Effect of Toxicity On Player Performance

Connor O'Toole, Cassandra Lee, Kirti Kumari, Ani Ganesh
University of Washington
Seattle, WA 98195 USA
{otool047, ckwl, kirtik2, anirud} @uw.edu

ABSTRACT

We conducted a study on the effect of toxicity within a cooperative game on player performance in an individual setting. We utilized the game, Snake, for a pre and post-test task as an indication for participant's individual mechanical skill where the difference in Snake score served as the player's performance score. The cooperative game, Keep Talking and Nobody explodes was played with the use of a confederate to expose the participants to toxic behavior. Participants in the control group, who were not exposed to toxicity had negative performance scores, whereas participants introduced to toxicity had positive scores.

Keywords
Toxic behavior, multiplayer gaming, cooperative gaming, individual player performance, stress

INTRODUCTION

People have been playing video games with others for as long as these games have existed. Even pong, the first video game ever created, was built for two people to play against each other. While competitive multiplayer games have always been enthralling, just as popular are cooperative multiplayer games, where strangers would need to team up to accomplish a common goal. Early 80s arcades saw multiple titles pitting multiple people against the game itself. The players could be tasked with cooperatively navigating a truck to save people falling out of a building, or pairing up to fight through waves of enemies together. With the advancement of computing, these cooperative experiences became more and more advanced, eventually allowing teams of players to face off against each other, thus combining the aspect of cooperation and competition among strangers. In recent years, some of the most popular online games have been built on this foundation. Popular titles include League of Legends, Dota 2, Fortnite, and Overwatch.

Despite the cooperative nature of these popular online multiplayer games, a growing issue seen across the game titles and communities is Toxic Behavior (TB). TB can be defined as a mental state of anger and frustration that can be enabled through real-time interaction and communication between players (Kordyaka et al, 2019). The level of competitiveness and nature of relative anonymity salient in the games makes them a likely environment for TB to develop and manifest (Kwak et al, 2015). As TB is a common occurrence observed across multiple populations of gamers, it has sparked interests in both industry and academic research to investigate the causes of TB, how to effectively identify and predict it, and combat it within various game settings (Kwak et al, 2015).

Regarding the causes of TB, researchers have investigated the underlying psychological process that causes certain players to engage in the behavior while others manage to avoid it. Anonymity plays a large role in many of the prevalent negative behaviors we typically see in online communities and group communication in general (Christopherson, 2007). Past social psychological research has shown anonymity to be a major factor in various phenomena such as group polarization (Isenberg, 1986), bystander apathy (Darley & Latané, 1968), and social loafing (Hoeksema-van Orden et al, 1998). We can these same phenomena occur in computer-mediated communication (CMC) as well. Sia et al. (2002) investigated how group polarization can occur in face-to-face (FtF) and CMC group discussions. Group polarization can be defined as the tendency for like-minded individuals to become more extreme in their thinking following a group discussion (Isenberg, 1986). By experimentally manipulating the social presence of individuals in a group meeting, Sia et al. were able to show not only that group polarization can occur within CMC settings, but

that it is most extreme in anonymous CMC conditions.

In addition to anonymity, the competitive nature of team-based online games is speculated to have an effect on the prevalence of TB within those games (Kwak & Blackburn, 2014). While competition is considered a good game design to increase enjoyment and investment in a game, it can also lead to intra- and inter-group conflicts that can escalate to aggressive or toxic behavior (Kwak & Blackburn, 2014). Using Tönnies (2012) classification schemes, the team structures in these games can be closely associated with task-oriented associations, where the relationship between players is impersonal and dependent upon progress towards an ultimate goal (Kwak et al, 2015). In these team-based games, when progress towards victory is halted, players tend to pick out and isolate individual members as scapegoats for the stunted progression, leading to TB (Kwak et al, 2015). These toxic players may feel justified in harassing teammates they see as a hurdle to winning the game (Kwak et al, 2015).

In this study, we attempt to bridge this gap to experimentally examine the effect that TB can have on a player's individual game performance. We believe there is still a substantial amount of work that needs to be done in regards to the effects TB can have on the players who are on the receiving end of such behavior. Our literature review tends to imply that being subjected to TB has a negative overall effect on individuals. Therefore, we hypothesize that when players are subject to TB in verbal form within a cooperative task-based environment, this will negatively affect the individual's performance in a simple, individual task.

LITERATURE REVIEW

While the causes of TB are certainly an area of interest for academics, much of the recent academic literature surrounding team-based games have been focused on identifying TB and predicting instances it will occur in games. Mora-Cantallops and Sicilia (2018) conducted a comprehensive review of the literature surrounding League of Legends and Dota 2 published since 2011 and found that the majority of published research (~56%) is focused on these aspects. This shows just how important researchers, both academic and industry, consider the topic of TB causes within the gaming landscape.

For instance, Kwak & Blackburn (2014) conducted a linguistic analysis of over 590 thousand League of Legends' in-game chat logs to gain better insight into the type of language toxic individuals typically use in their games. They found that there are three distinct phases to the game associated with different levels of chat volume, typically peaking at the beginning and end of a match (Kwak & Blackburn, 2014). Toxic players will usually be less active than normal players at the start of the match, but by the mid and end game, become much more active in the chat compared to the norm. They go on to discover 252 specific uni- and bi-gram elements used specifically by toxic players (Kwak & Blackburn, 2014). Most of these tend to be some variation of derogatory language or domain-specific attack based on the current circumstances of the game. 209 of these elements are found to occur in the late stages of the game, once again indicating that toxic players tend to ramp up their levels of toxicity towards the end of the game (Kwak & Blackburn, 2014).

To try and combat toxic behavior, gaming companies such as Riot Games and Valve have introduced in-game systems allowing players to report teammates and enemies for various toxic behaviors. However, these systems have inherent flaws, a major one being that they are reliant upon victimized players recognizing and taking action against the offending player (Kwak & Blackburn, 2014). This leads to many instances of TB going unreported, and thus players going unpunished, perhaps not even realizing that they were harassing or offending their teammates (Kwak et al, 2015). This phenomena of non-reporting can be attributed largely to the bystander effect, which describes the pattern of observers avoiding to help victims, particularly when they are immersed in a group (Kwak et al, 2015). In cases where the bystander effect is valid within a particular game, then most players will not report the offender even though they directly witnessed the abuse (Kwak et al, 2015).

As for the effect TB has on the victimized players, we need to look no further than the

real-life effect that cyberbullying has on people. There are a number of similarities between TB and cyberbullying, making it an apt comparison when looking for lasting behavioral effects (Kwak et al, 2015). In general, cyberbullying can cause far-reaching psychological problems, especially for younger members. Associated problems can include anxiety, depression, and has resulted in an individual committing suicide in extreme cases (Aggression & Roles, 2012).

However, while we can draw connections between TB and cyberbullying, this is mostly speculation and no specific evidence has shown a direct connection between the two. Because most of the research surrounding TB is focused on definition and identification, little research has been done specifically examining the effects that TB can have on victims, both within game and out. There is evidence showing that teams with toxic behaviors are more likely to lose compared to those without (Kwak et al, 2015), though no experiments have been run to identify a causal effect.

METHODS

For our study, we have used the true experiment method with 6 participants. We chose two games to design our research study, the first one is “Keep Talking and Nobody Explodes” used to induce verbal toxicity and observe the participant, and the second is the Snake game which was used to measure the player’s performance by keeping tracking of the score as pre and post-test. Initially, we observed how the participants react and perform while playing “Keep Talking and Nobody Explodes” with their partner. Then we asked them to play Snake in order to measure the score before and after the test. At the end of the session, participants were asked to fill out a post-test survey to describe the stress and their experience playing the game, followed up by a debrief session.

Participants (N = 6) were recruited from the University of Washington based social media groups (Facebook and Slack). All the participants were 20-29 years old and had previous experience with gaming, with a number of them have played some of the titles mentioned in the introduction (League of Legends, Dota 2). 6 out of the 8 participants were female.

During recruitment, the study goals were communicated as receiving feedback on how specific design concepts may affect key game mechanics and the player’s perception of the game. Participants were asked to describe their experience with videogames, including how often they play, what genre they play and how they approach playing. Gender and age were also tracked.

Participants were randomly assigned to a control group and treatment group. A confederate was used to interact with the participant in playing a cooperative based game. In the control group, the conversation flowed naturally with a neutral tone, while toxic statements were woven into the conversation between Confederate and participant in the treatment group.

When arriving at the lab, participants were asked to read and sign a consent form. They were then asked to perform a pre-study test in the form of a simple arcade game, Snake, where the player controls a snake that grows in length by collecting fruit, with the snake itself and perimeters of game space are primary obstacles. The score of Snake is equivalent to the amount of fruit collected. Participants were informed that this is a warm-up exercise prior to the primary game, “Keep Talking and Nobody Explodes”.

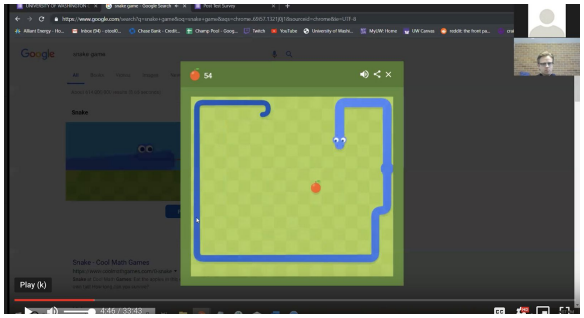

Figure 1. Screenshot from a participant playing Snake

Once the pre-test is completed, participants were shown a brief tutorial explaining that there are 2 main roles in the game, the Bomb Expert and the Bomb Diffuser, that need to work together to solve various modules to diffuse the bomb. The participant was assigned as the Expert and confederate was assigned as the Diffuser. The Expert was given a Bomb Manual which they used to instruct the Diffuser on how

to disable the bomb on the computer screen. The Expert did not see the bomb on the Diffuser's screen and vice-versa.

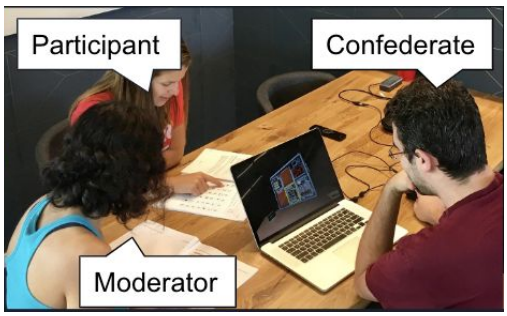

Figure 2. Study Setup Environment

The participant, as the Bomb Expert, played 3 rounds of the game with the confederate as the Bomb Diffuser. Each round had a 5-minute time limit and increased in difficulty (additional modules to solve). The first round was played naturally with the confederate using neutral terms for communication. The confederate continued using neutral language in the subsequent rounds when playing with participants in the control group. For those assigned in the treatment group, the confederate would use toxic phrases in the second round, continually increasing the volume of toxic phrases through the third round. The confederate had a list of predetermined toxic phrases and behaviors aimed to induce stress (Table 1). In the second round, the confederate used 5-8 toxic instances and increased the count to 12 for round three.

	Toxic Phrases/Derogatory statements
1	Can't you read any faster?
2	Yes! That's what I said, are you deaf?
3	You are so slow
4	Faster faster faster
5	We are both going to die because of you
6	6th graders are faster than you
7	How many times do I repeat myself?
8	I'm gonna throw this bomb at you
9	You suck
10	Arghhhhh!!!!
11	Uuuuffffff!!!!
12	Can I have a different partner?

Table 1. List of phrases that confederate used to induce toxicity in treated group

After playing the 3 rounds, participants in both groups were asked to play another round of Snake as a cool-down exercise. Lastly, participants completed a self-reported questionnaire to collect their stress levels and thoughts on their experience. Additionally, this was used to determine whether the confederate had a negative effect on the participant or not. Immediately after completion of the questionnaire, the moderator initiated the debrief session to introduce the confederate, explain the true purpose of the study, and answer any questions.

RESULTS

Manipulation Check

First, we wanted to look at the results from our manipulation check to ensure there was a perceived and noticeable difference between our experimental and control group. We ran two-tailed t-tests on each of the Likert scales from our post-task questionnaire. You can see the results in the graphs and table below.

Figure 3. Likert averages per round for perceived stress, difficulty, and performance

Here we can see that as we move to the later round of Keep Talking and Nobody Explodes, the participants in the treated condition experienced a rise in their perceived levels of stress, while those in control group went down as the rounds advanced.

Additionally, we can see a difference between the two groups in their perceptions of difficulty between the rounds. Oddly enough, we see that the treated group actually thought the second round was easier compared to the control group, but then jumps back up in the perception of difficulty for round three. This could either be

due to our small sample sizes, but could also indicate that toxicity does not raise perceptions of game difficulty when it is introduced, but rather after the toxicity has had some time to affect the individual. Additional research would need to be done to help clarify this point.

Finally, we can see that both control and treated groups perceived themselves as performing better throughout the rounds. This is most likely due to a practice effect, primarily from the fact that after three rounds of playing the game, they would become more comfortable with the format and mechanics. However, when we take this into consideration with their reported stress levels, this indicates that the treated group is potentially attributing that stress to an external cause. A good indication that our confederate was having an effect on them through their toxic behavior.

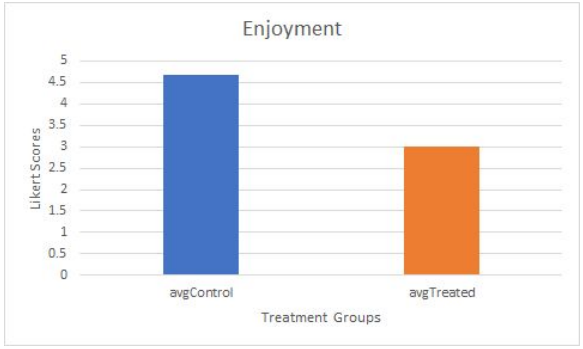

Figure 4. Likert averages for perceived enjoyment

Looking at the Likert scores for perceived enjoyment does seem to indicate a difference between the two groups. As expected, we see that the treated group enjoyed playing the game less overall compared to the control group.

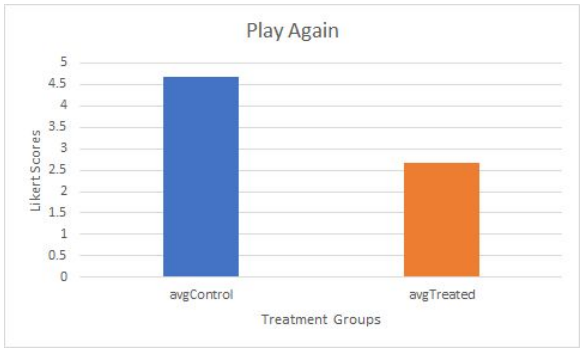

Figure 5. Likert averages for likelihood to play the game again

When asked how likely the participant would be to play the game again, we see that there also appears to be a difference between our two groups here as well.

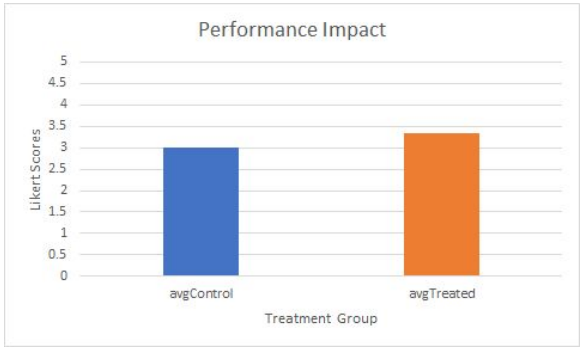

Figure 6. Performance impact of confederate on participant

Here we can see that participants in both groups indicated a similar level of impact from the confederate on their performance within the game. This is to be relatively expected because they were playing a cooperative game, so it makes sense that both groups, control and treated, would see their partner as having some level of influence on how well they performed.

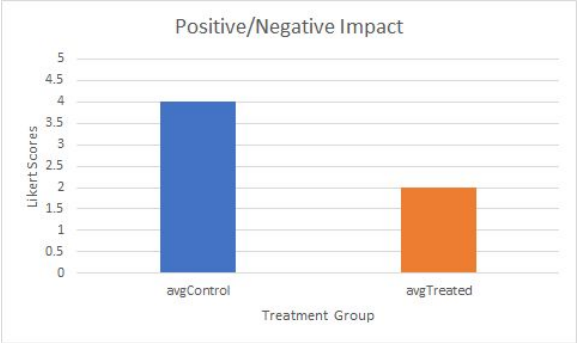

Figure 7. Positive/Negative impact on performance

As a follow up to the previous question, we asked participants whether the impact the confederate had on their performance was positive or negative. And as expected, we can see there appears to be a difference between the two groups, with the control indicating the confederate having a more positive impact and the treated group indicating a more negative impact. This question potentially gives the best insight into whether or not our manipulation was successful, so it's a good sign to see a potential difference here.

Manipulation Check	p-Value	Reject Null Hypothesis
Enjoyment	0.2322	No
Play Again	0.1394	No
Performance Impact	0.8194	No
Impact PosNeg	0.0742	No

Table 2. Table of significance for manipulation check

While many of the graphs from our manipulation check seemed to indicate a difference between our control and treated groups, after running two-tailed t-tests on each scale, we can see that there are no significant differences for any of the categories. However, there's a good chance that the lack of significance is due to our small sample size, and the results do indicate that there is a potential difference. So, while we can't say for certain that our experimental manipulation had a significant effect on our participants, we can make a reasonable assumption that the confederate's toxic behavior affected them on some level.

Main Effect

Keeping this in mind, we moved on to statistically testing our hypothesis, that being that the control group would perform better in their second round of snake compared to the treated group. Because we did predict that the treatment group would perform worse than the control group, we used a one tailed t-test as our primary statistical test for significance. Running the test at a 95% confidence interval yielded a value of p-value of 0.02445, meaning that there was a significant difference between our two groups and that we should reject the null hypothesis.

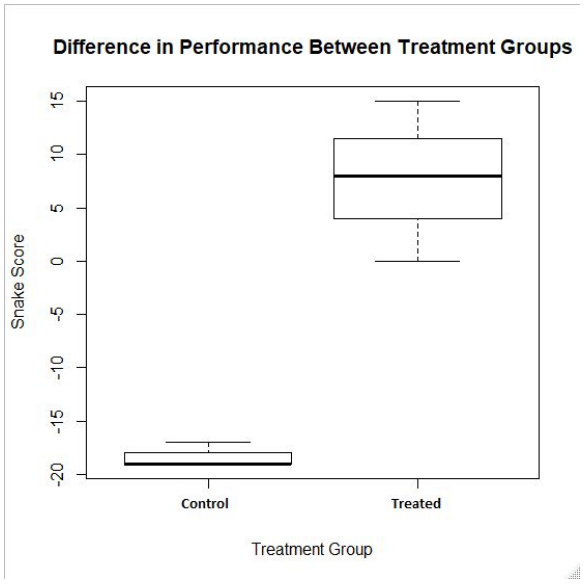

Figure 8. Box plot of scores between control and treated groups

Looking at a graph of the data shows something rather interesting; while there does appear to be an obvious difference between the two groups, it is not in the direction that we thought it would be. The participants who were exposed to the toxic behavior actually performed better than those in the control condition. So while we rejected the null hypothesis, we also rejected our alternative hypothesis.

Moderating Variables

In addition to our main dependent variable, we examined a number of possible moderating variables to see if there were any obvious correlations. The data we used for these tests were all based on the participants' answers from the screener survey we sent them.

First, we looked at the correlation between their snake scores and the amount of time they typically spend playing video games each week. Next, we examined whether or not the approach participants take in regards to gaming was correlated with their scores at all. Finally, we examined the correlation between their preferences in terms of who they play with and their snake scores. We used Pearson's product-moment correlation test as our main tool of examination. Refer to the table below for a list of results.

Moderating Variable	**r-Value**	**p-Value**
Hours played per week	0.4037	0.4273
Gaming approach	0.4859	0.3285
Gaming preference	0.5421	0.2665

Table 3. Table of correlation for potential moderating variables

Similar to the results generated for our manipulation check, we can see that none of the p-values are significant for any of the variables. Though again, with such a small sample size, each of the correlational coefficients seems large enough to indicate a potential trend, meaning that it could be worth exploring each of these variables further in a future study.

DISCUSSION

The current study evaluated the effects of toxicity on an individual's performance in mechanical movements in a video game. The main result showed in a positive performance in the player group exposed to toxicity. This could be explained by Yerkes–Dodson law, which describes the curvilinear relationship between mental arousal, in this case, stress, and performance. Stress and performance increases then performance drops off when there is too much stress (Yerkes and Dodson, 1908).

There was a single level of toxicity introduced to participants in this study. When comparing the phrases used in the study with the language commonly used in real game environments, the study's phrases used by the confederate can be considered relatively mild in comparison. While the study phrases follow patterns of toxic players, including increased aggression of non-strategic abusive language, it did not reach the level of derogatory and discriminative words that highly toxic players have (Blackburn and Kwak, 2014).

When analyzing performance based on the participant's experience in video games, operationalized by their hours spent playing in a week, we found that participants that play more had positive performances in the study task when subjected to toxic behavior than those that play less often (less than an hour a week). With more hours spent playing games, the likelihood of experiencing toxic behaviors is higher which could be a factor in their ability to handle toxic behaviors. For less experienced players, toxicity may have induced too much-added stress for them to handle, as reflected in poorer performance and modeled in the Yerkes-Dodson law.

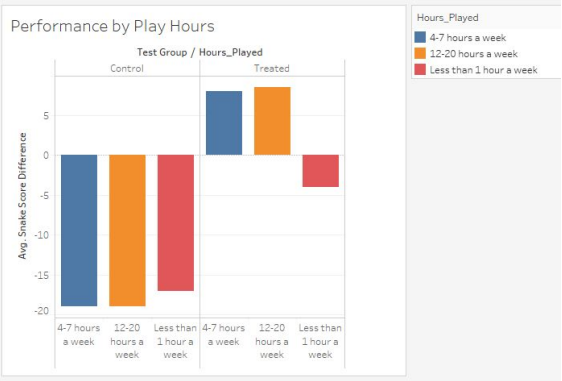

Figure 9. Impact of performance by hours spent playing video games in a week

Limitations

This study has some limitations that we must consider that prevents generalization of the results and weakens the validity.

Sample Size

The sample size for this study is too small to draw statistical significant in the results. Within the sample size, there was too much variance in each participant's background that was not controlled for. This included the gaming background in what types of games they played and how they approached gameplay. There was not enough representation in each group to clearly discern whether there was an effect.

Snake Performance

The participant's experience with the pre/post test game, Snake, was not considered. Previous experience could have affected the study's individual performance measure, where more versed players of snake could inflate the scores. As such, these results should disregard the magnitude of the performance score and focus on the sign of performance.

Game Environment

In order to easily and correctly measure the participants' individual performance in a cooperative setting, we separated these into two distinct parts with the pre/post task and main activity to isolate individual performance from the potential influence of the confederate's performance. Most gameplay experiences are not usually isolated in this way, where player performance is more difficult to attribute. Additionally, the artificially created toxicity does not necessarily represent the degrees of which a player may experience in-game, where the language is likely to be harsher. These lab experiences are not reflective of the in-game environments participants that players would normally encounter.

FUTURE WORK

We would like to explore conducting a similar goaled study to test the Yerkes-Dodson law with the relationship between arousal, in the form of stress formed from toxicity, and player performance. There have been previous studies exploring the Yerkes-Dodson law in video games, however, the focus was on arousal and memory (Jeong, & Biocca, 2011).

For this next study, we would select a different multiplayer cooperative game, one with a larger scale that would not only increase sample size but also create a more natural gameplay experience. The game would need roles or actions that can be easily attributed to each player so that individual performance can be isolated without removing the participant from the experience. Ideally, the game would have built in telemetry to track relevant data points for performance analysis. Additional points for data analysis outside of gameplay for further exploration include modeling stress behaviors in physical motions or changes in tone and volume in their voice.

For recruitment**,** we would like to consider participant's background experience such as preferred gameplay (solo, cooperative, competitive), approach to game (collab, compete, explore, express), etc. in order to understand their behavior and gaming experience. This would help us to understand what role does gaming experience and player's

behavior plays when it comes to the toxic game environment.

REFERENCES

1. Aggression, R., & Roles, P. M. (2012). Cyberbullying in Japan. Cyberbullying in the Global Playground: Research from International Perspectives, 183.
2. Blackburn, J., & Kwak, H. (2014, April). Stfu noob!: predicting crowdsourced decisions on toxic behavior in online games. In Proceedings of the 23rd international conference on World wide web (pp. 877-888). ACM.
3. Christopherson, K. M. (2007). The positive and negative implications of anonymity in Internet social interactions:“On the Internet, Nobody Knows You’re a Dog”. Computers in Human Behavior, 23(6), 3038-3056.
4. Darley, J. M., & Latané, B. (1968). Bystander intervention in emergencies: Diffusion of responsibility. Journal of personality and social psychology, 8(4p1), 377.
5. Foo, C. Y., & Koivisto, E. M. (2004, September). Defining grief play in MMORPGs: player and developer perceptions. In Proceedings of the 2004 ACM SIGCHI International Conference on Advances in computer entertainment technology (pp. 245-250). ACM.
6. Ho, S. S., & McLeod, D. M. (2008). Social-psychological influences on opinion expression in face-to-face and computer-mediated communication. Communication Research, 35(2), 190-207.
7. Hollingshead, A. B. (1996). Information suppression and status persistence in group decision making: The effects of communication media. Human Communication Research, 23(2), 193-219.
8. Isenberg, D. J. (1986). Group polarization: A critical review and meta-analysis. Journal of personality and social psychology, 50(6), 1141.
9. Kordyaka, B., Klesel, M., & Jahn, K. (2019, January). Perpetrators in League of Legends: Scale Development and Validation of Toxic Behavior. In Proceedings of the 52nd Hawaii International Conference on System Sciences.
10. Jeong, & Biocca. (2011). Are there optimal levels of arousal to memory? Effects of arousal, centrality, and familiarity on brand memory in video games. *Computers in Human Behavior,28*(2), Computers in Human Behavior.
11. Kwak, H., & Blackburn, J. (2014, November). Linguistic analysis of toxic behavior in an online video game. In International Conference on Social Informatics (pp. 209-217). Springer, Cham.
12. Kwak, H., Blackburn, J., & Han, S. (2015). Exploring Cyberbullying and Other Toxic Behavior in Team Competition Online Games.
13. Mora-Cantallops, & Sicilia. (2018). MOBA games: A literature review. Entertainment Computing, 26, 128-138.
14. Tönnies, Ferdinand. Studien zu Gemeinschaft und Gesellschaft. Springer-Verlag, 2012.
15. Yerkes RM, Dodson JD (1908). "The relation of strength of stimulus to rapidity of habit-formation". Journal of Comparative Neurology and Psychology. 18(5): 459–482. doi:10.1002/cne.920180503
16. Zhang, J., Chang, J. P., Danescu-Niculescu-Mizil, C., Dixon, L., Hua, Y., Thain, N., & Taraborelli, D. (2018). Conversations gone awry: Detecting early signs of conversational failure.