Proceedings of the Eighteenth International AAAI Conference on Web and Social Media (ICWSM 2024)

# Measuring Causal Effects of Civil Communication without Randomization

**Tony Liu**[1,2], **Lyle Ungar**[1], **Konrad Kording**[1], **Morgan McGuire**[2]

[1] University of Pennsylvania
[2] Roblox
liutony@seas.upenn.edu, lyle.ungar@cis.upenn.edu, koerding@gmail.com, morgan@roblox.com

## Abstract

Understanding the causal effects of civility is critical when analyzing online social communication, yet measuring causality is difficult. A/B tests and other randomized experiments are the gold standard for establishing causal effects but they are inapplicable in this setting due to 1) the inability to control civility levels in an experiment, and more importantly, 2) ethical constraints on intentionally randomizing civility levels. We develop a novel *quasi-experimental* approach to quantify the causal effect of civility in online communities on the Roblox social 3D platform without requiring explicit randomization. This method uses residual stochasticity in the "matchmaking" assignment of users to servers as a quasi-randomization mechanism in observational historical data. We find that assigning a user to a server with higher levels of civil communication could increase engagement time by as much as 1.5% in particular experiences. Given the 4.8B person hours spent monthly on the platform, this implies a potential increase of over 8,000 person years of social interaction every month. Furthermore, this effect is mis-estimated by non-causal methods. Quasi-experimental approaches promise new avenues for measuring the causal impact of user behavior in online communities without adversely affecting users through randomized experiments.

## 1 Introduction

Understanding the causal impact of civil communication in online social settings is critical to ensure a functioning community. These settings include text based social networks, multi-party 3D chat environments and video conferencing, and games. Previous work has focused on measuring and detecting uncivil/toxic behavior (Jigsaw 2021; Unitary 2021; Canossa et al. 2021). Though toxicity is an important piece of the equation, focusing solely on it misses the positive aspect of online social communication that constitutes the vast majority of interactions. If one's goal was solely to minimize incivility, then prohibiting *all* interaction would satisfy it, such as when Riot Games disabled /all chat in match-made queues for League of Legends (Riot Games 2021). The goal for supporting an online community must include maximizing civility as well. Moreover, while measurement can demonstrate correlational effects, it does not provide insight into the *causal effect* of such behaviors on the community; e.g., in revealed preferences, does civility actually cause changes in user engagement, and to what extent can fostering civility improve interaction?

For example, we might observe that a higher level of uncivil behavior, such as abusive text communication, is correlated with higher user engagement. Such observed relationships between civility and engagement may be *confounded* by the subject matter of the conversation (in the case of social media) or the game genre (in the case of video games), and may not be generally true – based on previous evidence, we expect that uncivil behavior causally reduces user engagement in most contexts (Fair Play Alliance 2020). Perhaps provoking arguments makes others engage, and a system designed to maximize engagement without context could yield a negative social outcome for the community (Munn 2020) – such as encouraging violent arguments and hate speech! With a causal result showing that a values-metric (civility) drives a goal-metric (engagement), one can then make implementation changes to optimize the values metric and still obtain the goal in a values-aligned manner. Thus, we need to go beyond correlation and quantify the causal impact of civil communication to ensure that online social platforms are optimized for a net positive effect on the community.

However, measuring the causal effects of civil communication in online social settings is challenging. Though the industry gold standard for causal inference is an A/B test, randomized experimentation is often not feasible when analyzing the effects of civility because of both implementation realities and ethical concerns.

The implementation challenge is the lack of control over the interventions of interest. For example, in a typical A/B test one may want to randomize on amount of civil communication a user encounters on the platform, but it is difficult to do so without substantially altering the natural setting of the online community. We cannot force a user to participate in a more civil discussion thread, because such manipulation of the user's actions undermines any finding about the user behavior. That is, active experiments need cleanly defined treatment arms, which are difficult to implement when the treatment of interest is the (civility of) user behavior itself.

The second challenge is that active experimentation on civility is often ethically untenable. One should not intention-

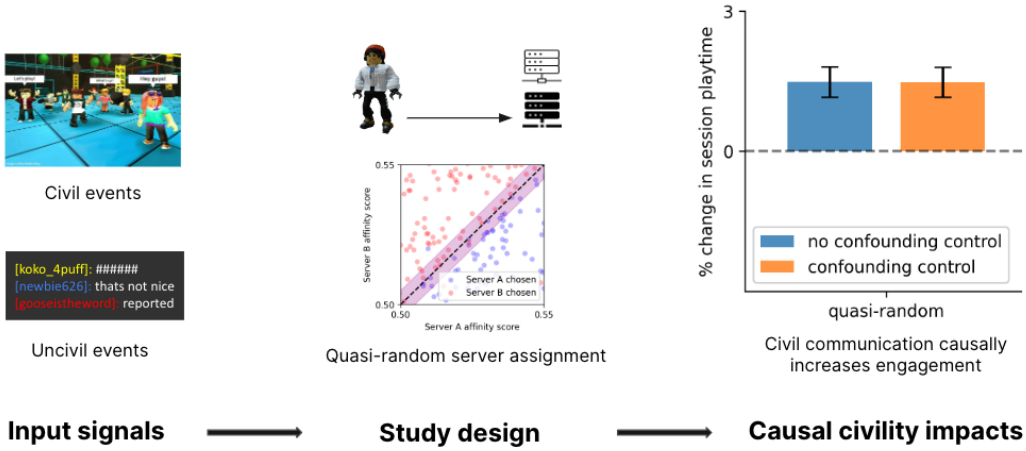



Figure 1: Our study examines the causal effect of civil communication on player engagement through a quasi-experimental design. We take objective input signals to construct measures of communication civility at the server level. We then leverage naturally occurring randomness in the matchmaking process to identify causal effects, allowing us to estimate the downstream impacts of civil communication.

ally expose users to uncivil behavior beyond what would be present in the absence of a study, nor should one deny users exposure to more civil behavior (Kramer, Guillory, and Hancock 2014). We advocate that online social studies adopt the concept of *clinical equipoise* from medicine: an active experiment should not be conducted if the experimenter believes a priori that a significantly beneficial or harmful experimental arm exists. Thus, study strategies that can establish causality of the impacts of online civility ethically, *without* active experimentation on the users, are needed.

When explicit randomization is not an option, there are numerous techniques for measuring causality from observational, historical data (Imbens and Rubin 2015), but many of these methods, such as matching (Falavarjani et al. 2017; Ribeiro, Cheng, and West 2022), rely on the strong assumption of *unconfoundedness*, which stipulates that all relevant variables that affect both the treatment (e.g., civil interactions) and outcome (e.g., subsequent user engagement) are measured. Unconfoundedness is difficult to justify in user behavior studies, where there can be numerous latent attributes that can confound the treatment-outcome relationship of interest (Feder, Riehm, and Mojtabai 2020). For example, online trolls tend to be both highly uncivil and yet quite engaged, and it can be challenging to measure such intrinsic factors that affect the treatment and outcome. However, econometricians have developed a class of study designs called *quasi-experiments* that address confounding by exploiting naturally occurring randomness present within the data. Quasi-experiments allow for credible causal claims without active experimentation (Angrist and Pischke 2009; Liu, Ungar, and Kording 2021), but it is an open question how to bring these approaches to civility studies.

Quasi-experimental methods are relatively underutilized in social media settings (Tian and Chunara 2020) despite having a rich history in economics, education, and epidemiology research (Leamer 1983; Campbell and Stanley 2015; Musci and Stuart 2019). The massive scale of online social media data can mitigate statistical power and sample size concerns, which are significant practical limitations of quasi-experiments (Angrist and Pischke 2009; Lal et al. 2023). Furthermore, in the case of studying the impact of civil communication, quasi-experiments have the key advantage of protecting the user base, which potentially includes minors, from unanticipated impacts by only using historical data from the online platform under normal operating conditions. Quasi-experiments have distinct benefits when studying user behavior in online settings, and should be applied more frequently to better understand the causal impact of civil communication.

Our contributions in this work are as follows:

- We demonstrate the advantages of a quasi-experimental design by conducting a novel study of communication civility on the Roblox platform (Figure 1).
- We use anonymized data from over millions play sessions comprising hundreds of millions of interaction events and construct a privacy-preserving civility metric from text features (Section 3.1).
- By using quasi-randomness present in online server matchmaking (Section 3.3), we are able to establish causal relationships of civil communication and their subsequent impact on engagement.
- We not only show the casual effects of civility on user engagement, but also illustrate pitfalls when performing non-causal analysis (Section 4).
- Our demonstration of a quasi-experimental approach promises to open new avenues for studying civility on online platforms without actively intervening on the user-base (Section 5).

## 2 Background

Here we provide an overview of the Roblox platform as well as a review of prior work on online civility.

## 2.1 The Roblox Platform

Our study focuses on user behavior on Roblox, a free-to-play online 3D social platform. Roblox has 67 million global daily active users interacting in groups of up to 700 people within user-generated 3D *experiences*. These experiences exhibit considerable heterogeneity, ranging from action and platformer games to chat rooms and roleplaying sandboxes. Within these experiences, users interact through global text chat (Figure 2), private chat, voice chat, 3D emote animations, and a social friending system. With 45% of users under the age of 13 (Roblox 2023a), maintaining a safe and civil environment on the platform is critical.

## 2.2 Related Work on Digital Civility

In order to measure civility in online interactions, we need a consistent definition of civil behavior on the platform. There are three broad communities that we draw upon for our study, which we review here to contextualize our approach.

**Digital civility as a social construct.** Social scientists study *digital* civility by examining the conduct of individual's computer-mediated communication, such as through social network sites (SNS). Civil behavior in these settings are often decomposed into behaviors 1) upholding *individual* societal norms of appropriate conduct, such as politeness, and 2) supporting *community* formation and upkeep (Bonotti and Zech 2021; Rowe 2015). Digital behaviors become uncivil when they violate norms of either individual conduct (such as using incendiary language) or community upkeep (such as discrimination or exclusion of some group members). Uncivil behavior in this framing is synonymous with *toxicity*, a term often used in online gaming mediums (Kowert 2020). Most research focuses on the effects of incivility in online discourse and mitigation of uncivil behavior at the individual and community level, rather than the effects and encouragement of more civil behavior (Munn 2020; Anderson et al. 2014; Gervais 2015). In our framework, we aim to map the individual and community components of civility onto our specific digital platform context while also emphasizing markers for positive, civil behavior.

**Game industry guiding principles.** Industry groups have proposed guiding principles on promoting civility or managing incivility in online gaming settings, such as Roblox's educational material on Digital Civility (Roblox 2022a), Electronic Arts Inc. (EA)'s Positive Play Charter (Electronic Arts 2022), and the Fair Play Alliance (FPA)'s Disruption and Harms in Online Gaming Framework (Fair Play Alliance 2020). These guidelines provide best practices that align with the individual ("follow online etiquette," "be the player you want to play with") and community ("show respect for other player's content," "find ways to participate") dimensions of civility. In particular alignment with our goals of civility metric development, the FPA points to a need for measures assessing the quality of both "individual interactions" as well as "community resilience/social cohesion." Indeed, as the bulk of evidence supporting these principles rely on end user surveys that indicate uncivil behavior being most disruptive to gaming experiences (Shi 2019; Fair Play Alliance 2020; Figueiredo 2022), there is a need for further work on how in-game signals contribute to more or less civil online environments. Our study is a first step towards quantifying such finer-grained civility measures by explicitly considering civil and uncivil events in online communities.


Figure 2: Civil and uncivil behavior can be measured through text communication on the platform. Uncivil chat (second speech bubble from the left) is detected through an automated moderation system and filtered out.

**Computer science research on incivility.** Finally, numerous subfields of computer science have studied incivility by primarily focusing on the measurement and detection of negatively valenced behaviors. From a human-computer interaction perspective, research has focused on mitigating uncivil behaviors through platform-level intervention and policy (Chandrasekharan et al. 2017; Jhaver et al. 2021), encouraging user and community self-moderation (Seering, Kraut, and Dabbish 2017), and by examining how platform and UX design promote desirable undesirable behavior (Munn 2020; Seering et al. 2019). Uncivil speech detection can be formulated as supervised machine learning and NLP tasks (Davidson et al. 2017), often with the aim of building automated systems for filtering such unwanted content either tailored to specific platforms (Masud et al. 2022; Nobata et al. 2016) or for general use across contexts (Jigsaw 2021; Lees et al. 2022). There is also a significant body of work on measuring and classifying uncivil behavior beyond textual communication in online gaming contexts (Canossa et al. 2021; Kou 2020). These methods that can measure and classify uncivil content are important to consider, but this same approach needs to be taken with civil signals in order to comprehensively capture civility.

# 3 Methodology

We operationalize a definition of civil communication through measurable signals on the Roblox platform (Section 3.1) and establish our study design (Section 3.2). We then formalize our quasi-experimental strategy tailored to online matchmaking (Section 3.3) that allows us to estimate causal effects without active randomization (Section 3.4).

## 3.1 Measuring Civil Communication

To construct a valenced civility measure that captures both civil and uncivil behavior, we use objective player signals

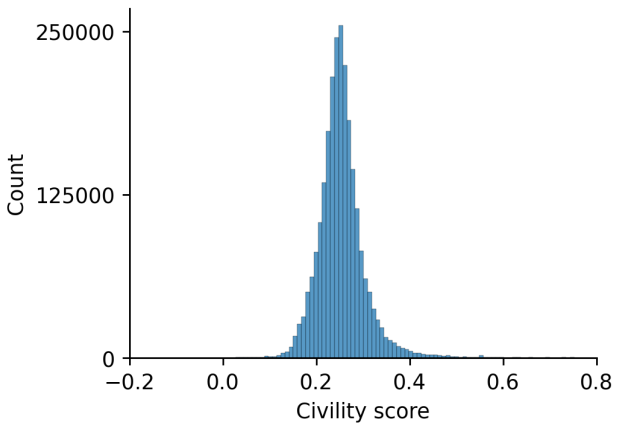



Figure 3: The vast majority of server instances on the platform we analyze have positive communication civility. Communication civility scores shown here are computed over millions of server instances, with over 100 million total communication events.


Figure 4: Words from de-identified empathetic chats sized by their relative frequency.

gathered from public chat text within a Roblox server instance. Our data fit into existing literature defining online social 3D interactions, and our work lies within the *verbal*[1] and *transient* quadrant of user actions (Kowert 2020). Because online civility is a challenging concept to both concretely define and measure directly, we lean on prior literature to determine text-based proxy metrics to best capture the nature of civil communication on a server. We measure civil events through empathy classification of allowed text chat on the platform (Figure 4). The intuition is that, in aggregate, empathetic chats on a server instance are a signal for higher levels of digital civility of the users on the server. In our review of existing work in Section 2.2, digital civility is largely viewed as behaviors that uphold norms of appropriate conduct and community formation. Though empathetic communication does not guarantee such norms are being upheld, we contend that it does provide a signal towards the "sense of belonging" that characterize civil environments (Fair Play Alliance 2020; Electronic Arts 2022; Bonotti and Zech 2021). Beauchere (2019) in particular state that digital civility comprises of "online interactions rooted in empathy, respect, and kindness," which is well-aligned with empathetic communication. In a particularly relevant prior study of online civility, Seering et al. (2019) target empathy in their intervention design to improve the civility of online forums, while numerous studies have establish a relationship between empathy and online pro-social behavior (Corkum and Shead 2023; Mistretta 2021; Prot et al. 2014). We thus propose that empathetic text can act as an objective signal for the constructive, community-building behavior we aim to capture in a civility metric.

To measure negative behavior, we use text violations identified by a custom automated text filter moderation system deployed on the Roblox platform. Though these chat violations are censored and are never displayed to other players (Figure 2, second from left speech bubble), they serve as an objective proxy measure of negative behavior on a particular server instance. Here, it is well-established that such communication can be categorized as uncivil behavior (Gervais 2015; Seering et al. 2019), and is thus an appropriate signal for our civility metric definition.

**Communication civility metric formulation.** We define a communication civility metric $\gamma$ for a given server instance (Equation 1):

$$\gamma := \frac{1}{n}\left(\sum_{m \in A_{\text{civil}}, A_{\text{uncivil}}} m\right) \tag{1}$$

$$\begin{aligned} m &= \text{Individual event civility score} \\ n &= \text{Total number of events} \\ A_{\text{civil}} &= \text{Civil events; varying } 0 \leq m < 1.0 \\ A_{\text{uncivil}} &= \text{Uncivil events; varying } -1.0 < m \leq 0 \end{aligned}$$

A non-filtered chat message weighted by its estimated empathy constitutes a positive event $A_{\text{civil}}$. Specifically, we use a pre-trained empathy (Omitaomu et al. 2022) classifier to predict empathy scores of each line of chat text, normalizing their output to lie within 0 and 1 (see Appendix B.1 for more details). We find that despite not being tuned specifically to the Roblox context, the classifiers qualitatively produce sensible results (Figure 4), with words such as "happy," "feel," and "help" appearing as the top frequency words in chats that are classified as empathetic (Table C.2). Using text classification for civil events produces a much better signal for predicting player engagement than simply using message volume due to the classifiers being able to capture finer grained communication context (Figure B.1).

Conversely, a message filtered by the chat moderation system constitutes a negative event $A_{\text{uncivil}}$ scaled by the moderation category precision, normalized to lie within -1 and 0. We note that the text filter system utilizes an ensemble of methods that include a DistilBERT model trained on toxic-labelled text (Sanh et al. 2019) and tuned over years of deployment, achieving upwards of 0.8 precision – thus

[1] Kowert's "verbal" is language and text; i.e., not behavior.

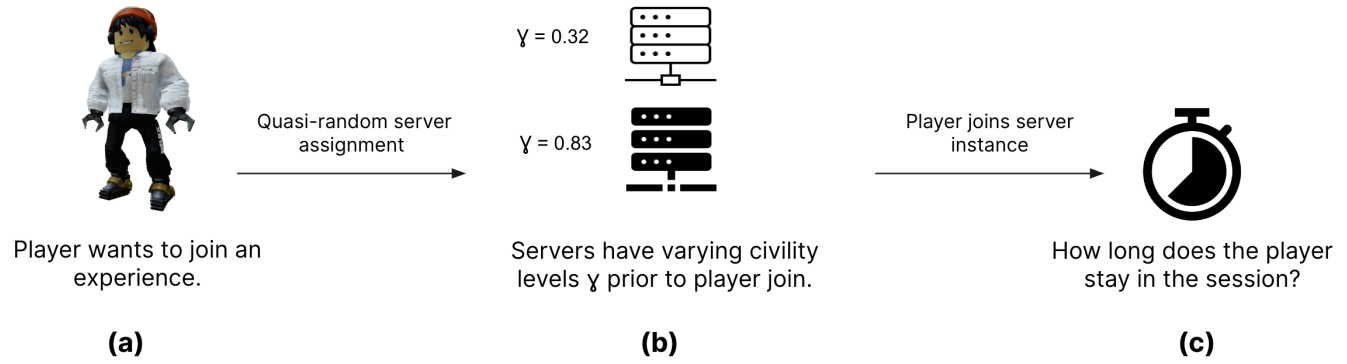



Figure 5: We leverage naturally occurring randomness from player-server matchmaking to measure the effects of civil communication. When a user chooses an experience to join (a), quasi-random server assignment that will vary the level of communication civility a player experiences (b), which allows for retroactive analysis of the causal impact of civility on engagement (c).

it is a reliable indicator of uncivil text. Overall, the metric encompasses both positive and negative behaviors, where higher values indicate higher rates of communication civility on a server, while lower values indicate less communication civility. By and large, most interactions on the platform can be considered civil (Figure 3). We note personally identifying information about users are sanitized from the source signals prior to the beginning of our analysis, ensuring data privacy. Our metric $\gamma$ provides an interpretable and anonymized measure of the nature of communication civility in a given server instance on the platform.

### 3.2 Study Design

In order to cleanly measure the effects of communication civility on user behavior, we consider the process of players being matched to Roblox server instances, where they experience varying levels of communication civility (Figure 5). The data we analyze are collected from a two week period in February 2023, with over 50 million play sessions considered in the initial dataset. We randomly sample to ensure that any individual player is only represented once. To account for the heterogeneity in experience mechanics and gameplay types, we control for the historical average session length for each individual experience in our analysis.

When users want to join a Roblox experience, such as a video game or chat room, there can be multiple candidate server instances of their chosen experience (Figure 5a). Players enter a matchmaking process where a backend algorithm determines the best server for them to join. The player-server matchmaking algorithm computes a continuous "affinity" score $s$ based on factors such as the number of players already on a candidate server and the network latency between the player and the servers. The algorithm then ranks candidate servers and assigns players to the server with the highest score.

Though most servers are not randomly assigned due to these preferred characteristics being clearly satisfied by a best-choice server, because there is stochasticity in the input signals to this algorithm (e.g., player occupancy and network latency can fluctuate in real time) as well as a hard thresholded decision of selecting the server with the best score, servers that are "barely chosen" and "nearly chosen" can be viewed as undergoing quasi-random assignment. We think of these barely chosen vs. nearly chosen servers as "almost coin flip" decisions where the server instance a player is placed on is independent of the factors that are considered in the matchmaking algorithm. Because of this *quasi-random assignment* (Titiunik 2021), the levels of civility that they experience on those servers will be plausibly random as well.

We define the communication civility of a server instance by its $\gamma$ in a 10 minute window *prior* to the player joining in order to prevent the joined player's behavior contaminating the server's civility levels (Figure 5b). We believe that this serves as a reliable measure of communication civility the joining player subsequently observes as chat history is preserved and because the other players who emitted the previous chat lines are likely to still be on the server. We then measure the player's subsequent play session time on the server instance, which serves as an indicator of their engagement with the experience (Figure 5c). By analyzing a selected data sample that satisfies this quasi-random assignment mechanism, we can measure the causal impact of civil communication on player session length from historical data under the platform's normal operating conditions.

### 3.3 Quasi-Random Server Pairs

To select a sample of servers that satisfy the quasi-random assignment mechanism, we need to determine a *window* region where the scores between a pair of candidate servers are statistically indistinguishable from one another. We would like this window region to be as large as possible while maintaining quasi-random assignment in order to maximize the sample size of our study. Given pairs $(A, B)$ of the top two candidate servers and their corresponding affinity scores $s$ for a player $(s_A, s_B)$, we want to find the largest window length $w$ such that we cannot reject the null hypothesis that $s_A = s_B$ for the set $\{(s_A, s_B) \mid s_A - s_B \leq w\}$, visualized

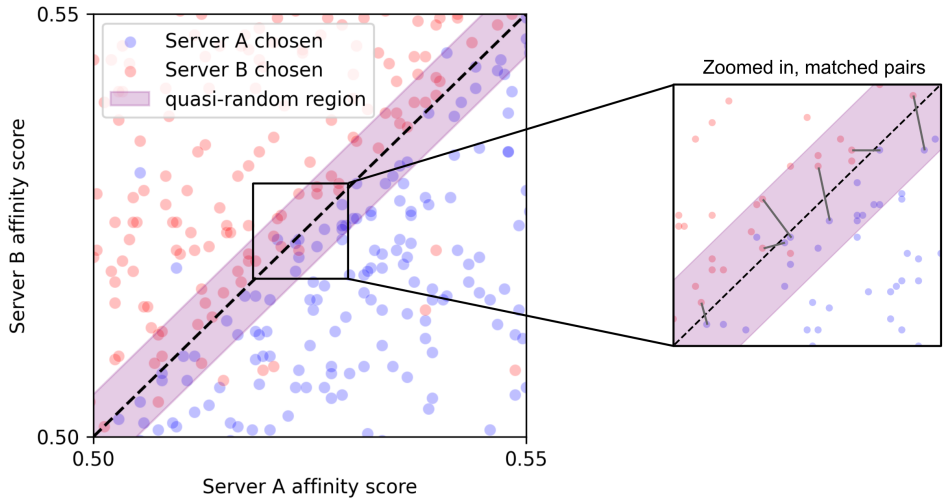



Figure 6: We leverage differences in the computed server affinity scores for players for quasi-randomization of civility exposure. By selecting pairs of chosen servers that are statistically indistinguishable from each other (gray pairs within purple region, right), we are able to estimate causal effects through the difference in outcomes for players that are assigned to them.

as the purple region in Figure 6.

Determining $w$ is akin to the window selection procedure in *regression discontinuities* (RDs), a particular class of quasi-experiments where treatment assignment is determined by a hard threshold of the score (Imbens and Lemieux 2008). Intuitively, scores that are close to the threshold can be thought of as-if randomly assigned, where scores just below the threshold are not treated and scores just above the threshold are treated, analogous to our "nearly chosen" and "barely chosen" servers. And in a similar fashion, a window around the threshold must be selected in order to implement an RD study. We therefore utilize the same testing procedure for selecting $w$ as recommended by the RD literature (Cattaneo, Frandsen, and Titiunik 2015), where successively smaller $w$ sizes are tested for whether the null hypothesis $s_A = s_B$ fails to be rejected. We follow Cattaneo, Frandsen, and Titiunik (2015)'s testing procedure and determine $w = 0.001$ for our data (see Appendix B.2 for details). This test-driven procedure for selecting $w$ ensures that the sample we analyze satisfy quasi-random server assignment, allowing us to draw causal conclusions on the effects of being placed on these servers.

However, in order to fully measure the causal effects of communication civility experienced when being assigned a server, we have to ensure that both servers in the quasi-randomized pair are represented in the sample. This is due to the "almost coin flip" decisions only occurring between a specific pair of servers for a given player and experience choice; we need to not only measure the "barely chosen" server $A$ but also the "nearly chosen" server $B$ to generate a "counterfactual" outcome for the pair i.e., what would have happened if the player were placed on server $B$ instead of server $A$? The challenge is that because only server $A$ was chosen, we never directly observe the effects of placing the player on server $B$; this is often known as the *fundamental problem of causal inference* (Imbens and Rubin 2015). We resolve this challenge in our causal strategy by also selecting players who have the reverse decisions where server $B$ was the "barely chosen" server, resulting in pairs of quasi-randomized server choices (Figure 6, gray pairs within zoomed panel). This unique paired design ensures that the civility experienced on each server is plausibly random and that outcomes can be measured for both the servers.

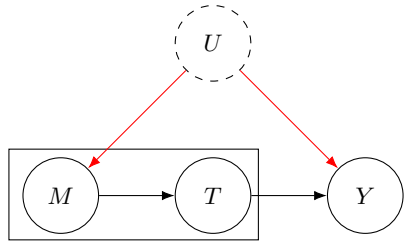
### 3.4 Causal Effect Estimation

To formalize our quasi-experimental strategy of analyzing "near coin flip" servers for effect estimation, we introduce the following *potential outcomes* notation (for a comprehensive review, see e.g., Imbens and Rubin (2015) or Hernán and Robins (2023)):

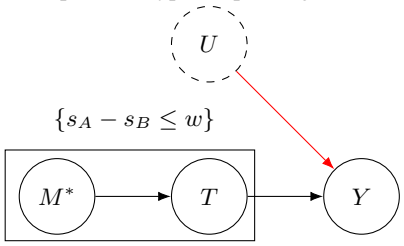
$$
\begin{aligned}
Y &= \text{outcome of interest: session playtime} \\
T &= \text{treatment: civil server placement} \\
Y(t) &= \text{potential outcome } t \\
U &= \text{confounder} \\
M^* &= \text{binary indicator of unit in matched server pair}
\end{aligned}
$$

Our outcome of interest $Y$ is the length of a player's session on a specific server. We log transform $Y$ for a more meaningful interpretation of the causal effect as a percentage change in session length (Figure B.3).

We define our treatment $T$ of interest as a dichotomization of the communication civility metric $\gamma$, where $T = 1$ represents being placed on a "more civil" server. $\gamma$ is dichotomized at a q50 level, where servers with civility scores above the 0.5 quantile are considered $T = 1$ while servers below the quantile are considered $T = 0$. As a robustness check, we perform causal effect estimation testing different cutoff levels and find the conclusions of our analysis remain consistent (Figure B.2). We note that this dichotomization necessarily results in some loss of information, but we make this methodological decision due to the modeling




(a) Graph under typical operating conditions.




(b) Graph for quasi-random pairs.

Figure 7: Graphical illustration of our quasi-random server pair causal inference strategy. Under typical circumstances (a), the matchmaking decision $M$ for placing a user on a server will likely be confounded by $U$ (red lines), with dotted circle around $U$ indicating that confounders may be unobserved. In our design (b), the quasi-randomized matched pairs from matchmaking decision $M^*$ (where the affinity score differences $s_A - s_B$ are less than $w$) between servers of varying civility exposure $T$ breaks any confounding links, allowing us to measure the causal relationship between civility and engagement. We draw a box around $M$ and $T$ as in practice they are coupled together – we can think of them acting as a single "node," where the matchmaking and server placement under $M$ is confounded while the matchmaking and server placement under $M^*$ is unconfounded.

advantages of defining a binary treatment in causal inference (Fong, Hazlett, and Imai 2018). Furthermore, though the metric exists on a continuous scale, dichotomization reflects the realities of decision-making as a discrete action (e.g. a player joining or not joining a server, or in medical contexts, treating or not treating a patient), and thus more appropriately informs *policies*. We leave further analysis of the continuous civility metric and policy design improvement as future work.

We then define *potential outcomes* $Y(t)$ as the player session lengths that would have been observed under a particular treatment assignment $t$ e.g., $Y(1)$ is the session length if a player were to be placed on a civil server. Since a player only undergoes a single treatment assignment, only one of these potential outcomes is observed. The goal is to estimate a causal *average treatment effect* $\tau$ of being placed on a civil server, conditioning on only quasi-random server pairs:

$$\tau := E[Y(1) - Y(0)|M^* = 1] \tag{2}$$

The unconfoundedness assumption (also known as *ignorability* or *exchangeability*) is often needed to *identify* causal effects such as $\tau$ i.e., convert causal quantities to statistically estimable quantities (Imbens and Rubin 2015):

$$Y(1), Y(0) \perp T \tag{3}$$

As we discuss in Section 1, unconfoundedness is typically hard to justify, but our use of quasi-random server assignment to construct our study provides a strong case for why the assumption could plausibly hold. In particular, because units where $M^* = 1$ are quasi-randomized within the server score window, $T$ can be treated as independent from a player's potential outcomes. We can think of $M^*$ being an *adjustment* criterion allowing for conditional unconfoundedness (Neal 2020):

$$Y(1), Y(0) \perp T \mid M^* \tag{4}$$

Our quasi-random paired server assignment mechanism can be seen as a robust form of a *matching* strategy (Hernán and Robins 2010; Mansournia, Hernán, and Greenland 2013). By construction of the matched pairs, experience and user characteristics are necessarily balanced between the paired servers (Figure 8). We can thus use the conditional unconfoundedness in Equation 4 to estimate $\tau$ as a conditional difference in means (see Appendix B.3):

$$\tau = E[Y|T = 1, M^* = 1] - E[Y|T = 0, M^* = 1] \tag{5}$$

Concretely, this quantity is estimated by performing a difference-in-means regression (Angrist and Pischke 2009; Wooldridge 2010) on pairs of servers with $M^* = 1$ and different treatment statuses ($T = 1$ and $T = 0$). The quasi-experimental effect estimation can also be represented graphically (Steiner et al. 2017), where quasi-randomized server pairs breaks any incoming links to confounders, regardless of whether the confounders are measured or observed (Figure 7b). Through our study strategy, we are able to quantify the causal effect communication civility has on player engagement within our study sample.

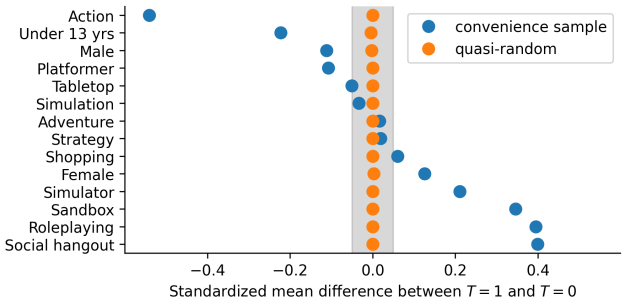



Figure 8: Our quasi-randomized server matching (orange) ensures balance among server and user characteristics. We show a "Love" plot (Ahmed et al. 2006) to visually assess covariate balance between the $T = 1$ and $T = 0$ groups. Ideally, the standardized mean differences between the two groups should fall within a bandwidth of 0.05 around zero (shaded region) (Ho et al. 2007), which the quasi-randomized pairs achieve for all shown covariates but a random convenience sample (blue) does not.

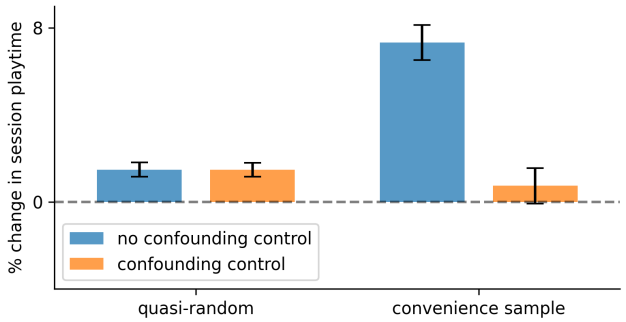



Figure 9: Separating correlation from causation is critical as confounders can bias the relationship between civility measures and user engagement. A randomly drawn convenience sample produces starkly different effect estimates with/without controlling for confounders, while the quasi-random estimates remain stable after controlling for confounders. Error bars are 95% confidence intervals, and full regression details can be found in Appendix B.4.

## 4 Results

To summarize our study strategy, our final sample consists of server pairs where the respective matchmaking decisions are quasi-randomized due to the server affinity score being statistically indistinguishable from one another. We then utilize a regression framework (Angrist and Pischke 2009) to estimate the causal effects of communication civility on player session length.

**Civil communication causally increases engagement.** We find that being placed on a civil server based on our dichotomization increases a player's session length by 1.5% (CIs $[1.2, 1.8]$, $p < 0.0001$), shown on the left in Figure 9. To contextualize the size of this effect, given the 4.8 billion person hours spent monthly on the platform (Roblox 2023b), a 1.5% increase could yield over 8,000 person years of increased social interaction every month. By providing causal evidence of the link between civility and player engagement, we can better inform policy that could proactively improve the online community.

**Non-causal analyses produce inconsistent results, while causal analyses are stable.** Furthermore, we demonstrate the pitfalls of using observational data without a study design that considers causality. As a concrete example, we perform the same statistical analysis utilizing a convenience sample randomly drawn from all play sessions instead of only quasi-randomized server assignments and find an instability in the estimated effect. The correlational estimate without controlling for confounding shows a positive relationship between civil communication and user playtime (7.3%, CIs $[6.5, 8.1]$, $p < 0.0001$) (Figure 9, right). However, once we control for the potential confounding factors of age, gender, and experience genre, the estimated effects for our convenience sample drastically shift (0.7%, CIs $[-0.001, 0.016]$, $p > 0.05$), indicating that confounding factors are likely present.

Critically, the results from our quasi-random sample remain stable when controlling for potential confounding factors, (1.5% CIs $[1.2, 1.8]$, $p < 0.0001$), with confidence intervals that almost identically match the previous interval without controls (Figure 9, left in orange). This provides evidence that the treatment assignment is plausibly randomized. Analyses using observational data that do not account for confounding risk potentially erroneous conclusions, underscoring the need for causal study design.

## 5 Discussion

In this work, we examine civility in online peer interactions and quantify its causal impact on user behavior through a novel quasi-experimental design. We now discuss limitations alongside future work and generalizable insights.

### 5.1 Limitations and Future Work

We recognize that there are generalizability constraints of our causal results due to the quasi-experimental framing (a tradeoff for not needing to perform active interventions) as well as aspects of user behavior that our civility metric does not encompass.

First, we note that because we are not conducting an active experiment, we cannot control our study population of interest. It is possible that there are Roblox experiences that rarely have quasi-random server assignment e.g., experiences that are single-player or are only ever played with friends, and so our analysis would not capture nor apply to these cases. Indeed, there is evidence that our matched server pair sample that allows for quasi-random analysis differs in baseline characteristics from other server experiences (see Table C.3). As with any causal analysis, it is important to consider both the *internal* validity of the study (do we believe the conclusions drawn from the data sample?) as well as the *external* validity (to what extent do the conclusions drawn apply to other populations?). We also note that interference on social network platforms (e.g., friend and peer network effects) is a concern that could influence our causal conclusions. An important shortcoming of most quasi-experimental studies is that though they have greater internal validity than typical observational studies, their external validity can be quite limited.

In particular, our design has limitations in interpretability analogous to the limitations of regression discontinuities we describe in Section 3.3. Much like how regression discontinuity estimates only apply locally around the threshold of interest, our causal estimates will only apply to those units that are a part of those "coin flip" matchmaking decisions. Future work can explicitly address external validity concerns (Liu et al. 2022; Wu et al. 2022) for our quasi-experimental design in order to identify who these effects apply to and the relative magnitude of the effects.

We also acknowledge that our communication civility metric can be further refined. Here we relied on two measures when quantifying civility: 1) signal from existing online incivility filtering metrics that is robust, due to its extensive tuning during actual deployment, and 2) an off-the-shelf model of empathetic interactions that do not benefit from the

same level of context on the platform. Important future work is bringing the positive model to the same level of robustness as the negative model. This will not change our core contribution's analysis framework and civility definition, but will increase its precision and applicability.

There are also broader opportunities for improving the measurement of civility in online communication settings. Though we choose to rely on text classifiers to provide a more empirical grounding in the definition of civil and uncivil events, in order to better understand the nature of civility on specific platforms, future measurement efforts should couple this classification with user surveys and crowdsourced annotations of communication events. Furthermore, we acknowledge that the dichotomization of communication civility that facilitates our quasi-experimental strategy loses information when compared to the continuous metric. Future work could formulate study designs that allow for a causal effect interpretation of the continuous metric through applying methodologies like those presented in Fong, Hazlett, and Imai (2018) and Kennedy, Lorch, and Small (2019).

A particularly fruitful direction for future work would be to go beyond effect measurement to explore policy decisions such as deploying nudge interventions or presenting educational material that encourage players to engage in more socially constructive and prosocial behavior (Lin 2013, 2015; Jones, Mitchell, and Beseler 2023). By better understanding the relationship between civil communication and player behavior, we could design mechanisms within online communities that not only maintain safety and civility but also organically improve peer interaction.

### 5.2 Generalizable Insights

We highlight study process and methodological insights that can extend to the wider social media academic community. Experiments are still needed to establish causal links with the greatest degree of certainty, but quasi-experiments fill a middle ground between active experimentation and purely observational analyses.

In particular, the phenomenon we identify with "near coin flip" servers producing quasi-randomization can be adapted to other online platforms with decisions based on score thresholds, which could be used as supporting analysis ahead of a more expensive A/B test. For example, online video games that match players to lobbies based on a similarity score could leverage an analogous design when there is a toss-up between two lobbies for a pair of players – this could allow for causal analysis comparing properties of the lobbies. Another situation where this strategy may be applied would be "recommended" product placements in online shopping contexts. If there are a set number of recommended products visually displayed to a user based on a score, one could perform causal evaluation on e.g., click-through rates by comparing products that just barely made the cutoff for the recommended list against ones that barely missed the cutoff. Though we develop our paired server approach for our specific data domain, the overall strategy can still be leveraged where pairs of units are placed in different "treatment" categories based on the threshold of a continuous score.

Furthermore, we note that there are there numerous other quasi-experimental designs available, such as *instrumental variables*, *regression discontinuities*, and *differences-in-differences*, that can be used to establish causality in a variety of data contexts, including longitudinal data. We refer the readers to Angrist and Pischke (2009) as a reference and advocate for familiarity with these designs in order to apply them in their own work.

However, we want to emphasize that though we believe that quasi-experimental designs should be used more frequently where possible, they should not be thought of as replacements to active experiments. Rather, they are supplemental tools in a scientist's toolkit that should be utilized whenever possible because of their lower cost relative to randomized experiments and greater causal validity relative to correlational studies.

We hope that our methodological approach encourages more usage of quasi-experimental methods within the community to provide more credible causal insights when experimentation is difficult.

## 6 Conclusion

Here we have conducted a quasi-experimental study on the effects of civil communication on the Roblox platform, finding that civil communication causally increases player engagement. Critically, our quasi-experimental design only utilized passively collected historical data, allowing us to study civility without actively intervening on the user base. We advocate for the increased use of quasi-experiments in online social media user studies to complement existing methodologies, as these approaches promise to inform future interventions in large-scale online communities that can not only reduce undesirable behavior but also encourage socially constructive behavior.

## Broader Perspective and Ethics

We protected the personal information of the studied Roblox community (including minors) by designing the civility metric to be privacy preserving. We only report aggregated statistics publicly and do not publish personally identifying information from our analysis. The motivation and largest contribution of our work is establishing an ethical causal framework (via quasi-experiments, under the equipoise principle) for civility research. The broader perspective served by our analysis is studying the causal link between engagement and civility, which can help inform the design of value-aligned social platforms.

## Acknowledgments

We would like to thank bc Wong, Shyna Khurana, Henry Lin, Ujwal Kharel, and Colin Dillard for their insights and discussion throughout the development of this project. We also thank the anonymous reviewers whose feedback helped greatly improve our work.

## References


Ahmed, A.; Husain, A.; Love, T. E.; Gambassi, G.; Dell'Italia, L. J.; Francis, G. S.; Gheorghiade, M.; Allman, R. M.; Meleth, S.; and Bourge, R. C. 2006. Heart failure, chronic diuretic use, and increase in mortality and hospitalization: an observational study using propensity score methods. *European heart journal*, 27(12): 1431–1439.

Anderson, A. A.; Brossard, D.; Scheufele, D. A.; Xenos, M. A.; and Ladwig, P. 2014. The "Nasty Effect:" Online Incivility and Risk Perceptions of Emerging Technologies*. *Journal of Computer-Mediated Communication*, 19(3): 373–387.

Angrist, J. D.; and Pischke, J.-S. 2009. *Mostly harmless econometrics: An empiricist's companion*. Princeton university press.

Beauchere, J. F. 2019. Encouraging Digital Civility: What Companies and Others Can Do. In *Internet and Technology Addiction: Breakthroughs in Research and Practice*, 748–759. IGI Global. ISBN 978-1-5225-8900-6.

Bonotti, M.; and Zech, S. T. 2021. Understanding Civility. In Bonotti, M.; and Zech, S. T., eds., *Recovering Civility during COVID-19*, 37–64. Singapore: Springer. ISBN 978-981-336-706-7.

Campbell, D. T.; and Stanley, J. C. 2015. *Experimental and quasi-experimental designs for research*. Ravenio books.

Canossa, A.; Salimov, D.; Azadvar, A.; Harteveld, C.; and Yannakakis, G. 2021. For Honor, for Toxicity: Detecting Toxic Behavior through Gameplay. *Proceedings of the ACM on Human-Computer Interaction*, 5(CHI PLAY): 1–29.

Cattaneo, M. D.; Frandsen, B. R.; and Titiunik, R. 2015. Randomization inference in the regression discontinuity design: An application to party advantages in the US Senate. *Journal of Causal Inference*, 3(1): 1–24.

Chandrasekharan, E.; Pavalanathan, U.; Srinivasan, A.; Glynn, A.; Eisenstein, J.; and Gilbert, E. 2017. You Can't Stay Here: The Efficacy of Reddit's 2015 Ban Examined Through Hate Speech. *Proceedings of the ACM on Human-Computer Interaction*, 1(CSCW): 31:1–31:22.

Corkum, M.; and Shead, N. W. 2023. Online Moral Disengagement: An Examination of the Relationships Between Electronic Communication, Cognitive Empathy, and Antisocial Behavior on the Internet. *Psychological Reports*, 00332941231216415.

Davidson, T.; Warmsley, D.; Macy, M.; and Weber, I. 2017. Automated Hate Speech Detection and the Problem of Offensive Language. *Proceedings of the International AAAI Conference on Web and Social Media*, 11(1): 512–515. Number: 1.

Electronic Arts. 2022. Positive Play Charter - Electronic Arts Official Site.

Fair Play Alliance. 2020. Disruption and Harms in Online Gaming Framework.

Falavarjani, S. M.; Hosseini, H.; Noorian, Z.; and Bagheri, E. 2017. Estimating the effect of exercising on users' online behavior. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 11, 734–738.

Feder, K. A.; Riehm, K. E.; and Mojtabai, R. 2020. Is there an association between social media use and mental health? the timing of confounding measurement matters—reply. *JAMA psychiatry*, 77(4): 438–438.

Figueiredo, C. 2022. Trust and Safety and Fair Play in Video Games: Intentionally Designing Positive Communities. In *Game Usability*. CRC Press, second edition. ISBN 978-1-00-310938-9.

Fong, C.; Hazlett, C.; and Imai, K. 2018. Covariate balancing propensity score for a continuous treatment: Application to the efficacy of political advertisements. *The Annals of Applied Statistics*, 12(1): 156–177.

Gervais, B. T. 2015. Incivility Online: Affective and Behavioral Reactions to Uncivil Political Posts in a Web-based Experiment. *Journal of Information Technology & Politics*, 12(2): 167–185. Publisher: Routledge _eprint: https://doi.org/10.1080/19331681.2014.997416.

Hernán, M. A.; and Robins, J. M. 2010. Causal inference.

Hernán, M. A.; and Robins, J. M. 2023. *Causal Inference: What If*. Boca Raton: Chapman & Call/CRC.

Ho, D. E.; Imai, K.; King, G.; and Stuart, E. A. 2007. Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference. *Political analysis*, 15(3): 199–236.

Imbens, G.; and Rubin, D. B. 2015. *Causal inference: for statistics, social and biomedical sciences : an introduction*. ISBN 978-1-139-02575-1. OCLC: 985493948.

Imbens, G. W.; and Lemieux, T. 2008. Regression discontinuity designs: A guide to practice. *Journal of econometrics*, 142(2): 615–635.

Jhaver, S.; Boylston, C.; Yang, D.; and Bruckman, A. 2021. Evaluating the Effectiveness of Deplatforming as a Moderation Strategy on Twitter. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW2): 381:1–381:30.

Jigsaw. 2021. Perspective API. https://www.perspectiveapi.com/.

Jones, L. M.; Mitchell, K. J.; and Beseler, C. L. 2023. The impact of youth digital citizenship education: Insights from a cluster randomized controlled trial outcome evaluation of the be internet awesome (BIA) curriculum. *Contemporary School Psychology*, 1–15.

Kennedy, E. H.; Lorch, S.; and Small, D. S. 2019. Robust causal inference with continuous instruments using the local instrumental variable curve. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 81(1): 121–143.

Kou, Y. 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. Virtual Event Canada: ACM. ISBN 978-1-4503-8074-4.

Kowert, R. 2020. Dark Participation in Games. *Frontiers in Psychology*, 11.

Kramer, A. D. I.; Guillory, J. E.; and Hancock, J. T. 2014. Experimental Evidence of Massive-Scale Emotional Contagion through Social Networks. *Proceedings of the National Academy of Sciences*, 111(24): 8788–8790.

Lal, A.; Lockhart, M.; Xu, Y.; and Zu, Z. 2023. How much should we trust instrumental variable estimates in political science? Practical advice based on over 60 replicated studies. *arXiv preprint arXiv:2303.11399*.

Leamer, E. E. 1983. Let's Take the Con Out of Econometrics. *The American Economic Review*, 73(1): 31–43.

Lees, A.; Tran, V. Q.; Tay, Y.; Sorensen, J.; Gupta, J.; Metzler, D.; and Vasserman, L. 2022. A New Generation of Perspective API: Efficient Multilingual Character-level Transformers. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 3197–3207. Washington DC USA: ACM. ISBN 978-1-4503-9385-0.

Lin, J. 2013. The Science Behind Shaping Player Behavior in Online Games.

Lin, J. 2015. More Science Behind Shaping Player Behavior in Online Games.

Liu, T.; Lawlor, P.; Ungar, L.; and Kording, K. 2022. Data-driven exclusion criteria for instrumental variable studies. In *Conference on Causal Learning and Reasoning*, 485–508. PMLR.

Liu, T.; Ungar, L.; and Kording, K. 2021. Quantifying Causality in Data Science with Quasi-Experiments. *Nature Computational Science*, 1(1): 24–32.

Mansournia, M. A.; Hernán, M. A.; and Greenland, S. 2013. Matched designs and causal diagrams. *International journal of epidemiology*, 42(3): 860–869.

Masud, S.; Bedi, M.; Khan, M. A.; Akhtar, M. S.; and Chakraborty, T. 2022. Proactively Reducing the Hate Intensity of Online Posts via Hate Speech Normalization. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '22, 3524–3534. New York, NY, USA: Association for Computing Machinery. ISBN 978-1-4503-9385-0.

Mistretta, S. 2021. The new netiquette: Choosing civility in an age of online teaching and learning. In *International Journal on E-Learning*, 323–345. Association for the Advancement of Computing in Education (AACE).

Munn, L. 2020. Angry by design: toxic communication and technical architectures. *Humanities and Social Sciences Communications*, 7(1): 1–11. Number: 1 Publisher: Palgrave.

Musci, R. J.; and Stuart, E. 2019. Ensuring Causal, Not Casual, Inference. *Prevention science : the official journal of the Society for Prevention Research*, 20(3): 452–456.

Neal, B. 2020. Introduction to causal inference. *Course Lecture Notes (draft)*.

Nobata, C.; Tetreault, J.; Thomas, A.; Mehdad, Y.; and Chang, Y. 2016. Abusive Language Detection in Online User Content. In *Proceedings of the 25th International Conference on World Wide Web*, WWW '16, 145–153. Republic and Canton of Geneva, CHE: International World Wide Web Conferences Steering Committee.

Omitaomu, D.; Tafreshi, S.; Liu, T.; Buechel, S.; Callison-Burch, C.; Eichstaedt, J.; Ungar, L.; and Sedoc, J. 2022. Empathic conversations: A multi-level dataset of contextualized conversations. *arXiv preprint arXiv:2205.12698*.

Prot, S.; Gentile, D. A.; Anderson, C. A.; Suzuki, K.; Swing, E.; Lim, K. M.; Horiuchi, Y.; Jelic, M.; Krahé, B.; Liuqing, W.; et al. 2014. Long-term relations among prosocial-media use, empathy, and prosocial behavior. *Psychological science*, 25(2): 358–368.

Ribeiro, M. H.; Cheng, J.; and West, R. 2022. Post Approvals in Online Communities. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 16, 335–346.

Riot Games. 2021. Disabling /All Chat - League of Legends. https://www.leagueoflegends.com/en-us/news/game-updates/disabling-all-chat/.

Roblox. 2022a. Intro to Digital Civility. https://create.roblox.com/docs/education/resources/intro-to-digital-civility. Accessed 2022-12-22.

Roblox. 2022b. Roblox Privacy Policy 2022-12-31. https://en.help.roblox.com/hc/en-us/articles/115004630823-Roblox-Privacy-and-Cookie-Policy. Accessed: 2023-02-01.

Roblox. 2023a. Roblox Corporation 10-K 2022. https://d18rn0p25nwr6d.cloudfront.net/CIK-0001315098/007d50a0-89bf-486d-9850-1eeb50827af0.pdf. Accessed: 2023-09-08.

Roblox. 2023b. Roblox Reports March 2023 Key Metrics. https://ir.roblox.com/news/news-details/2023/Roblox-Reports-March-2023-Key-Metrics/. Accessed: 2023-05-13.

Rowe, I. 2015. Civility 2.0: a comparative analysis of incivility in online political discussion. *Information, Communication & Society*, 18(2): 121–138. Publisher: Routledge _eprint: https://doi.org/10.1080/1369118X.2014.940365.

Sanh, V.; Debut, L.; Chaumond, J.; and Wolf, T. 2019. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Seering, J.; Fang, T.; Damasco, L.; Chen, M. C.; Sun, L.; and Kaufman, G. 2019. Designing User Interface Elements to Improve the Quality and Civility of Discourse in Online Commenting Behaviors. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 1–14. Glasgow Scotland Uk: ACM. ISBN 978-1-4503-5970-2.

Seering, J.; Kraut, R.; and Dabbish, L. 2017. Shaping Pro and Anti-Social Behavior on Twitch Through Moderation and Example-Setting. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*, CSCW '17, 111–125. New York, NY, USA: Association for Computing Machinery. ISBN 978-1-4503-4335-0.

Shi, J. 2019. What Inclusion Means to Players. *Medium*.

Steiner, P. M.; Kim, Y.; Hall, C. E.; and Su, D. 2017. Graphical models for quasi-experimental designs. *Sociological methods & research*, 46(2): 155–188.

Tian, Y.; and Chunara, R. 2020. Quasi-experimental designs for assessing response on social media to policy changes. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 14, 671–682.

Titiunik, R. 2021. Natural Experiments. In Green, D. P.; and Druckman, J. N., eds., *Advances in Experimental Political*

*Science*, 103–129. Cambridge: Cambridge University Press. ISBN 978-1-108-47850-2.

Unitary. 2021. Unitary.Ai. *Unitary - Specialists in visual content moderation.*

Wooldridge, J. M. 2010. *Econometric analysis of cross section and panel data*. MIT press.

Wu, H.; Tan, S.; Li, W.; Garrard, M.; Obeng, A.; Dimmery, D.; Singh, S.; Wang, H.; Jiang, D.; and Bakshy, E. 2022. Interpretable Personalized Experimentation. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '22, 4173–4183. New York, NY, USA: Association for Computing Machinery. ISBN 978-1-4503-9385-0.

## A Ethics Checklist

1. For most authors...
   (a) Would answering this research question advance science without violating social contracts, such as violating privacy norms, perpetuating unfair profiling, exacerbating the socio-economic divide, or implying disrespect to societies or cultures? Yes, the motivation and largest contribution of our work is establishing an ethical framework (via quasi-experiments, under the equipose principle) for civility research that enables causal estimation *without* active experimentation. The intent is exactly in line with the goal of this question.
   (b) Do your main claims in the abstract and introduction accurately reflect the paper's contributions and scope? Yes, we discuss the scope of our claims as well as limitations to the generalizability of our study conclusions in Section 5.
   (c) Do you clarify how the proposed methodological approach is appropriate for the claims made? Yes, we justify our observational causal methodology in Section 1 and elaborate on the methodological details for causal identification in Section 3.4.
   (d) Do you clarify what are possible artifacts in the data used, given population-specific distributions? Yes, we provide population-level summaries of the observational data we use to show how they might differ from convenience samples, and discuss generalizability limits in Section 5.1.
   (e) Did you describe the limitations of your work? Yes, see Section 5.1.
   (f) Did you discuss any potential negative societal impacts of your work? No; in any likely application it is inherently neutral or positive.
   (g) Did you discuss any potential misuse of your work? No; there is no obvious misuse potential.
   (h) Did you describe steps taken to prevent or mitigate potential negative outcomes of the research, such as data and model documentation, data anonymization, responsible release, access control, and the reproducibility of findings? Yes, we discuss data anonymization and construction of a privacy-preserving civility metric in Section 3.1.
   (i) Have you read the ethics review guidelines and ensured that your paper conforms to them? Yes.
2. Additionally, if your study involves hypotheses testing...
   (a) Did you clearly state the assumptions underlying all theoretical results? Yes, we discuss the assumptions needed for causal identification in Section 3.4 and Appendix B.3.
   (b) Have you provided justifications for all theoretical results? Yes, we justify our causal identification mechanism in Sections 3.3-3.4.
   (c) Did you discuss competing hypotheses or theories that might challenge or complement your theoretical results? NA
   (d) Have you considered alternative mechanisms or explanations that might account for the same outcomes observed in your study? Yes, we discuss threats to causal validity that may influence the outcomes observed in our study in Section 5 and Appendix B.3.
   (e) Did you address potential biases or limitations in your theoretical framework? Yes, we discuss limitations of quasi-experimental frameworks in Section 5.1.
   (f) Have you related your theoretical results to the existing literature in social science? Yes, see Section 5 and 2.2.
   (g) Did you discuss the implications of your theoretical results for policy, practice, or further research in the social science domain? Yes, see Section 5.
3. Additionally, if you are including theoretical proofs...
   (a) Did you state the full set of assumptions of all theoretical results? NA
   (b) Did you include complete proofs of all theoretical results? NA
4. Additionally, if you ran machine learning experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? NA
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? NA
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? NA
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? NA
   (e) Do you justify how the proposed evaluation is sufficient and appropriate to the claims made? NA
   (f) Do you discuss what is "the cost" of misclassification and fault (in)tolerance? NA
5. Additionally, if you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? Yes, see Section 3.1 for the citation on the pre-existing sentiment classifiers we use for our study.
   (b) Did you mention the license of the assets? Yes, see Appendix C.

(c) Did you include any new assets in the supplemental material or as a URL? No, as the data we use are proprietary.

(d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? Yes, see Appendix C.

(e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? Yes, see Section 3.1 and Appendix C.

(f) If you are curating or releasing new datasets, did you discuss how you intend to make your datasets FAIR? NA

(g) If you are curating or releasing new datasets, did you create a Datasheet for the Dataset? NA

6. Additionally, if you used crowdsourcing or conducted research with human subjects...

(a) Did you include the full text of instructions given to participants and screenshots? NA

(b) Did you describe any potential participant risks, with mentions of Institutional Review Board (IRB) approvals? We discuss the potential participant risks of being exposed to uncivil content through active experiment, and describe how our methodology mitigates this risk in Section 1.

(c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? NA

(d) Did you discuss how data is stored, shared, and deidentified? Yes, see Appendix C.

# B Technical Appendix

## B.1 Civility metric construction details

To construct the score for a civil event $A_{\text{civil}}$, we normalize the empathy output to fall in the range of $[0, 1]$. For our qualitative textual analysis in Figure 4, we classify a chat as *empathetic* if it is scored above a 3, as the empathy classifer was trained on a survey data scaled 1 (least empathetic) to 5 (most empathetic) (Omitaomu et al. 2022).

## B.2 Selecting window for quasi-random region

We follow the window selection procedure recommended by Cattaneo, Frandsen, and Titiunik (2015) to determine the bandwidth of analysis. The intuition behind this procedure is to find the largest such difference between the matchmaking affinity score differences $s_A - s_B$ such that as-if randomization holds. To find this "window" size $w$, we sweep through candidate bandwidths and test the difference-in-means between $k$ input covariates $c_{A1}...c_{Ak}$ of server $A$ and server $B$ for the matchmaking affinity score, with tests that *fail to reject the null* indicative of as-if randomization. Note that in this situation, we are concerned with Type II errors i.e., failure to reject the null hypothesis when it is false, as opposed to Type I errors in the typical hypothesis testing case. We thus must be mindful of the test being too "conservative" in the sense of allowing for Type II errors; Cattaneo, Frandsen, and Titiunik (2015) recommend a significance value of

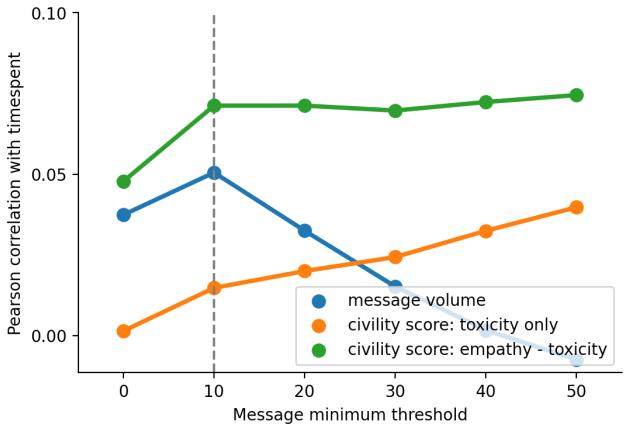



Figure B.1: Including empathy improves the signal between the communication civility metric and engagement. Message volume (with no sentiment or empathy weighting) in blue is used as a baseline. The x-axis corresponds to a minimum message inclusion threshold for data. We choose a minimum message threshold of 10 for our analyses (grey vertical line) to balance data availability and signal strength.

$\alpha = 0.15$ for testing the window size for as-if randomization. We consider window sizes from 0.01 to 0 with a step size of 0.001 and test for the difference-in-means between the input server scores of server $A$ and server $B$ in a given matchmaking decision. We follow Cattaneo, Frandsen, and Titiunik (2015) in that we consider the minimum p-value among all of covariates as the candidate p-value for a given window, and apply an additional acceptance criteria for robustness by only considering a window $w$ if the candidate p-value is $> 0.15$ across all data subsamples corresponding to distinct dates. We identify $w = 0.001$ as the largest window that satisfies this quasi-randomization testing criteria.

## B.3 Quasi-random server pair identification

We make the typical causal assumptions for an observational study of SUTVA (stable unit treatment value assumption) and positivity (Imbens and Rubin 2015; Neal 2020)), though we note that interference, which is part of SUTVA, on social network platforms (e.g., friend and peer network effects) is a concern that could influence our causal conclusions. Under the conditional unconfoundedness of Equation 4, our treatment effect can be *identified*, which entails converting our expression with causal quantities (e.g. potential outcomes) to an expression with only statistical quantities:

$$\begin{aligned}\tau &= E[Y(1) - Y(0)|M^* = 1] \\ &= E[Y(1)|M^* = 1] - E[Y(0)|M^* = 1] \\ &= E[Y(1)|T = 1, M^* = 1] - E[Y(0)|T = 0, M^* = 1] \\ &= E[Y|T = 1, M^* = 1] - E[Y|T = 0, M^* = 1] \qquad (6)\end{aligned}$$

## B.4 Regression Details

We show details of the four regressions run for average treatment effect estimation from Figure 9 below in Table B.1. The general form of the regression is:

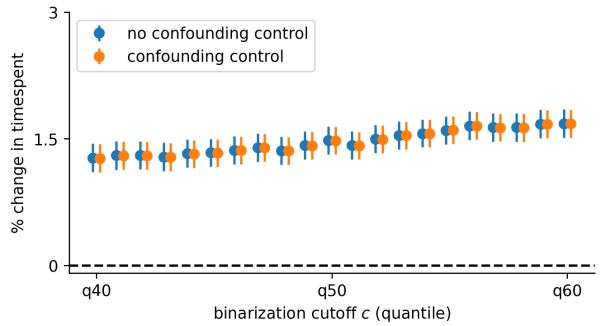



Figure B.2: The estimated effect of civil communication on engagement remains consistent at different binarizations of the civility metric. We re-run our matched server pair analysis for different quantile binarizations. Error bars displayed are standard errors. We find that the effect size stays consistent across different binarizations, and we observe the same consistency of estimates when controlling for confounders.

$$log(Y) = \beta_0 + \beta_1 T + \beta_2 H + ...$$

Where $Y$ is player session length, $T$ is the binary treatment of being placed on a civil server, and $H$ is the historical playtime for an experience as described in Section 4. Additional control regressors $C$ included for the observed confounding case include: indicator variables for gender, experience genre (action, simulation, roleplaying, simulator, sandbox, shopping, platformer, tabletop, social hangout, tycoon, adventure, idle, strategy, sports, minigames), as well as whether a user is over 13 or under 13.

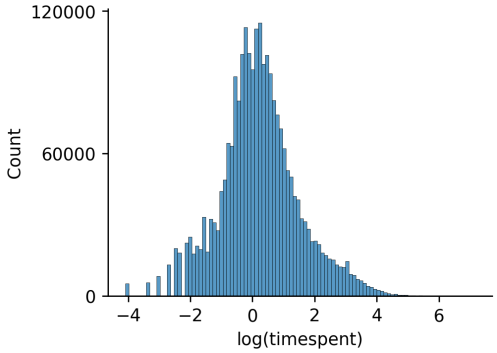



Figure B.3: Distribution of log(timespent) for our sample.

| Data | $\hat{\tau}$ | 95% CIs | P-value | $R^2$ |
|---|---|---|---|---|
| quasi-random, no ctl | 0.0148 | [0.012, 0.018] | $< 0.0001$ | 0.132 |
| quasi-random, ctl | 0.0147 | [0.012, 0.018] | $< 0.0001$ | 0.151 |
| convenience, no ctl | 0.0733 | [0.032, 0.065] | <0.0001 | 0.099 |
| convenience, ctl | 0.0074 | [-0.001, 0.017] | 0.077 | 0.108 |

Table B.1: Regression details on estimating the effect of civil server assignment on player engagement.

# C Data Appendix

## C.1 Data Collection

We use observational data under normal operating conditions of the Roblox platform during the analysis period of our study, in compliance with the Terms of Service and Privacy Policy of the platform (Roblox 2022b). All data used are anonymized and stored securely in accordance to Roblox's Privacy Policy.

**Licenses**. The empathy classifier is released under an "academic license" and is used with permission from the authors (Omitaomu et al. 2022).

| Empathetic Words | Relative Frequency |
|---|---|
| good | 1.0 |
| happy | 0.97 |
| feel | 0.70 |
| want | 0.65 |
| help | 0.60 |
| make | 0.56 |
| much | 0.52 |
| oh | 0.52 |
| will | 0.51 |
| omg | 0.45 |
| amazing | 0.43 |
| need | 0.42 |
| love | 0.42 |
| life | 0.42 |
| beautiful | 0.41 |

Table C.2: Top 15 empathetic words by relative frequency. Frequencies are normalized by the most common word i.e., "will" occurs roughly half as frequently as "good" within chats classified as empathetic.

| | Quasi-random | Convenience sample |
|---|---|---|
| Gender % | | |
| Male | 57.3 | 54.4 |
| Female | 32.1 | 37.7 |
| Unknown | 10.6 | 8.9 |
| Age Group % | | |
| Under 13 | 47.0 | 39.2 |
| Over 13 | 53.0 | 60.8 |
| Genre % | | |
| Action | 53.7 | 43.9 |
| Roleplaying | 12.3 | 26.5 |
| Strategy | 6.7 | 2.4 |
| Sandbox | 6.5 | 6.6 |
| Simulator | 6.0 | 5.7 |
| Platformer | 3.7 | 3.5 |
| Simulation | 2.9 | 1.7 |
| Tabletop | 2.7 | 1.6 |
| Social Hangout | 2.3 | 4.3 |
| Shopping | 1.6 | 1.1 |

Table C.3: Summary descriptive statistics for our analysis sample. Shown experience genres are above a 1% threshold.