classes: unsanctioned, warned, and banned. We use the full dataset of 24,513 players who have at least 50% of the features available in the game data and apply stratified sampling for balancing the 3 classes limiting the strata size by the size of the least common class (the accuracy baseline is 33.3%).

As seen in Table 5, our RF models reach an accuracy value of 85% on average with the 95% confidence interval across 100 runs lying between 84% and 86%. Note here that the model is able to distinguish well between toxic vs. non toxic behavior; the model, however, seems to misclassify banned players as warned almost as often as it correctly classifies them as banned.

5.3 Modeling Types of Toxic Behavior

So far our findings suggest that both toxic behavior and its severity can be predicted with sufficiently high accuracy through a small set of in-game behavioral features. In this section, we examine the degree to which our RF models can predict the type of toxic behavior. Such a predictor is critical to community managers as it provides a more nuanced and detailed information for any of their data-informed decisions about specific players and their behavior in the game.

Table 6: Predicting Toxicity Type: Confusion matrix for the 3-class stratified RF model based on sampled players across 100 runs. The dataset contains 24, 499 **players; the** 30% **testing portion reported here is** 7, 350 **players. Rows are the actual unsanctioned (U) players, offensive (O) players, and players seeking an unfair advantage (A); columns are the corresponding predicted classes (*U*, *O*, *A*). Unseen U O and A players are predicted correctly** 88%, 86% **and** 76.5% **of the time, respectively (in bold). The average accuracy of the 3-class stratified RF model is** 87.5%**.**

	U	*O*	*A*	Total
U	5, 997.4(**88%**)	448.3(6.5%)	371.7(5.5%)	6, 819.4
O	27(9%)	257.9(**86%**)	15.2(5%)	300.1
A	31.4(13.5%)	23.4(10%)	177.6(**76.5%**)	232.4
Total	6, 055.8	729.6	564.5	7, 350

The two types of toxic actions considered are a) *offensive behavior* and *unfair advantage*. The former toxic action is related to any type of offensive behavior observed during gameplay whereas the latter is related to behaviors that lead to unfair play. Note that these two toxic actions are not inclusive of all possible action types for which a player could be sanctioned in *For Honor*; however, they cover the vast majority of toxic behavior in the game (approx. 99% of sanctioned players which makes them representative) and we assume they would be easier to predict based on in-game behavioral features. We use the same data sample and the approach as in the previous sections; in this experiment, however, we omit 14 sanctioned players as they received sanctions both for *offensive behavior* and *unfair advantage*.

Across 100 independent runs, we manage to construct RF models that predict the 3 classes (*unsanctioned*, *offensive behavior*, and *unfair advantage*) with an accuracy of 87.5% on average with the 95% confidence interval lying between 86.7% and 88.5%. Table 6 shows the confusion matrix for the best RF classifier obtained. These results suggest that the type of toxic action can be predicted with a very high degree of accuracy.

6 DISCUSSION

We discuss here the outcomes of our effort to detect toxicity in online games through gameplay and the implications it has for game community management. We further argue for the need to better define what toxicity means, and explain how the HCI game community can help with this effort. Finally, we report the limitations of our work, which will help in understanding how our results can be generalized to other game contexts.

6.1 Toxic Players Are Behaviorally Distinguishable

For this initial study, which uses a dataset from the *For Honor* (FH) game, we aimed to see if we can distinguish 'sanctioned players' from 'unsanctioned players' first (H1). Then, we proceeded by evaluating if we can distinguish between different levels of severity (warned vs. banned) of toxic behavior (H2) and between different types (unfair advantage vs. offensive behavior) of toxic behavior (H3). In short, our work supports all three hypotheses: our random forest models can predict with a high level of accuracy (on average at least 82%) which players have been labelled as toxic, the severity of their behavior, and what type of behavior they committed, respectively. Typically, trying to predict more granular outcomes reduces the accuracy of a prediction models. Our models instead gain accuracy when trying to predict more precise outcomes, moving from an accuracy of 82% when predicting sanctioned and unsanctioned players to 85% when predicting the severity of the sanctions and to 87.6% when predicting the type of toxic behavior.This suggests that for predicting toxicity more precise outcomes can be added, such as severity and type of toxic behavior, without losing much predictability but gaining information about players. More importantly, it provides evidence that not only are toxic players distinguishable among other players, we can even rather accurately distinguish *between* toxic players based on their in-game behaviors.

In this initial study we selected players across the two classes whose behavioral characteristics (selected features) are similar using the k-NN approach. We followed this method in order for our sanctioned players to follow the underlying distribution of unsanctioned players, at least behaviorally. While the method produced datasets that contain similar types of players the RF method was still able to successfully distinguish between them. A potential avenue for further research is to examine alternative methods for populating the minority class of sanctioned players. Moreover, additional dataset balancing methods—including the Synthetic Minority Over-sampling Technique (SMOTE) [13] and the adaptive synthetic sampling approach (ADASYN) [31]—could be tested and compared against the results obtained with the stratified sampling.

Furthermore, it is important to note that selecting random forests (RFs) as our machine learning algorithm in this initial study has certain advantages with regards to our aims. Random forests do not only offer robust predictive capacity across any dataset we tried; notably, they offer a white-box, expressive method that is *transparent* to any community manager of the game. RFs—being a selection of decision trees—can inform any relevant stakeholder about the features involved in distinguishing between toxic vs. non-toxic players and their corresponding importance. This is one of the reasons we chose RFs over other machine learning techniques such as support vector machines (SVMs), which achieved a similar performance.

A critical part of our effort presented here was to first determine how we can compare toxic players (i.e., sanctioned players) with other players, especially as they make up a small percentage of the entire population (0.22%). To address this, we opted to use a *k*-nearest neighbors algorithm (*k*-NN) to find a set of unsanctioned players who are behaviorally similar to the sanctioned players based on more general input features that we believe best describe a *FH* player (e.g., total playtime, campaign progression rate). The next critical step in our method concerns the feature extraction and selection to compare these players. For this, we engaged with a team of experts working with the game, in order to reduce the number of features under examination and adjust their level of granularity. In our case, we originally extracted 36 features across five behavioral data types (see Section 4.2) and then selected 13 features, 4 features picked as critical by *FH* experts and 9 features through statistically examining the frequency distributions between sanctioned and unsanctioned players. While the exact process may differ, our method is reproducible for other MOBA games. The steps are as follows:

(1) **Player selection:** determine how to compare toxic players with other players. We suggest using *k*-NN to ensure behavioral similarity for key game features. We found that using a large sample (even with missing data, which we fixed with generalization) provides better explanatory power.
(2) **Game feature extraction and selection:** determine what in-game behavior to compare toxic players with other players on. We opted to work directly with the experts as well as statistically evaluating the frequency distributions between the two groups (using Welch's t-tests).
(3) **Model prediction with machine learning:** predict if toxic players are distinguishable. We suggest using random forests because of their transparency and thus explainability to community managers. We also gradually added classes to systematically examine the accuracy of different models.

In terms of misclassification, we find that there is less chance of misclassifying banned vs. warned players. This seems to suggest that banned players are behaviorally more distinct than warned players. As banned players' behavior is more severe, this aspect of the toxicity detection is desirable. Additionally, we find that unsanctioned players are easy to separate from banned and warned players. All in all, our work demonstrates a robust approach to detecting toxic behavior through gameplay.

6.2 Implications for Game Community Managers

Our work suggests above all that using gameplay data to detect toxicity is feasible. In fact, our results indicate that this can be done with a relatively low number of in-game behavioral features and reach a high amount of accuracy. However, misclassification can still occur in a few cases; therefore, it is strongly recommended that any automated effort to detect toxicity in the player community should not be deployed independently of human verification and a final confirmation that a certain player, classified automatically as toxic, did indeed break some of the rules stated in the code of conduct. As we stated in Section 1.4, we propose this study as a blueprint to create a tool to support community managers, not to replace them. This tool would allow the community managers to be more proactive and avoid relying on players reporting offending individuals, which, as we have seen, happens in less than half of the cases [5] and is often not used as intended [40]. Specifically, it would provide a faster response time for community managers; a wider reach in terms of the number of problematic players examined; more objective red flags and potentially help catch a much larger number of toxic players than what is usually the case by relying on players' reports. Detection based on gameplay would not represent the absolute solution to the problems of toxicity, but it would supplement the methods already employed by companies such as Riot or Blizzard with systems apt at eliciting pro-social behaviors.

More to the point, we imagine that if a detection method as described in this work would become part of the toolbox of community managers, they can deploy this to:

(1) **Identify extent and type of toxic behavior and determine mitigating actions:** being able to more objectively ascertain the extent and the type of toxic behavior in the community, it will provide community managers the ability to consider and discuss with the game designers how to mitigate this, for example through eliciting pro-social behaviors or suggesting changes in the match-making process.
(2) **Verify player reports on toxic behavior:** if the player reports match the outcomes from the detection method, then community managers can more rapidly respond to toxic behavior and more easily assess whether the player reports are accurate. While this mixed approach is a verification, we strongly recommend that community managers provide a final confirmation before sanctioning players.
(3) **Proactively identify toxic players:** instead of waiting for players reports, community managers can now actively monitor players that are likely to display toxic behaviors, until such behaviors are actually displayed. As stated, we advocate that community managers verify if players actually conducted toxic behavior before sanctioning them.

Besides the implication of complementing the toolbox of community managers in their fight against toxicity, our results from *FH* provide direct insights into issues that may help alleviate toxicity. First, descriptive statistics of the sanctioned players indicate that they practice less and run more but score/win less in Dominion matches, and play significantly more in vs. AI modes. This suggests that sanctioned players are not proficient in Dominion matches, possibly due to a lack of practice against human opponents and rushing through the Dominion game mode, which happens when playing in vs. AI matches. Game designers could provide players with a low win rate with tips or match-make them with less experienced players to keep them engaged and not frustrated with the game. More broadly, this suggests how game modes are used and players progress through them should be evaluated in order to decrease player frustration, which is a major contributing factor to toxic behavior.

Second, the excessive chat behavior among sanctioned players (excluding strategic messages, see Table 2) shows a strong will for connection and communication among sanctioned players. While abusing this feature may not be the ideal manifestation of that will, game designers can include more communication options but also incorporate frequency caps for messages per match to avoid spamming and disruption of other players' experience.

6.3 Toward Defining and Addressing Toxicity

For our work, we used the sanction matrix defined in Section 3.2.1, which is based on the Code of Conduct of the game, and essentially sanctions player according to four labels: severity (warned v.s banned) and type of toxic behavior (unfair advantage vs. offensive behavior). We used this sanction matrix as the starting point for mapping in-game behaviors, with the help of game designers, and to build our RF models around. Thus, the consideration and inclusion of game features is based on subject-matter expertise of this particular game. While our approach proved to be successful, the prediction results are only as good as the toxicity labels that were used. We believe the industry is in need of a more robust understanding of toxicity in order to address it better (through player reports and/or detection methods as presented here). This understanding starts with clear definitions of toxicity derived from qualitative and survey studies, but also necessitates grounded practical operationalization to bridge the 'descriptive' realities of toxicity with the 'predictive' models of toxicity. Although the taxonomies by Kou [39] and Kowert [42] provide a solid foundation, additional work is needed, including the input and help of the industry and player community, especially those who suffer the most from toxicity (i.e., female players, LGBTQIA players, players of color). Current work in the HCI game community has understandably first focused on describing player experiences [4, 74] and industry perspectives [3], how players make use of player reports [40] or perceive player behavior [9], but we call here on this academic community to take a leading role in establishing conversations and shared perspectives on what toxicity really means and how it should be dealt with by the community—with an emphasis on the perspectives of underrepresented groups of players. By leveraging HCI work concerned with amplifying the voices of underrepresented and vulnerable communities [38, 56], as well as leveraging inclusive participatory design [47] and Feminist HCI [7] practices, this will not only help amplify the voices of underrepresented communities in online game communities on how to address toxicity but also contribute to designing and keeping a more safe, healthy environment. We note that the careful consideration of underrepresented groups is especially critical when automated tools, such as proposed here, will be included to address the problem, which tend to be biased [82] or have difficulty handling underrepresented and vulnerable communities [61]. Last but not least, we advocate for more diverse and inclusive community managers, including underrepresented groups of players among their ranks.

6.4 Limitations and Generalization

Due to the focus on gameplay data, our work inherently touches on a subset of toxic behavior. This means we focused on *behavioral actions* and excluded *verbal actions* [42], the latter which has received far more attention in the literature (e.g., [26, 54, 55, 69, 72]). Future work should look into the overlap between behavioral and verbal actions, which is an area that Blackburn and Kwak [10] first investigated by combining in-game performance with linguistic analysis of chat data (in addition to user reports): are players who commit behavioral toxic actions also more inclined to take verbal toxic actions? Findings from such an effort may be able to shed further light on toxic player types, as well as what set of techniques are needed to comprehensively detect toxic players (e.g., combining random forest classifier on gameplay data with NLP classifier on chat data). We note that our work does include 'chat actions' but for this we only looked at the messages sent per minute of playtime and thus not the content of the messages. However, as stated, it is clear that toxic players make more frequently use of excessive chat messages compared to other players.

Another inherent limitation is that for modeling toxic behavior we had to reduce the number of features. As discussed above, defining toxicity in games first of all requires further scrutiny to improve classification. Second, detection methods are reductive and thus remove certain subtleties in toxic behavior. It is important to fully understand the implications of classifying behaviors at a more abstract level, i.e. does it matter what kind of cheating behavior a player does? To some extent, such questions are the domain of community managers, but it helps such managers in setting community rules and making decisions on warning or banning players if there is a better understanding about toxic behavior. Future research focused on the impact of specific toxic actions on player experience and game communities may facilitate such understanding.

The work presented here is demonstrated with the *For Honor (FH)* game. While this game has unique aspects (see Section 3), it is a fairly traditional MOBA game that shares many features and characteristics of this genre. As shown by Johnson et al. [34], MOBA games have some clear common features: (1) in terms of motivations, they cater less for needs of autonomy and relatedness compared to MMORPGs; (2) MOBA games seem to stimulate less immersion and presence compared to MMORPGs and RPGs; (3) MOBA games have less intuitive controls that RPGs and other genres, hence challenge and frustration are significantly higher in this genre; and (4) MOBA players get a sense of satisfaction from teamwork, competition and mastery of complex gameplay interactions. *FH* shares all these features, including a strong tendency for generating toxic behavior. In fact, the aspects of the gameplay data used in our work can easily be mapped to similar games such as *LoL* or *DotA 2* (e.g., average time spent running, average win rate, or chat messages per minute of playtime), suggesting that at least our method can be generalized to such games. Toxic behavior is of course not limited to the MOBA genre and for such other genres our method may not work; however, our work can still inspire others to consider how gameplay data can be used to identify toxic players. Regardless of the limitations and possibilities to generalize our work, it should be noted that efforts to generalize insights from one context

to another are welcome in the field as most current work on toxicity in games (whether qualitative or quantitative) is basing their findings on a single game.

In terms of generalization, and achieving our general aims to build *trustworthy* and *explainable* models of toxicity, we also need to consider the general limitations of data-driven methods of machine learning and the inherent biases (gender, racial, cultural, among many) that they carry through the data. All the RF toxicity models built for the purposes of this paper are trained on and predict data labels from a particular dataset. While the dataset is large and representative and the method appears to be robust in *FH*, future work will need to examine the degree to which the method we propose can generalize across dissimilar, potentially larger, and more representative datasets within this game. Once toxicity labels are available for other games, we would be able to test to which degree we can identify general patterns of toxicity across games and game genres.

7 CONCLUSION

We started this work with two questions: Is it possible to detect toxicity in games just by observing in-game behavior? If so, what are the behavioral factors that will help machine learning to discover the unknown relationship between gameplay and toxic behavior? From our study of players of *For Honor* we find that we can behaviorally distinguish toxic players from other players, and are even able to distinguish among toxic players in terms of the level of severity as well as the type of their toxic behavior. We obtained these results by (1) carefully selecting a sample of players, (2) extracting and then selecting game features based on input from designers and statistical results, and (3) deploying machine learning algorithms to predict toxic behavior. Altogether, this sums up our method for detecting toxic behavior through gameplay, which is scalable and generalizable to other MOBA games. Because this is, to our knowledge, the first study that attempts to detect toxicity through gameplay on a large scale, there are many opportunities to continue the research presented here. This is also pertinent because every day a few toxic players negatively affect the player experience and psychological well-being for a large amount of players.

ACKNOWLEDGMENTS

The authors would like to acknowledge the contribution of colleagues at Ubisoft Montreal for the labeled dataset and ongoing feedback on the work.

REFERENCES

[1] 2016. For Honor Code Of Conduct. https://forums.ubisoft.com/showthread.php/1490092-Code-of-Conduct. Accessed: 20-02-10.
[2] 2018. Dev Blog: Toxicity. ubisoft.com/en-us/game/rainbow-six/siege/news-updates/4mPLa195TWiXREafM4T4oC. Accessed: 20-02-10.
[3] Lucy A. Sparrow, Martin Gibbs, and Michael Arnold. 2021. The Ethics of Multiplayer Game Design and Community Management: Industry Perspectives and Challenges. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–13.
[4] Sonam Adinolf and Selen Turkay. 2018. Toxic behaviors in Esports games: player perceptions and coping strategies. In *Proceedings of the 2018 Annual Symposium on Computer-Human Interaction in Play Companion Extended Abstracts*. 365–372.
[5] Anti Defamation League. 2020. Free to Play? Hate, Harassment and Positive Social Experience in Online Games 2020. https://www.adl.org/free-to-play-2020. Accessed: 20-02-10.
[6] Mary Elizabeth Ballard and Kelly Marie Welch. 2017. Virtual warfare: Cyberbullying and cyber-victimization in MMOG play. *Games and culture* 12, 5 (2017), 466–491.
[7] Shaowen Bardzell. 2010. Feminist HCI: taking stock and outlining an agenda for design. In *Proceedings of the SIGCHI conference on human factors in computing systems*. 1301–1310.
[8] Christopher Barlett and Sarah M Coyne. 2014. A meta-analysis of sex differences in cyber-bullying behavior: The moderating role of age. *Aggressive behavior* 40, 5 (2014), 474–488.
[9] Nicole A Beres, Julian Frommel, Elizabeth Reid, Regan L Mandryk, and Madison Klarkowski. 2021. Don't You Know That You're Toxic: Normalization of Toxicity in Online Gaming. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–15.
[10] Jeremy Blackburn and Haewoon Kwak. 2014. STFU NOOB! predicting crowdsourced decisions on toxic behavior in online games. In *Proceedings of the 23rd international conference on World wide web*. 877–888.
[11] Chris Bruzzo. 2020. EA and the war against toxicity. https://www.gamesindustry.biz/articles/2020-06-23-ea-and-the-war-against-toxicity. Accessed: 20-02-10.
[12] Lindsey A Cary, Jordan Axt, and Alison L Chasteen. 2020. The interplay of individual differences, norms, and group identification in predicting prejudiced behavior in online video game interactions. *Journal of Applied Social Psychology* 50, 11 (2020), 623–637.
[13] Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. 2002. SMOTE: synthetic minority over-sampling technique. *Journal of artificial intelligence research* 16 (2002), 321–357.
[14] Valve Corporation. 20013. League Of Legends.
[15] Amanda C Cote. 2017. "I Can Defend Myself" Women's Strategies for Coping With Harassment While Gaming Online. *Games and Culture* 12, 2 (2017), 136–155.
[16] Pascal Debroek. 2020. The Impact of Toxicity on Retention and LTV. https://www.gamasutra.com/blogs/PascalDebroek/20201125/374272/The_Impact_of_Toxicity_on_Retention_and_LTV.php. Accessed: 20-02-10.
[17] Blizzard Entertainment. 2015. Overwatch.
[18] Niklas Ericsson and Hampus Bergström. 2020. How toxicity differ between male and female players in competitive Overwatch.
[19] David R Ewoldsen, Cassie A Eno, Bradley M Okdie, John A Velez, Rosanna E Guadagno, and Jamie DeCoster. 2012. Effect of playing violent video games cooperatively or competitively on subsequent cooperative behavior. *Cyberpsychology, Behavior, and Social Networking* 15, 5 (2012), 277–280.
[20] Fair Play Alliance. 2021. https://fairplayalliance.org/
[21] Chek Yang Foo and Elina MI Koivisto. 2004. Defining grief play in MMORPGs: player and developer perceptions. In *Proceedings of the 2004 ACM SIGCHI International Conference on Advances in computer entertainment technology*. 245–250.
[22] Jesse Fox and Wai Yen Tang. 2017. Women's experiences with general and sexual harassment in online video games: Rumination, organizational responsiveness, withdrawal, and coping strategies. *New Media & Society* 19, 8 (2017), 1290–1307. https://doi.org/10.1177/1461444816635778 arXiv:https://doi.org/10.1177/1461444816635778
[23] Julian Frommel, Valentin Sagl, Ansgar E Depping, Colby Johanson, Matthew K Miller, and Regan L Mandryk. 2020. Recognizing Affiliation: Using Behavioural Traces to Predict the Quality of Social Interactions in Online Games. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–16.
[24] Riot Games. 2009. League Of Legends.
[25] Spiros V Georgakopoulos, Sotiris K Tasoulis, Aristidis G Vrahatis, and Vassilis P Plagianakos. 2019. Convolutional neural networks for twitter text toxicity analysis. In *INNS Big Data and Deep Learning conference*. Springer, 370–379.
[26] Ayushi Ghosh. 2021. Analyzing Toxicity in Online Gaming Communities. *Turkish Journal of Computer and Mathematics Education (TURCOMAT)* 12, 10 (2021), 4448–4455.

[27] Kishonna L Gray. 2012. Deviant bodies, stigmatized identities, and racist acts: Examining the experiences of African-American gamers in Xbox Live. *New Review of Hypermedia and Multimedia* 18, 4 (2012), 261–276.

[28] Kishonna L Gray. 2012. Intersecting oppressions and online communities: Examining the experiences of women of color in Xbox Live. *Information, Communication & Society* 15, 3 (2012), 411–428.

[29] Kishonna L Gray. 2014. *Race, gender, and deviance in Xbox live: Theoretical perspectives from the virtual margins*. Routledge.

[30] Emily Jane Hayday, Holly Collison, and Geoffery Z. Kohe. 2020. Landscapes of tension, tribalism and toxicity: configuring a spatial politics of esport communities. *Leisure Studies* 0, 0 (2020), 1–15. https://doi.org/10.1080/02614367.2020.1808049 arXiv:https://doi.org/10.1080/02614367.2020.1808049

[31] Haibo He, Yang Bai, Edwardo A Garcia, and Shutao Li. 2008. ADASYN: Adaptive synthetic sampling approach for imbalanced learning. In *2008 IEEE international joint conference on neural networks (IEEE world congress on computational intelligence)*. IEEE, 1322–1328.

[32] Zorah Hilvert-Bruce and James T Neill. 2020. I'm just trolling: The role of normative beliefs in aggressive behaviour in online gaming. *Computers in Human Behavior* 102 (2020), 303–311.

[33] Jessica M Jerabeck and Christopher J Ferguson. 2013. The influence of solitary and cooperative violent video game play on aggressive and prosocial behavior. *Computers in Human Behavior* 29, 6 (2013), 2573–2578.

[34] Daniel Johnson, Lennart E. Nacke, and Peta Wyeth. 2015. *All about That Base: Differing Player Experiences in Video Game Genres and the Unique Case of MOBA Games*. Association for Computing Machinery, New York, NY, USA, 2265–2274. https://doi.org/10.1145/2702123.2702447

[35] Daniel Johnson, Xiang Zhao, Katherine M White, and Varuni Wickramasinghe. 2021. Need satisfaction, passion, empathy and helping behaviour in videogame play. *Computers in Human Behavior* 122 (2021), 106817.

[36] Jeff Kaplan. [n.d.]. Our progress so far. https://us.forums.blizzard.com/en/overwatch/t/our-progress-so-far/159046. Accessed: 20-02-10.

[37] Bastian Kordyaka, Katharina Jahn, and Bjoern Niehaves. 2020. Towards a unified theory of toxic behavior in video games. *Internet Research* 30, 4 (2020), 1081–1102. https://doi.org/10.1108/INTR-08-2019-0343

[38] Lindah Kotut and D Scott McCrickard. 2020. *Amplifying the Griot: Design Fiction for Development as an Inclusivity Lens*. Technical Report. EasyChair.

[39] Yubo Kou. 2020. Toxic Behaviors in Team-Based Competitive Gaming: The Case of League of Legends. In *Proceedings of the Annual Symposium on Computer-Human Interaction in Play*. 81–92.

[40] Yubo Kou and Xinning Gui. 2021. Flag and Flaggability in Automated Moderation: The Case of Reporting Toxic Behavior in an Online Game Community. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–12.

[41] Yubo Kou and Bonnie Nardi. 2013. Regulating anti-social behavior on the Internet: The example of League of Legends. In *iConference 2013 Proceedings*. iSchools, 616–622. https://doi.org/doi:10.9776/13289

[42] Rachel Kowert. 2020. Dark Participation in Games. *Frontiers in Psychology* 11 (2020), 2969. https://doi.org/10.3389/fpsyg.2020.598947

[43] Jeffrey H. Kuznekoff and Lindsey M. Rose. 2013. Communication in multiplayer gaming: Examining player responses to gender cues. *New Media & Society* 15, 4 (2013), 541–556. https://doi.org/10.1177/1461444812458271 arXiv:https://doi.org/10.1177/1461444812458271

[44] Haewoon Kwak, Jeremy Blackburn, and Seungyeop Han. 2015. *Exploring Cyberbullying and Other Toxic Behavior in Team Competition Online Games*. Association for Computing Machinery, New York, NY, USA, 3739–3748. https://doi.org/10.1145/2702123.2702529

[45] Maud Lemercier-Dugarin, Lucia Romo, Charles Tijus, and Oulmann Zerhouni. 2021. "Who Are the Cyka Blyat?" How Empathy, Impulsivity, and Motivations to Play Predict Aggressive Behaviors in Multiplayer Online Games. *Cyberpsychology, Behavior, and Social Networking* 24, 1 (2021), 63–69.

[46] Roger J Lewis. 2000. An introduction to classification and regression tree (CART) analysis. In *Annual meeting of the society for academic emergency medicine in San Francisco, California*, Vol. 14.

[47] Rachael Luck. 2003. Dialogue in participatory design. *Design studies* 24, 6 (2003), 523–535.

[48] Daniel Madden, Yuxuan Liu, Haowei Yu, Mustafa Feyyaz Sonbudak, Giovanni M Troiano, and Casper Harteveld. 2021. "Why Are You Playing Games? You Are a Girl!": Exploring Gender Biases in Esports. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–15.

[49] Brendan Maher. 2016. Can a video game company tame toxic behaviour? *Nature* 531, 7596 (2016), 568–571.

[50] E Makuch. 2014. MOBAs are "notoriously toxic," says Strife developer. Retrieved July 6, 2015.

[51] Lavinia McLean and Mark D. Griffiths. 2019. Female Gamers' Experience of Online Harassment and Social Support in Online Gaming: A Qualitative Study. *International Journal of Mental Health and Addiction* 17, 4 (2019), 970–994. https://doi.org/10.1007/s11469-018-9962-0

[52] "David Melhart, Antonios Liapis, and Georgios N. Yannakakis". "2021". "The Affect Game AnnotatIoN (AGAIN) Dataset". *"IEEE Transactions on Affective Computing"* ("2021"), "1–14". "in review".

[53] Marçal Mora-Cantallops and Miguel-Ángel Sicilia. 2018. MOBA games: A literature review. *Entertainment computing* 26 (2018), 128–138.

[54] Shane Murnion, William J Buchanan, Adrian Smales, and Gordon Russell. 2018. Machine learning and semantic analysis of in-game chat for cyberbullying. *Computers & Security* 76 (2018), 197–213.

[55] M. Märtens, S. Shen, A. Iosup, and F. Kuipers. 2015. Toxicity detection in multiplayer online games. In *2015 International Workshop on Network and Systems Support for Games (NetGames)*. 1–6. https://doi.org/10.1109/NetGames.2015.7382991

[56] Larissa Vivian Nägele, Merja Ryöppy, and Danielle Wilde. 2018. PDFI: Participatory design fiction with vulnerable users. In *Proceedings of the 10th Nordic Conference on Human-Computer Interaction*. 819–831.

[57] Joaquim AM Neto, Kazuki M Yokoyama, and Karin Becker. 2017. Studying toxic behavior influence and player chat in an online video game. In *Proceedings of the International Conference on Web Intelligence*. 26–33.

[58] Adewale Obadimu, Esther Mead, Muhammad Nihal Hussain, and Nitin Agarwal. 2019. Identifying toxicity within youtube video comment. In *International Conference on Social Computing, Behavioral-Cultural Modeling and Prediction and Behavior Representation in Modeling and Simulation*. Springer, 214–223.

[59] Stephanie M Ortiz. 2019. "You can say I got desensitized to it": How men of color cope with everyday racism in online gaming. *Sociological Perspectives* 62, 4 (2019), 572–588.

[60] Hunter L Paul, Nicholas David Bowman, and Jaime Banks. 2015. The enjoyment of griefing in online games. *Journal of Gaming & Virtual Worlds* 7, 3 (2015), 243–258.

[61] Ari Schlesinger, Kenton P O'Hara, and Alex S Taylor. 2018. Let's talk about race: Identity, chatbots, and AI. In *Proceedings of the 2018 chi conference on human factors in computing systems*. 1–14.

[62] Konstantinos Sechidis, Grigorios Tsoumakas, and Ioannis Vlahavas. 2011. On the stratification of multi-label data. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, 145–158.

[63] Sercan Sengün, Joni Salminen, Soon-gyo Jung, Peter Mawhorter, and Bernard J Jansen. 2019. Analyzing Hate Speech Toward Players from the MENA in League of Legends. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–6.

[64] Sercan Sengün, Joni Salminen, Peter Mawhorter, Soon-gyo Jung, and Bernard Jansen. 2019. Exploring the relationship between game content and culture-based toxicity: a case study of league of legends and MENA players. In *Proceedings of the 30th ACM Conference on Hypertext and Social Media*. 87–95.

[65] Cuihua Shen, Qiusi Sun, Taeyoung Kim, Grace Wolff, Rabindra Ratan, and Dmitri Williams. 2020. Viral vitriol: Predictors and contagion of online toxicity in World of Tanks. *Computers in Human Behavior* 108 (2020), 106343. https://doi.org/10.1016/j.chb.2020.106343

[66] Kenneth B Shores, Yilin He, Kristina L Swanenburg, Robert Kraut, and John Riedl. 2014. The identification of deviance and its impact on retention in a multiplayer game. In *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing*. 1356–1365.

[67] Noah Smith. 2019. Racism, misogyny, death threats: Why can't the booming video-game industry curb toxicity? https://www.washingtonpost.com/technology/2019/02/26/racism-misogyny-death-threats-why-cant-booming-video-game-industry-curb-toxicity/. Accessed: 20-02-10.

[68] Lucy Sparrow, Martin Gibbs, and Michael Arnold. 2019. Apathetic villagers and the trolls who love them: Player amorality in online multiplayer games. In *Proceedings of the 31st Australian Conference on Human-Computer-Interaction*. 447–451.

[69] Wessel Stoop, Florian Kunneman, Antal van den Bosch, and Ben Miller. 2019. Detecting harassment in real-time as conversations develop. In *Proceedings of the Third Workshop on Abusive Language Online*. 19–24.

[70] Wai Yen Tang, Felix Reer, and Thorsten Quandt. 2020. Investigating sexual harassment in online video games: How personality and context factors are related to toxic sexual behaviors against fellow players. *Aggressive Behavior* 46, 1 (2020), 127–135.

[71] Haydn Taylor. 2020. Valve introduces new measures to shut down team fortress 2 racist bot problem. https://www.gamesindustry.biz/articles/2020-06-18-valve-introduces-new-measures-to-shut-down-team-fortress-2s-racist-bot-problem. Accessed: 20-02-10.
[72] Joseph J Thompson, Betty HM Leung, Mark R Blair, and Maite Taboada. 2017. Sentiment analysis of player chat messaging in the video game StarCraft 2: Extending a lexicon-based model. *Knowledge-Based Systems* 137 (2017), 149–162.
[73] Arjen Traas. 2017. *The Impact of Toxic Behavior on Match Outcomes in DotA*. Ph.D. Dissertation. Tilburg University.
[74] Selen Türkay, Jessica Formosa, Sonam Adinolf, Robert Cuthbert, and Roger Altizer. 2020. See No Evil, Hear No Evil, Speak No Evil: How Collegiate Players Define, Experience and Cope with Toxicity. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3313831.3376191
[75] Ubisoft. 2016. For Honor.
[76] Rebekah Valentine. 2020. Infinity Ward pledges to do more to moderate racist content in its games. https://www.gamesindustry.biz/articles/2020-06-03-infinity-ward-pledges-to-do-more-to-moderate-racist-content-in-its-games. Accessed: 20-02-10.
[77] Mark Verschoor. 2016. *Eating seeds as a pastime activity: Predicting toxicity in online game chat using in-game events*. Master's thesis. Tilburg University.
[78] Julia Crouse Waddell and Wei Peng. 2014. Does it matter with whom you slay? The effects of competition, cooperation and relationship type among video game players. *Computers in Human Behavior* 38 (2014), 331–338.
[79] Wargaming. 2010. World of Tanks.
[80] Tony Xiao. 2019. Confronting Toxicity in Gaming: Going Beyond "Mute". https://www.nytimes.com/2019/06/06/learning/confronting-toxicity-in-gaming-going-beyond-mute.html. Accessed: 20-02-10.
[81] Georgios N Yannakakis and Julian Togelius. 2018. *Artificial intelligence and games*. Vol. 2. Springer.
[82] James Zou and Londa Schiebinger. 2018. AI can be sexist and racist—it's time to make it fair. *Nature* (2018), 324–326. https://doi.org/10.1038/d41586-018-05707-8

APPENDIX: FEATURE SELECTION

Group	Variable	Explanation	Feature selection status
Activity Modes	Custom Game	Participation rate in a custom game in which players can adjust detailed parameters of the match including number of players and AI-controlled heroes per team, rank, gear type, and severity of attacks, among others	Not selected
	Activity (vs. Practice)	Participation rate in Practice mode which is a tutorial of basic movement and combat which is used to increases kill or remind oneself of the game rules and how it should be played	Frequency distribution
	Private Match	Participation rate in Private matches which are a variation of custom matches that limits the matchmaking options to only include friends (1v1 game modes) and will be hosted by the group leader (as opposed to game servers)	Not selected
	Ranked Match	Participation rate in Ranked matches (added to the game in June 2017) in which players need to participate in qualifying matches then they are assigned to tiers and compete for a better placement	Not selected
	Tournament	Participation rate in Tournaments, in which Players need to participate in qualifying matches then assigned to tiers and to be crowned in their respective tier	Not selected
	Tournament - Unranked	Participation rate in Tournaments without the tiers	Not selected
	Activity (vs. AI)	Participation rate in activities where the opposing team is controlled by AI	Frequency distribution
	Activity (vs. Player)	Participation rate in activities where the opposing team is controlled by a player	Not selected
Movement Modifiers	ddl_run_rate	Percentage of time spent running in Brawl mode	Not selected
	ddl_spr_rate	Percentage of time spent sprinting in Brawl mode	Not selected
	ddl_stc_rate	Percentage of time spent standing still in Brawl mode	Not selected
	ddl_wlk_rate	Percentage of time spent walking in Brawl mode	Not selected
	dl_run_rate	Percentage of time spent running in Duel mode	Not selected
	dl_spr_rate	Percentage of time spent sprinting in Duel mode	Not selected
	dl_stc_rate	Percentage of time spent standing still in Duel mode	Not selected
	dl_wlk_rate	Percentage of time spent walking in Duel mode	Not selected
	dmn_run_rate	Percentage of time spent running in Dominion mode	Frequency distribution
	dmn_spr_rate	Percentage of time spent sprinting in Dominion mode	Frequency distribution
	dmn_stc_rate	Percentage of time spent standing still in Dominion mode	Not selected
	dmn_wlk_rate	Percentage of time spent walking in Dominion mode	Not selected
Match Performance	ddl_pts	Points scored in Brawl mode	Not selected
	dl_pts	Points scored in Duel mode	Not selected
	dmn_pts	Points scored in Dominion mode	Frequency distribution
	Win rate	Percentage of matches won	Expert suggestion
Chat Actions	total_playtime_min	Time passed from the start of the game	Not selected
	msg_pm_all	Number of messages sent to all players (per minute)	Frequency distribution
	msg_pm_team	Number of messages sent to own team (per minute)	Frequency distribution
	msg_pm_group	Number of messages sent to a specific group (per minute)	Frequency distribution
	msg_pm_courtesy	Number of messages of courtesy (per minute)	Not selected
	msg_pm_goto	Number of messages of directions (per minute)	Not selected
	msg_pm_help	Number of messages requesting help (per minute)	Not selected
	msg_pm_objective	Number of messages directing players to game objectives (per minute)	Not selected
	msg_pm_strategic	Number of strategics messages (per minute)	Frequency distribution
Disengagement and AFK	AFK rate	Ratio of the rounds where the player was kicked out by the game for inactivity (away from keyboard rate)	Expert suggestion
	Abandon rate (opponent)	Ratio of rounds where at least one opponent exited manually	Expert suggestion
	Abandon rate (self)	Ratio of rounds where the player exited the game manually and on own volition	Expert suggestion

Received February 2021; revised June 2021; accepted July 2021