Doing Good in the Digital World*

Jeffrey Da-Ren Guo†

November 21, 2023

[CLICK HERE FOR MOST RECENT VERSION]

Abstract

Though digital interactions between people have become more commonplace and sophisticated, behavior in digital settings remains underresearched. A distinctive feature of the digital world is the ability to calibrate or withhold one's identifier: a person can be identified by a string of letters, an avatar, their real name, or even nothing at all. That digital identifiers allow a person to mask their physical identity also makes it difficult to attribute digital actions to a physical person, even when the actions are observed. I embed these features in an experiment where subjects play a finitely repeated, linear public goods game. Treated subjects are identified in one of three ways—by their photograph, by a random number, or by a self-designed cartoon avatar—and their individual choices are revealed and either attributed to, or decoupled from, their identifier. In line with the previous literature, identifying subjects and increasing the precision of attribution increases contributions relative to a baseline condition without identifiers or revealed individual choices. Remarkably, however, the largest impact on behavior comes from having an identifier in the first place: for a given level of attribution, the experimental data suggest that being identified by a number or by an avatar is as powerful as being identified by one's photograph.

JEL codes: C72, C90, D91, H41

*I am indebted to Alessandra Casella and Mark Dean for their guidance and support. I thank Pietro Battiston, Sharon Harrison, Michelle Jiang, Navin Kartik, Judd B. Kessler, Victoria Mooers, Suresh Naidu, Shin Oblander, Jacopo Perego, Susie Scanlan, Silvio Ravaioli, Qianyang Zhang, as well as the attendees of the 2023 SWEET Workshop, the 2023 CTESS Workshop on Theory-Driven Experiments, and many seminars at Columbia for their comments. The Columbia Program for Economic Research and the Columbia Experimental Laboratory for the Social Sciences provided generous financial support. The experiment was approved under Columbia IRB Protocol AAAU1106. All remaining errors are my own.

†Columbia University, jeffrey.guo@columbia.edu

1 Introduction

The internet has given rise to digital spaces in which individuals gather and interact, from AOL chat rooms in the 1990s to modern-day forums like Reddit that attract over 50 million daily active users.[1] Technological advancements have facilitated digital interactions that more closely resemble those of the physical world around us. Massively multiplayer video games like *World of Warcraft* attract millions of players who walk around to meet each other's characters and complete tasks together. Platforms like *Second Life* and *VRChat* bring together thousands of users simply to hang out with each other, but as avatars in virtual worlds. And in recent years, developments in augmented and virtual reality technology have brought the physical and digital worlds closer together than ever before. As the line between the physical and digital continues to blur, it has become imperative to understand how individuals behave in the digital world, particularly because myriad anecdotes of online toxicity, harassment, and abuse remain pervasive.[2] How can the digital world be designed so as to induce people to do good unto one another?

One key distinction between interactions in the digital and physical worlds is how participants are identified to one another. In most physical interactions—running into someone on the street, gathering with friends for dinner—each participant is identified by their literal, physical person. Simply by showing up to an interaction in the physical world, one inherently brings along an identifier that uniquely pins down, or identifies, who they are. On the other hand, identifiers in the digital world can take many forms and vary in how precisely they identify the underlying person. In digital interactions, one could be identified by a random string of letters and numbers (imprecise), by a picture of their face (very precise), or even by nothing at all. And in recent years, lifelike avatars have become an increasingly common digital identifier, further blurring the line between the digital and physical.[3]

An inextricable consequence of the wide range of digital identifiers is that attributing actions to the underlying actor becomes more difficult. In most physical interactions, when an action

[1] https://www.redditinc.com/blog/reddits-2020-year-in-review/

[2] A 2021 *New York Times* article opens with the headline, "The Metaverse's Dark Side: Here Come Harassment and Assaults." In the same year, the Pew Research Center released a report that found that 41% of Americans had experienced some form of online harassment.

[3] The realism of avatars lends credence to the popular argument that an avatar is an extension of one's person. For example, Meta writes on their website at https://www.meta.com/avatars/ that "[a]vatars are a digital expression of you, letting you freely express your identity, personality and appearance."

is observed, it is naturally attributed to the person who performed it. Past studies find that unmasking identities and attributing actions have a complementary, positive effect on the level of prosocial behavior with physical identifiers (Andreoni & Petrie 2004; Rege & Telle 2004). But, it is unclear how this effect translates to digital settings, where actions need not be attributed to an identifier. Moreover, even if actions can be attributed to an identifier, the identifier itself may not identify the person behind the keyboard.[4] This complication raises the main questions that the paper seeks to answer: *Can less precise identifiers and less precisely attributed actions still induce prosocial behavior? Which channel is more powerful in inducing prosocial behavior?*

Thus, I conduct an experiment whose two arms of variations are in identification and the attribution of individual actions. First, subjects have identifiers that differ in the precision with which they identify the physical person. Treated subjects are identified to each other by either their photograph, a self-designed cartoon avatar, or a random three-digit number; *Baseline* (control) subjects are not identified to each other at all. The second arm of variation is in the attribution scheme used to publicly attribute subjects' choices in the game to their identifiers. Treated subjects' individual choices are publicly revealed and either attributed to (*Full Attribution*), or decoupled from (*Partial Attribution*), their identifier; *Baseline* subjects' individual choices are kept private.

I embed this variation in a linear public goods game (Groves & Ledyard 1977) played in fixed groups of four subjects each, finitely repeated for thirty-two rounds. Public goods games have been studied extensively in the laboratory to address questions about prosocial behavior in group settings.[5] My experiment implements a $(3\times2)+1$, between-groups design. The *Baseline* treatment replicates the canonical implementation of the public goods game in the lab: subjects are not identified to each other, and only the aggregate contribution to the public good is revealed after each round.[6] In each of the treatment conditions, two elements are added to the game. First, each subject has an identifier: *Number*, *Avatar*, or *Photo*. Second, each subject's contribution choice is also revealed to the group after each round—the breakdown of the aggregate contribution is made known. When the *Full Attribution* scheme is used to reveal choices, each subject's contribution choice appears underneath the subject's identifier. In contrast, when the *Partial Attribution* scheme

[4] Interactions on Reddit are an example of this. Each Reddit user's history of posts and comments are logged under their username. However, Reddit does not require its users to use their real name (orthonym) as their username.

[5] Ledyard (1995) and Chaudhuri (2011) survey results from public goods experiments in the laboratory.

[6] See, for example, the experimental implementation of the public goods game in Isaac *et al.* (1984).

is used, subject contributions are presented in descending order, from largest to smallest, without being linked to any specific identifier. Each of the six treatment conditions is characterized by the Identifier and the Attribution scheme shared by all members in a group. This dictates how I abbreviate the treatment conditions: for example, I will abbreviate the treatment condition with *Photo* and *Full Attribution* as *PhotoFull*.

Strictly speaking, neither adding identifiers nor revealing individual choices changes the subgame perfect Nash equilibrium prediction of the game, which states that players contribute zero ("free ride") in all periods of the game.[7] However, I analyze two standard channels that may induce positive contributions. The first is image concerns.[8] In a simple theoretical framework, I derive two predictions on equilibrium contributions when agents also derive utility from their image. These predictions revolve around two parameters. One parameter (b) captures the intensity with which an agent cares about, identifies with, or invests in their identifier[9]; the other parameter (p) captures the probability with which the agent's action can be attributed to their identifier.[10] The first prediction is that equilibrium contributions are increasing in b and p. Utility-maximizing agents contribute more when their contribution is more likely to be attributed to their identifier. Taken to the experiment, this predicts higher contributions under *Full Attribution* than under *Partial Attribution*. Additionally, assuming that subjects identify less more with their photograph than with a number (*i.e.*, $b_{\text{Photo}} > b_{\text{Number}}$), this also predicts higher contributions under *Photo* than under *Number*. The second prediction is that contributions have increasing differences in (b, p): the impact of a change in Identifier should be larger under *Full Attribution* than under *Partial Attribution*, and the impact of a change in Attribution should be larger under *Photo* than under *Number*. This follows from b and p being complements, as is consistent with the literature.

The second channel is the possibility that the game extends outside the laboratory. Imagine that group members are able to sanction each other *outside* the laboratory, even after the public goods

[7] It is well known that experimental subjects do not play the Nash equilibrium of contributing zero: rather, subjects contribute some (but not all) of their endowment in early rounds, and contribution levels deteriorate over time (Ledyard 1995; Fehr & Schmidt 1999).

[8] Impure altruism, or warm glow, is a commonly cited source of self image utility (Andreoni 1989, 1990). Various social effects—prestige, recognition, shame—are sources of utility derived from one's social image (Gächter & Fehr 1999; Samek & Sheremeta 2014).

[9] There is evidence that one's identifier can affect the level of self-image concerns one faces. Falk (2021) provides experimental evidence that a decisionmaker faces higher levels of self-image concerns when they see themselves in a mirror when making their decision, compared to seeing nothing or a neutral stimulus.

[10] Bénabou & Tirole (2006) model decisionmaking with image concerns. In their model, the reputational payoff of taking an action is increasing in the "visibility" of the action—the "probability that it will be observed by others."

game in the laboratory is over.[11] Could the possibility of being punished outside the public goods game induce positive contributions during it? I consider an infinite horizon supergame that consists of two games: a finitely repeated public goods game, followed by an infinitely repeated costly punishment game.[12] I show that the threat of punishment can indeed induce positive contributions in the public goods game, but that the size of the threat depends on whether group members can identify and coordinate punishment on a free rider. I show that the expected punishment from free riding is strictly higher when contribution choices are (1) attributed with certainty to an identifier that (2) can identify the free rider outside the public goods game. This result, when applied to the experiment, predicts higher contributions in the *PhotoFull* condition than in all other conditions.[13]

Relative to the *Baseline* condition with neither identifiers nor revealed individual choices, I find more prosocial behavior—higher contributions—in all six treatment conditions. The increases are statistically significant at the 5% level for five of the six treatment conditions; moving from *Baseline* to *PhotoFull* increases contributions by 68%. However, changes in Identifier and Attribution do not affect behavior equally. The main finding of the paper is that Attribution is the more powerful channel in inducing prosocial behavior. While average contributions are indistinguishable across the three different Identifiers (pooling across Attribution schemes), there is a significant increase in contributions from *Baseline* to *Partial Attribution*, and again from *Partial Attribution* to *Full Information* (pooling across Identifiers). In fact, under *Full Attribution*, contributions are just as high when subjects are identified by a random number or a cartoon avatar as when they are identified by their photograph. Put another way, the data suggest that mere presence of an Identifier

[11] Many studies consider the effect of being able to punish free-riders *within* the public goods game itself by incorporating a punishment scheme into the game. In such schemes, subjects choose if they want to punish other members of their group in between rounds of the public goods game (after observing the contribution choices of all group members), usually at a cost. Ostrom *et al.* (1992) implement a similar sanctioning scheme within a common pool resource game. The addition of a punishment scheme has been found to be an effective mechanism for inducing prosocial behavior (Fehr & Schmidt 1999; Fehr & Gächter 2000; Fehr & Gächter 2002; Denant-Boemont *et al.* 2007; Nikiforakis 2010).

[12] Imagine a group of friends which gathers one evening to play a board game. Actions taken in the game may spill over into their larger friendship: a person who behaves selfishly in the game may be shunned by the other friends afterward.

[13] Note that what I model here is different from an infinitely repeated public goods game. Analysis of an infinitely repeated public goods game using the Folk Theorem suggests that positive contributions can be sustained in equilibrium for sufficiently large discount rate. Notably, there exist common "trigger" strategies, such as Grim Trigger, that can sustain positive contributions in equilibrium with only feedback about the *aggregate* contribution between rounds—neither identifiers nor revealing individual contribution choices is necessary. If subjects mistakenly believe the public goods game in the laboratory to be infinitely repeated, and play such trigger strategies *as if* the game were infinitely repeated, then we would expect no difference in behavior between the *Baseline* condition and the treatment conditions.

is sufficient to induce higher contributions, as long as the Identifier is paired with information about individual choices.

These results are interesting in light of Andreoni & Petrie (2004), whose 2×2 design incorporates variation on the same two dimensions: whether subjects are identified, and whether their individual contribution choices are revealed. Their experiment has a condition in which subjects are not identified and only the aggregate contributions are revealed (analogous to my *Baseline* condition), as well as a condition where subjects are identified by their photograph and their individual contributions are attributed to their photo (analogous to my *PhotoFull* condition). In their two intermediate conditions, one of the two arms is entirely shut off: in one condition, subjects are identified by their photographs but only aggregate contributions are revealed; in the other condition, subjects have no identifiers, and individual contribution choices are displayed from largest to smallest. A significant increase in contributions (relative to their Baseline) comes only when identification with photographs is combined with revealing individual contribution choices. As they write, "adding just information on generosity has no significant effect on giving, and neither does adding just the identity of the giver. However, a substantial impact comes from using both in combination." The treatment conditions in my experiment maintain the presence of both identifiers and individual contribution choices, and instead vary the precision of the identifiers and the precision of the attribution of the choices to identifiers. Indeed, in five of my six treatment conditions, contributions are significantly higher than in the *Baseline*: this suggests that Andreoni and Petrie's result is robust to settings where identifiers and attribution are imprecise, but present.

Additionally, the experimental data are inconsistent with the punishment channel, and not fully consistent with the image channel. As a result, I conduct exploratory analysis to identify what is driving contribution behavior. I find that the increase in contributions from *Baseline* to *Partial Attribution* to *Full Attribution* is driven primarily by an increase in the rate of subjects contributing their full endowment: the rate of full contribution rises from 16.6% in *Baseline* to 38.6% in *Partial Attribution* to 62.3% in *Full Attribution*. Note that, when individual contribution choices are revealed, subjects are also given pieces of "relative" information: for example, the rank of their contribution in the group, the minimum and maximum contributions, whether an other member of the group contributed their full endowment or contributed nothing.[14] To test whether

[14] "Aggregate" information, on the other hand, can be known from having only the aggregate contribution to the

relative information influences future contribution behavior, I estimate a dynamic panel model with the experimental data using the Arellano & Bond (1991) estimator. In particular, I estimate the model separately for *Baseline*, *Partial Attribution*, and *Full Attribution*: not only do I check that a piece of relative information has a significant effect on future contributions under *Partial Attribution* and *Full Attribution*, I also check that it has no significant effect in the *Baseline*.

Indeed, the dynamic panel estimation shows that subjects condition future contributions on "relative" information gleaned from past rounds, but only when it is available. Under *Partial Attribution* and *Full Attribution*, subjects significantly reduce their contributions in period t when another group member free rides in period $t-1$; this does not happen in the *Baseline*. Moreover, the amount by which subjects increase their contributions from period $t-1$ to period t is increasing in the number of group members that contributed strictly more than the subject in period $t-1$—but only under *Partial Attribution* and *Full Attribution*. Taken together, these results suggest that when individual contribution choices are available, subjects use the additional information to establish a norm of reciprocity for their group.[15] When a subject knows that they contributed a relatively low amount, they increase their contribution in the next period; yet, when a subject realizes someone else violated the norm by free riding, they reduce their contribution in the next period. Interestingly, the analysis shows that the effects of reciprocity are stronger under *Full Attribution* than under *Partial Attribution*, even though the amount of relative information is the same under both Attribution schemes. This suggests that not just reciprocity, but also an interaction between reciprocity and image, drives behavior.

The experimental data also shed light on how individuals choose to identify and represent themselves in the digital world. From the *Avatar* condition, I observe the cartoon avatars created by 100 subjects (56 assigned to *AvatarFull*, 44 assigned to *AvatarPartial*), as well as the amount of time each subject spent on customizing their avatar. Subjects were able to choose visual attributes from between twelve dropdown menus; in total, over 2.8 trillion permutations of attributes were

group account: the total amount contributed by other group members, whether one's contribution was above or below the mean contributed by other group members.

[15] Croson (2007) also finds contribution behavior in public goods games that is consistent with reciprocity models (over commitment or altruism models). In particular, they find that subjects in round t try to match the median contribution from round $t-1$, as opposed to the minimum or maximum contribution. In a field experiment, Shang & Croson (2009) find evidence of reciprocity in giving. For a general discussion of the theory of reciprocity, see Meier (2007).

available to be chosen.[16] Additionally, after the public goods game ends, *Avatar* subjects reported their race and gender, as well as the intensity with which they felt represented by their avatar.[17] I find that subjects invested substantial time in customizing their avatar: on average, subjects spent just over three minutes making their avatars.[18] However, I do not find a correlation between a subject's customization time and their contribution in the first round of the game. I do find a positive, albeit weak, correlation between self-reported representation and first round contribution among *AvatarPartial* subjects (but no such correlation for *AvatarFull* subjects). Finally, there is evidence that some subjects created avatars that look different from their physical person, based on their reported race and gender.

1.1 Literature

This paper contributes primarily to a rich literature that studies the effect on prosocial behavior of identification and attribution of individual actions. Much experimental work finds a positive effect of both channels on prosocial behavior in both lab and field settings.[19] Charness & Gneezy (2008) identify some players in dictator games by their family names, and find that dictators allocate more of the endowment to the other player when names are known. Ariely *et al.* (2009) find that subjects exert more effort in a real-effort task for charity when they are compelled to reveal the amount of money they earned to other participants in the lab. In the field, Soetevent (2005) studies charitable giving at churches in Denmark by manipulating the visibility of others' donations in the collection vessel; churchgoers give more when the collection vessel allows others' contributions to be seen. Karlan & McConnell (2014) pair a field experiment conducted through an on-campus charitable organization with a lab experiment that reveals donation decisions of some subjects to others in the room. Evidence from their experiment suggests that individuals give in order to improve their social image, rather than purely altruistically. Indeed, the results of my experiment are consistent with this pattern: subjects behave more prosocially when they are identified, and when their actions can be attributed back to them.

[16] The cartoon avatars looked like passport or ID photographs: a "face" from the "shoulders" up. Figure 1 shows the dropdown menus available, as well as a sample avatar.

[17] Subjects rated their agreement on a 7 point Likert scale (1 being Strongly Disagree, 7 being Strongly Agree) with the following statement: *I felt like my avatar represented me.*

[18] I find no significant difference in customization time between *AvatarPartial* and *AvatarFull* subjects.

[19] Dufwenberg & Muren (2006) do not find a positive effect of identification and attribution in their lab experiment. However, they note that their experimental implementation introduces confounding factors.

A similar effect has been found in public goods games: identification and attribution have a positive effect on contributions. In the lab, variation in attribution was first implemented by revealing individual contribution choices in addition to the aggregate contribution. Results from past studies are mixed: Sell & Wilson (1991) find a significant increase in contributions in the last five rounds of their experiment, while Weimann (1994), Wilson & Sell (1997), and Croson (2001) do not find an effect of revealing individual choices on contributions.[20] In a related study, Jones & McKee (2004) find that presenting "relative" information to subjects—including the maximum and minimum contribution choices and the ranking of one's contribution in the group—increases contributions. More recently, experimenters have also identified the players themselves, along with revealing their choices. Rege & Telle (2004) find an increase in contributions when players are compelled to count out their contribution and write the amount on a blackboard in front of the group. Samek & Sheremeta (2014) identify players using their photographs and first names: they find that revealing the two lowest contributors and their choices each round induces the same increase in contributions as revealing all players and their contributions. On the other hand, Andreoni & Petrie (2004)'s treatment with photographs, but only aggregate contributions, doesn't find a statistically significant increase in contributions compared to their Baseline. Moreover, their treatment with individual contribution choices, but without identifiers, finds a decrease in contributions relative to their Baseline. The results from my experiment suggest that identifiers and individual choices are needed in conjunction to increase contributions, as is consistent with previous findings. The novel finding from my experiment is that this result still holds, even when both the identifier and the attribution scheme are imprecise.

Also related is a smaller literature that studies the relationship between one's avatar and behavior. Lim & Reeves (2009) establish a physiological effect of being able to choose one's avatar: in their experiment, subjects who customized an avatar had a 10% faster heart rate compared to those who were assigned their avatar. Other studies have shown an effect of a person's avatar on their behavior. Yee & Bailenson (2007) find that subjects who were assigned tall avatars in virtual

[20] It should be noted that in the Sell & Wilson (1991) and Wilson & Sell (1997) experiments, individual contribution choices were presented in a way that allowed a subject's contribution history to be tracked over time. In contrast, Croson (2001) presents individual contribution choices in ascending order. Weimann (1994) does not describe how individual contribution choices were presented, but does note that "[s]ubjects have no contact before, after and during the game: they act in strict anonymity."

reality behaved more confidently than subjects assigned short avatars.[21] The effect of an avatar has also been shown to extend into the physical world: Groom *et al.* (2009) find that subjects who were assigned a black avatar in virtual reality subsequently exhibited greater racial bias in the physical world than those assigned a white avatar. Moreover, avatars have been shown to be effective substitutes for photographs in reducing social distance, and increasing cooperation and trust between individuals.[22] On the other hand, there is evidence that people don't simply create avatars that look like themselves. Vasalou & Joinson (2009) and Martey & Consalvo (2011) find that users in virtual worlds consider the types of people they will interact with as well as the nature of the interactions themselves when customizing their own avatars.[23] Recent experimental work in economics has found that strategic considerations could also drive such behavior. Charness *et al.* (2020) find that, when competing to be "hired" for a math task, female subjects choose female avatars significantly less often than when the nature of the task is unknown. Abraham *et al.* (2023) also find a gender effect in a marketplace experiment: they find that buyers trust female avatars more than male avatars. In response, sellers with "genuine" male avatars are more likely to switch to female avatars than vice versa.

Compared to past studies, my experiment gives subjects many more degrees of freedom in creating their avatars: while subjects in Charness *et al.* (2020) chose one of three avatars and subjects in Abraham *et al.* (2023) chose one of eight avatars, subjects in my experiment had billions of avatars available to them via the dropdown menus. Given this freedom, I find that subjects do invest time in customizing their avatars, though the time spent customizing does not depend on whether actions will be attributed to their avatars with certainty or only probabilistically. Furthermore, I find some evidence that subjects don't make avatars that look like their physical person, though my design does not allow me to identify the reason.

The paper proceeds as follows. Section 2 describes the experiment. Section 3 discusses the theoretical framework. Section 4 presents the results, and Section 5 concludes. The formal theo-

[21] In their experiment, subjects wore virtual reality 'helmets' and played multiple rounds of an ultimatum game against a confederate. Subjects assigned tall avatars proposed splits that were significantly more in their own favor than subjects assigned short avatars. Moreover, subjects assigned tall avatars were half as likely to accept an unfair offer from a confederate as subjects assigned short avatars.

[22] See, for example, Bente *et al.* (2008), Fiedler & Haruvy (2009), and Bente *et al.* (2014).

[23] Vasalou & Joinson (2009) study subjects who were asked to create an avatar on the Yahoo! Avatars site either for blogging, dating, or gaming. Martey & Consalvo (2011) analyze observational data, survey responses, and interviews of users in the *Second Life* virtual world.

retical framework is left to the Appendix. Experimental instructions and screenshots appear in the Appendix.

2 Experimental Design

The game in the experiment is the finitely repeated, linear public goods game (also known in the literature as the voluntary contribution mechanism). I begin by describing the one-shot stage game and the finitely repeated game, as well as the standard theoretical predictions for those games. Next, I describe the experimental design and implementation.

2.1 The Linear Public Goods Game

The stage game is parameterized by $\{N, m, Y\}$, where N agents are in a group together, and each is endowed with Y points. Each agent $i \in \{1, ..., N\}$ privately chooses an amount of points $g_i \in [0, Y]$ to contribute to a group account (the public good), and keeps the remaining $Y - g_i$ points in a private account. Every point contributed toward the group account is multiplied by a factor m, where $m > 1$ but $m < N$. Total contributions to the public good are redistributed to the agents in equal shares. Hence, agent i's payoff in the stage game is given by:

$$\pi_i(g_1, ..., g_N) = (Y - g_i) + \frac{m}{N}(g_i + \sum_{j \neq i} g_j) \tag{1}$$

Each agent's marginal per capita return (MPCR) from contributing a point to the group account is $\frac{m}{N} < 1$, while the MPCR from keeping a point in their private account is 1. Assuming an agent seeks to maximize only their own payoff, the agent has a dominant strategy of contributing zero to the group account in the stage game: I call such a strategy *free riding*. Therefore, all agents free ride in the unique Nash equilibrium of the stage game: $(g_1, ..., g_N) = (0, ..., 0)$; the equilibrium payoff for all agents is equal to Y—the initial endowment.[24]

The stage game can also be repeated. Suppose the public goods game is repeated finitely for $T > 1$ discrete periods indexed by t. Denote agent i's contribution in period t as g_{it}. I assume that the aggregate contribution to the group account in period t is made known to all group members,

[24]Note that this outcome is not Pareto optimal because $m > 1$: each point becomes bigger when contributed to the group account. The unique Pareto optimal outcome in the stage game occurs when all agents contribute their entire endowment to the group account—that is, $(g_1, ..., g_N) = (Y, ..., Y)$, and all agents earn a payoff equal to mY.

before contribution decisions are made in the next period $t+1$. Formally, denote the aggregate contribution in period t as $G_t \equiv \sum_{i=1}^{N} g_{it}$. Thus, for all periods $t > 1$, the history of past aggregate contributions $\boldsymbol{G}_t \equiv \{G_1, ..., G_{t-1}\}$ is common knowledge.

Standard analysis of the repeated game is straightforward. Because the game is finitely repeated, and it is common knowledge that the game is finitely repeated, the game in the final period T can be analyzed in the same manner as the stage game: the unique Nash equilibrium in the final period is $(g_1, ..., g_N) = (0, ..., 0)$. Backward induction then yields that the unique subgame perfect Nash equilibrium remains: $(g_1, ..., g_N) = (0, ..., 0)$ in each round.

Note that if the game were instead infinitely repeated, then the Folk Theorem would apply: positive contributions could be sustained in equilibrium. In particular, note that if the game were infinitely repeated, it is possible to sustain positive contributions while knowing only the level of *aggregate* contributions in previous periods.[25]

2.2 Experimental Design

In all sessions, subjects played the finitely repeated, linear public goods game just described in fixed groups. Each subject participated in exactly one session and belonged to exactly one group: subjects stayed in the same group for the duration of the experiment. The game was parameterized as follows: groups were of size $N = 4$ and the game was finitely repeated for $T = 32$ rounds. In each round, each subject was endowed with $Y = 20$ points and privately chose an integer amount $g \in \{0, ..., 20\}$ to contribute to the group account; the remaining $20 - g$ points would be kept in their private account. All contributions to the group account were doubled ($m = 2$), so that $\frac{m}{N} = 0.5$. After each round, subjects were told the aggregate contribution to the group account (G), as well as their own earnings from that round. At the end of the experimental session, subjects were paid in cash based on the total number of points accumulated across all 32 rounds, converted at a rate

[25] Per the Folk Theorem, it is possible to sustain strictly positive contributions in equilibrium as long as agents are sufficiently patient (*i.e.*, the discount factor $\delta \in (0, 1)$ is sufficiently large). In context of the public goods game, it has been shown that such an equilibrium can be sustained by many different types of "trigger" strategies, which condition future contribution(s) on the history of past aggregate contributions ($\boldsymbol{G}_t$). Such trigger strategies take the following general form: *Contribute a strictly positive amount $g > 0$. If an other player deviates and contributes less than g, then punish the deviation by contributing $g = 0$ in the next period. Continue to contribute $g = 0$ for some number of periods, after which return to contributing the strictly positive amount $g > 0$.* A common trigger strategy is Grim Trigger, which says to contribute one's entire endowment ($g = Y$) until an other player deviates, at which point one should contribute zero forever. But it can be shown that other trigger strategies, which dictate a different amount to contribute g and/or a different duration of punishment, can be supported in equilibrium as well (Lugovskyy *et al.* 2017).

of 50 points to \$1.00 (or 1 point to \$0.02).[26]

The experiment used a between-groups, $(3 \times 2) + 1$ design. The *Baseline* condition implements exactly the finitely repeated game as described above: only the aggregate contribution to the public good is made public after each round, and subjects are not identified to each other in any way.[27] The six treatment conditions introduce two additional elements to the game. First, each subject is identified by an identifier in the game: either a randomly selected three-digit number (*Number*)[28], a photograph taken of the subject (*Photo*)[29], or a cartoon avatar customized by the subject (*Avatar*).[30] I abbreviate each treatment condition in two words: one for the Identifier, and one for the Attribution scheme. For example, I write the treatment condition with *Photo* and *Full Attribution* as *PhotoFull*.

In *Avatar* sessions, subjects were given five minutes before the start of the game (but after instructions were read) to customize their avatar by selecting different visual attributes from drop-down menus. The avatar customization screen was coded in a way that allowed subjects to see, in real time, how changing a particular attribute would affect the appearance of the avatar. Hence, subjects could experiment with different attributes—and see how they would look—before finalizing their avatar. Figure 1 shows screenshots of the avatar customization interface.

Second, subjects are *additionally* given, after each round, the individual contribution decisions made by each group member. These individual contribution decisions are either attributed to the identifier of the subject that made the contribution (*Full Attribution*), or decoupled from the identifiers (*Partial Attribution*). *Partial Attribution* of individual contribution decisions was implemented by presenting the contribution amounts from highest to lowest after each round, while the order of identifiers remained constant throughout the session. Figure 2.2 shows how *Full Attribution* and *Partial Attribution* were implemented in the experimental interface.

The combinations of different Identifiers and different Attribution schemes create different levels

[26] Subjects were paid one at a time, in private. Subjects saw only their own earnings, and were instructed not to discuss their earnings with others.

[27] The Baseline condition also corresponds to the canonical implementation of the public goods game in experiments.

[28] In *Number* sessions, no two subjects in the same group had the same number.

[29] In *Photo* sessions, each group member was identified by photograph taken at the start of the experimental session. The photographs were taken one person at a time by the experimenter, using a digital camera. All photographs were taken in public view of all participating subjects.

[30] The universe of cartoon avatars used in the experiment are from an open source Sketch library designed by Pablo Stanley, available at https://www.avataaars.com/. The avatar customization interface and dropdown menus used in the experiment were based on, and supported by, an open source web-based app developed by Fang-Pen Lin, available at http://getavataaars.com.

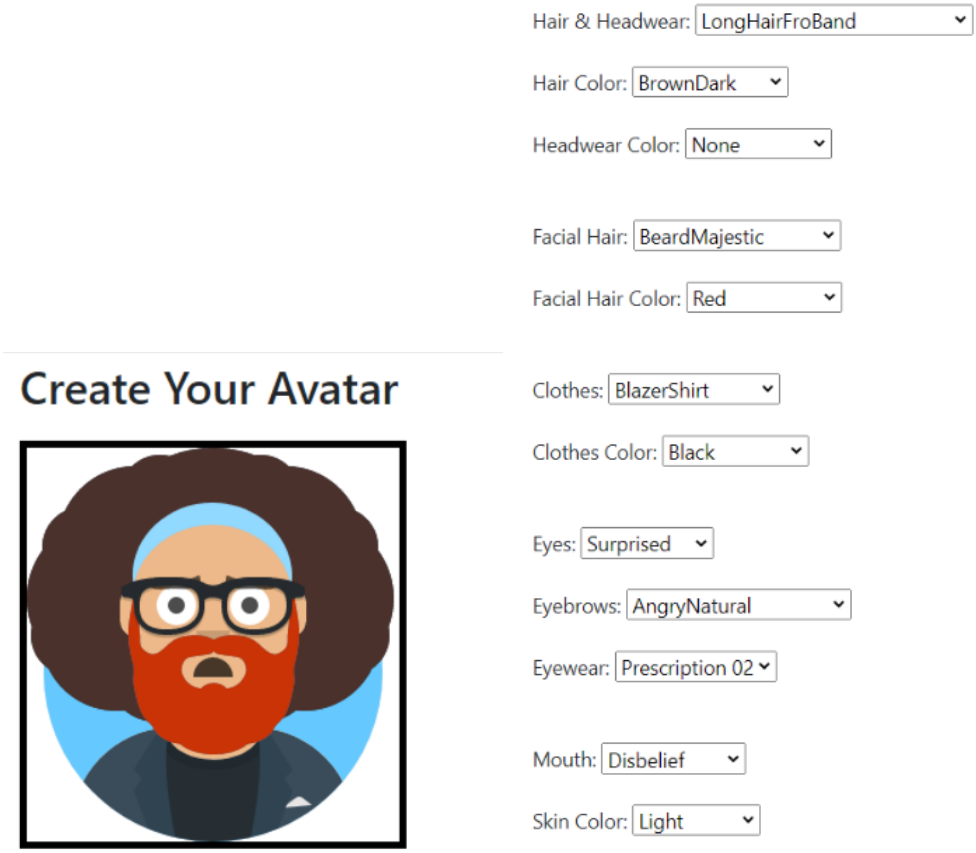


Figure 1: Experimental Interface, Avatar Customization. The image on the right shows the different attribute categories and dropdown menus available to subjects. The image on the left shows the avatar that corresponds to the specific attributes chosen in the image on the right.

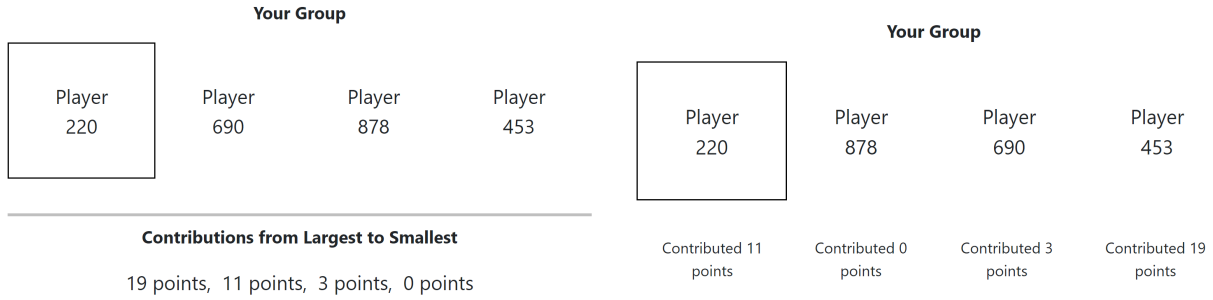


Figure 2: Experimental Interface, Round Results. *Partial Attribution* (left) and *Full Attribution* (right).

	Aggregate Only	*Partial Attribution*	*Full Attribution*
No Identifier	10		
Number		10	12
Avatar		11	14
Photo		10	10

Table 1: Number of Groups Assigned to Each Condition

of anonymity across treatment conditions. A subject who is identified by their photograph in the game can be definitively identified by others outside of the lab, whereas an avatar typically will (and always can) look different from the corresponding subject, and a number never allows a physical identification. Similarly, one can definitively attribute a contribution decision to an identifier under *Full Attribution*, whereas one can only do so probabilistically under *Partial Attribution*. In the experiment, the condition with the most anonymity is the *Baseline* condition: subjects have no identifiers whatsoever, and individual contribution choices are not revealed. The least anonymous condition is *PhotoFull*: actions taken inside the lab can (1) be attributed with certainty to an identifier that (2) allows the subject to be identified outside the lab. In all remaining conditions, one of these links is weakened, so the level of anonymity sits somewhere in between *Baseline* and *PhotoFull*. For example, the *NumberFull* condition still attributes actions taken in the lab to an identifier with certainty, but the identifier itself—the random number—doesn't serve to identify the subject beyond the experiment. On the other hand, the *PhotoPartial* condition identifies subjects in a way that extends outside the lab, but does not allow actions taken in the lab to be attributed with certainty to the identifier.

One treatment condition was assigned to each session, and subjects within a session were randomly assigned to groups of four. Results from a power analysis dictated that each condition have at least 10 groups. Some conditions had more than 10 groups due to higher-than-expected turnout by subjects who signed up for some experimental sessions. Table 1 summarizes the number of groups assigned to each treatment.

The number of groups in each session varied based on the number of subjects that signed up and turned out. For *Baseline*, *Number*, and *Avatar* sessions, between two and five groups participated in each session: this ensured that subjects could not identify their group members outside the lab with certainty, just by seeing who else was in the lab with them. Between one and three groups participated in each *Photo* session. Recall that each subject belongs to and interacts with only

one group for the duration of the experiment—groups are never reshuffled or remade—so the total number of subjects in a session does not affect any strategic aspect of the game. In total, thirty sessions were run.

The experiment was conducted in person at the Columbia Experimental Laboratory for Social Sciences (CELSS) between September 2022 and March 2023. 308 subjects were recruited from the CELSS subject pool using the Laboratory's ORSEE recruitment system (Greiner 2015). Most subjects were undergraduate students at Columbia University or Barnard College. The experiment was programmed in oTree (Chen *et al.* 2016) and subjects accessed the experimental interface on desktop computers at CELSS. Each experimental session lasted about 60 minutes. Average subject earnings were $22.07, including a $5 show-up payment. Instructions were read aloud to all subjects by the experimenter at the front of the room before the start of the game, and all subject questions were answered publicly.[31]

3 Theoretical Framework for Anonymity

A key finding of Andreoni & Petrie (2004) is that attributing individuals' contribution choices to their photographs induces higher contributions in a finitely repeated public goods game. This finding is commonly understood to reflect the fear of being identified as a low contributor in the group; indeed, Samek & Sheremeta (2014) find that revealing the two lowest contributors and their contributions between rounds raises contributions as much as revealing all players and their contributions. However, since the game is finitely repeated, adding identifiers or the individual contribution choices does not change the equilibrium prediction of free riding in each round. Moreover, even if players mistakenly believed that the game were infinitely repeated, fear of punishment in the game cannot explain the higher contributions, since punishment cannot be individually targeted (unlike in a prisoners' dilemma).

It follows that we must consider other motivations that can explain this effect. In this section, I consider two channels through which anonymity (or a lack thereof) could affect contribution behavior: image concerns and punishment outside the game. I derive predictions for how these channels would influence contributions in the public goods game. While these channels are standard

[31] The Appendix contains representative instructions, as well as additional screenshots of the experiment interface.

in the literature, they will not be sufficient to fully explain the results I present in Section 4.

3.1 Image Concerns

One possible consequence of introducing identifiers or revealing individual contribution choices is that subjects have image concerns within the game. Further, suppose subjects derive utility not only from their payoff in the public goods game, but also from their self image and social image within that interaction. I model self image utility as coming a warm glow effect of contributing à la Andreoni (1989) and Andreoni (1990), and I model social image utility as coming from the various social effects of contributing—recognition, prestige, shame. I introduce additional parameters capture the intensity with which players 'care about' their identifier and the probability with which their action is attributed to the identifier. In the model, these parameters capture the level of anonymity in the game, and serve to amplify or attenuate image utility. I assume all agents are symmetric, and write agent i's utility function as follows:

$$U_i(g_i, g_{-i}) = \pi_i(g_i, g_{-i}) + (1 + b)W(g_i) + bpR(g_i) \tag{2}$$

Let $g_i \in [0, Y]$ denote the amount that agent i contributes to the public good.[32] The first term π_i is agent i's payoff from the public goods game, as defined in Equation 1. The second and third terms capture the agent's self image utility and social image utility, respectively. $W : [0, Y] \to \mathbb{R}$ maps from contribution choice to self-image (warm glow) utility; analogously, $R : [0, Y] \to \mathbb{R}$ maps from contribution choice to corresponding social image (recognition) utility. I assume that W and R are weakly increasing in g_i, concave, and smooth: the more the agent contributes, the stronger the warm glow, and the more recognition they receive.

The parameters b and p model the level of anonymity in the game. $p \in [0, 1]$ parameterizes the probability with which the agent's action can be attributed to their identifier. The parameter $b \geq 0$ captures the intensity with which the agent cares about, invests in, or identifies with their identifier. I normalize $b = 0$ when players do not have identifiers. In the absence of identifiers, $p = 0$. Note that b appears in both image utility terms: self image utility and social image utility are amplified by the extent to which the agent cares about their identifier. Moreover, note that b

[32] I allow non-integer contributions here for tractability.

and p are complements in the social image utility term. Even if one is identified by their photograph (large b), there is little social image to be gained from contributing if the contribution is unlikely to be attributed to one's photograph (small p). Similarly, there is little social image to be gained from contributing if the contribution will be attributed to an identifier one cares little about (small b, large p).

In the experiment, $b = p = 0$ in the *Baseline* condition. Under *Full Attribution*, $p = 1$, and under *Partial Attribution*, $p = \frac{1}{3}$.[33] While impossible to ascribe values of b to a random number, cartoon avatar, or photograph, it seems reasonable to suppose that the following holds: $b_{\text{Number}} \leq b_{\text{Avatar}} \leq b_{\text{Photo}}$, with at least one inequality being strict.

As I show formally in the Appendix, sufficiently strong image concerns can induce positive contributions in equilibrium.[34] I state here two key results from the image concerns framework, and their associated predictions for the experiment.

Result 1. Equilibrium contributions g^* are increasing in b and in p: $\frac{\partial g^*}{\partial b} > 0$ and $\frac{\partial g^*}{\partial p} > 0$.

Prediction 1a. Contributions in the Baseline will be lower than contributions in all six treatment conditions.

Prediction 1b. $g^*_{\text{Baseline}} < g^*_{\text{Partial Attribution}} < g^*_{\text{Full Attribution}}$, for all identifiers.

Prediction 1c. $g^*_{\text{Number}} \leq g^*_{\text{Avatar}} \leq g^*_{\text{Photo}}$, with at least one inequality being strict, for all Attribution schemes.

Result 2. g^* has increasing differences in (b, p), since b and p are complements.

Prediction 2a. $(g^*_{\text{PhotoFull}} - g^*_{\text{PhotoPartial}}) > (g^*_{\text{NumberFull}} - g^*_{\text{NumberPartial}})$. Contributions are more sensitive to a change in p when b is larger.

Prediction 2b. $(g^*_{\text{PhotoFull}} - g^*_{\text{NumberFull}}) > (g^*_{\text{PhotoPartial}} - g^*_{\text{NumberPartial}})$. Contributions are more sensitive to a change in b when p is larger.

3.2 Punishment outside the Game

Another possible consequence of introducing identifiers or revealing individual contribution choices is that it enables players to be punished outside of the public goods game itself. Imagine a group

[33] I state these probabilities from the perspective of a subject in the game. Assuming that subjects recall their own contribution and identifier, there remain three contribution choices that need to be attributed to identifiers. Under *Partial Attribution*, the probability that a contribution choice will be attributed to the correct identifier is $\frac{1}{3}$.

[34] I show formally in the Appendix that equilibrium contributions can be strictly positive, and can be as large as the endowment Y under certain conditions.

of friends which gathers one evening to play a board game. Actions taken in the game may spill over into their larger friendship: a person who behaves selfishly in the game may be shunned by the other friends afterward, even after the game itself is over. I model this intuition in my setting as follows: I consider an infinite horizon supergame, in which a finitely repeated public goods game is followed by an infinitely repeated, costly punishment game. In every period of this punishment game, each player can punish as many (or few) of their group members as they want to; punishing another player, as well as being punished, are both costly. Can the threat of being punished deter free riding in the preceding public goods game? I show formally in the Appendix that the answer is yes.

This framework is relevant to the experiment because the equilibrium of the supergame depends on whether the identity of a free rider in the public goods game can be common knowledge. If a free rider's identity can be commonly known, the other players are able to select the equilibrium that inflicts maximal punishment on the free rider. If not, then the free rider escapes maximal punishment with strictly positive probability. Thus, when a free rider's identity cannot be commonly known, the expected punishment from free riding is strictly lower. This further implies that there exists some discount factor that can sustain positive contributions when a free rider's identity can be commonly known, but not when a free rider's identity cannot be known; moreover, the reverse does not hold.

When can the identity of a free rider, or deviator, be commonly known in the public goods game? Note that for any given profile of contribution strategies in the public goods game, observing the aggregate contribution after each round is sufficient to reveal the *presence* of a deviator.[35] But only when a deviation can (1) be attributed with certainty to an identifier that (2) allows the deviator to be identified outside the game, can the deviator's identity be common knowledge. In the experiment, only in the *PhotoFull* treatment are conditions (1) and (2) both satisfied.

Result 3. There exists some discount factor $\delta \in (0, 1)$ that can sustain positive contributions in the public goods game if (1) a deviation can be attributed with certainty to an identifier that (2) allows the deviator to be identified outside the game, but cannot sustain positive contributions otherwise. The reverse does not hold.

[35] For example, consider a profile of strategies that dictates that all N group members contribute their entire endowment. The presence of a deviator is revealed if the aggregate contribution is observed to be less than NY.

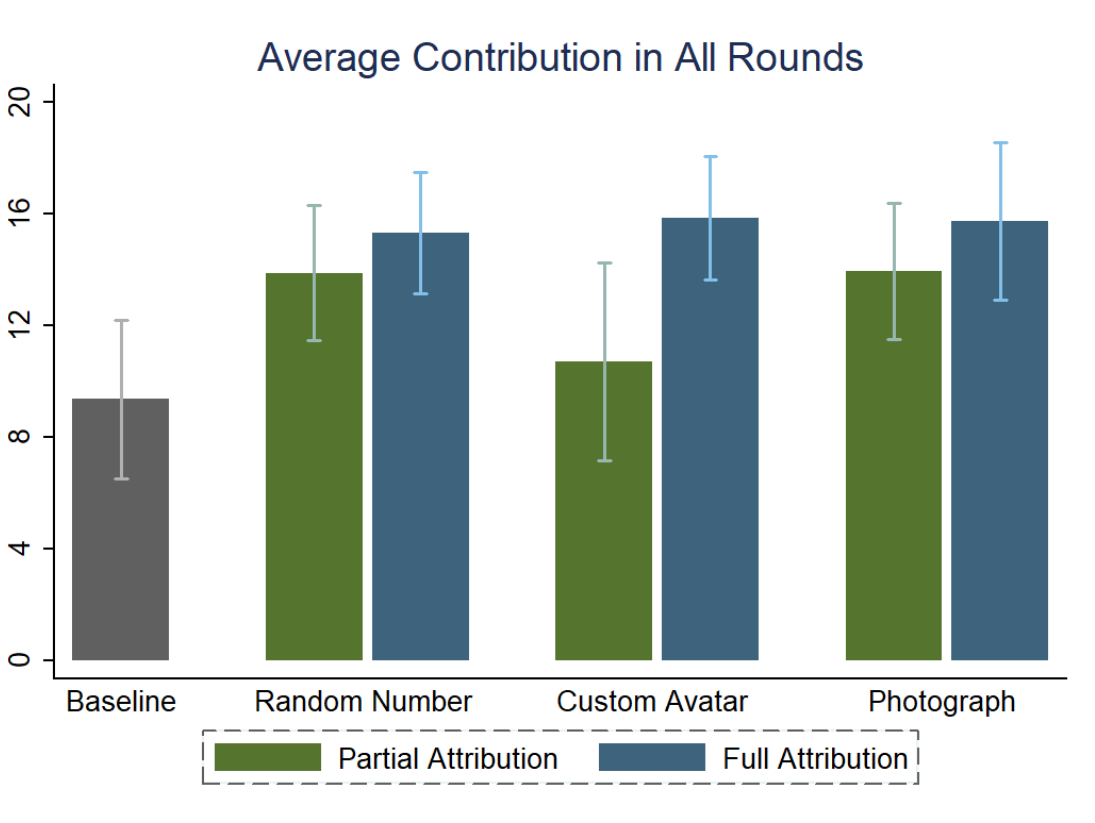


Figure 3: Average Contribution by Treatment. Error bars show a 95% confidence interval, with standard errors clustered at the group level.

Prediction 3. $g^*_{\text{PhotoFull}} > g^*_{\text{All Other Conditions}}$.

4 Experimental Results

4.1 Overview

Average contributions aggregated across all rounds are displayed in Figure 3, while Figure 4 shows the average contribution over time. While subjects do not play the Nash equilibrium of contributing zero, contributions decline over time in all treatment conditions. Both observations are consistent with behavior of subjects in past experiments with public goods games.

4.2 The Treatment Effect of Identifiers and Attribution

Does introducing identifiers or revealing individual contribution choices increase contributions? Figure 3 shows that the answer is yes. Average contributions are higher in all six treatment

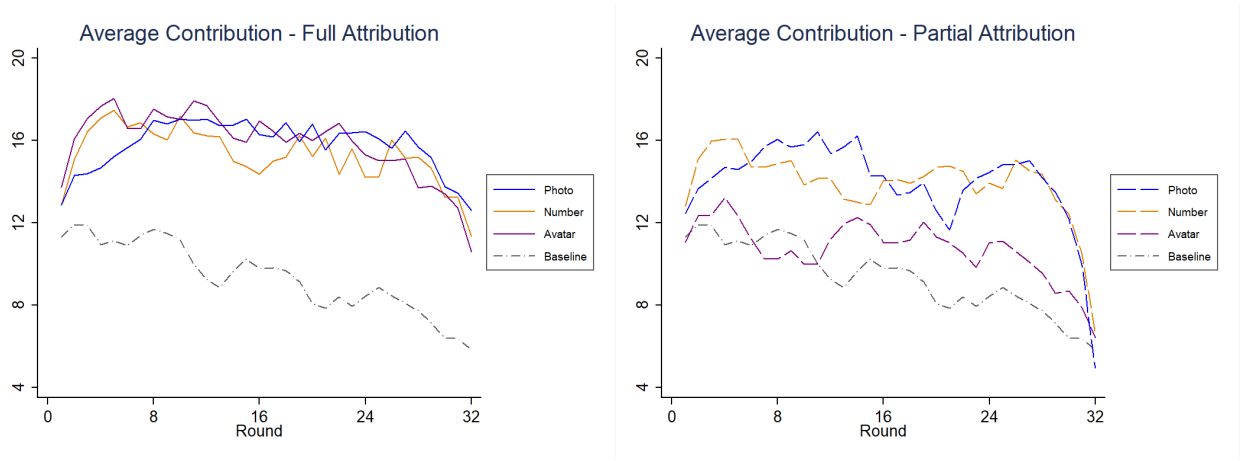


Figure 4: Average Contribution over Time.

Condition	Average Contribution (as % of Endowment)	Mann Whitney Test Statistic (vs Baseline)	p-value
Baseline	47%	-	-
NumberPartial	69%	2.26	0.023
NumberFull	77%	2.57	0.010
AvatarPartial	54%	0.63	0.526
AvatarFull	79%	2.86	0.004
PhotoPartial	70%	2.04	0.041
PhotoFull	79%	2.79	0.005

Table 2: Average Contribution and Mann Whitney Test by Treatment.

conditions than in the *Baseline* condition. Moreover, Figure 4 shows that contributions under *Full Attribution* lie strictly above contributions in the *Baseline* condition, even in the final round of the game. To test whether contributions in the treatment conditions are statistically significantly different from contributions in the Baseline, I perform a Mann Whitney U test, where each group of four subjects is treated as one observation.[36] The U test shows that the difference in average contributions is significant at the 5% level for five of the six treatment conditions (*AvatarPartial* being the exception). Table 2 summarizes the results.

As contributions in the treatment conditions are higher than in the *Baseline*, *Prediction 1a* from the image concerns framework is satisfied. On the other hand, contributions do not exhibit increasing differences. In fact, across all Identifiers, contributions stay equally high under *Full*

[36]Recall that the public goods game is played in fixed groups: each subject had strategic interactions only with the other three members of their group. Thus, I treat each group as an independent observation.

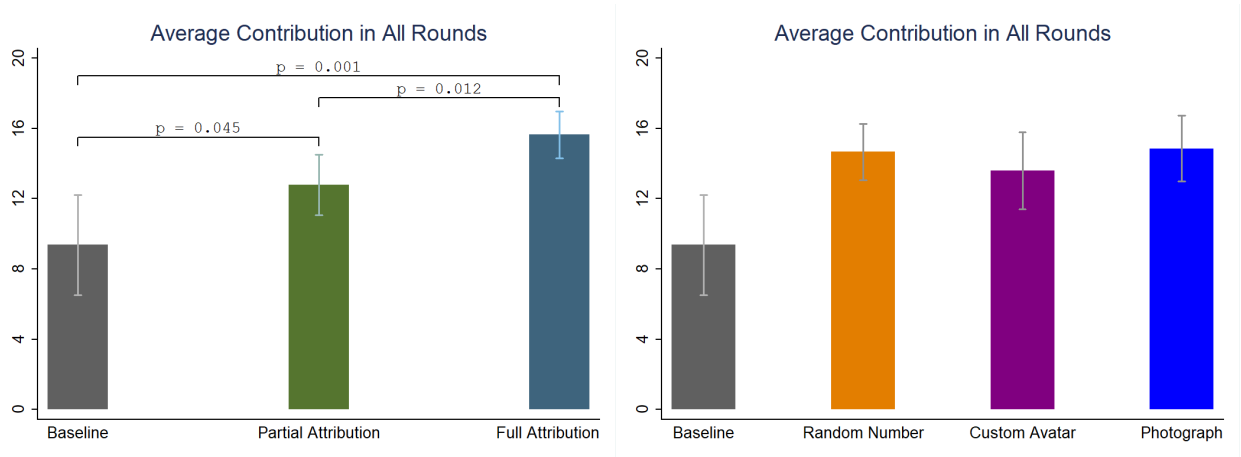


Figure 5: Average Contribution by Attribution (left) and by Identifier (right). Error bars show a 95% confidence interval, with standard errors clustered at the group level. p-values are shown for a Mann Whitney U test, where each group of four subjects is treated as one observation.

Attribution. Hence, the data are not consistent with *Prediction 2a* and *Prediction 2b* from the image concerns framework. Moreover, contributions in *PhotoFull* are not significantly greater than all other conditions: *Prediction 3* from the punishment framework is also not satisfied.

To isolate the effect of changing Identifier (Attribution) on contributions, I pool the data from the treatment conditions across Attribution (Identifier). The data show that changing the level of Attribution has a stronger effect on contributions than changing Identifier. The left panel in Figure 5 shows average contributions for the different levels of Attribution, pooling across Identifiers; the right panel shows average contributions for the different Identifiers, pooling across levels of Attribution. Pooling across Identifiers, the average contributions in Baseline, Partial Attribution, and Full Attribution are all significantly different from each other at a 5% level: *Prediction 1b* from the image concerns framework is satisfied. On the other hand, when pooling across Attribution levels, average contributions in Number, Avatar, and Photograph are not significantly different from each other: the data are not consistent with *Prediction 1c*.

To understand what is causing these differences, I examine extreme behavior. How often do subjects contribute nothing or contribute their full endowment, and how does that vary across conditions? Table 3 reports the percentage of subject-round observations in which the contribution was zero; Table 4 reports the percentage of subject-round observations in which the entire endowment was contributed. There is no significant difference in the rates of zero contribution

	Aggregate Only	*Partial Attribution*	*Full Attribution*	Total
No Identifier	14.3%			14.3%
Number		7.2%	11.7%	9.6%
Avatar		14.6%	7.1%	10.4%
Photo		11.1%	3.8%	7.5%
Total	14.3%	11.1%	7.7%	9.9%

Table 3: Percentage of Subject-Round Observations with Zero Contribution

	Aggregate Only	*Partial Attribution*	*Full Attribution*	Total
No Identifier	16.6%			16.6%
Number		45.1%	60.9%	53.7%
Avatar		30.1%	66.4%	50.4%
Photo		41.5%	58.1%	49.8%
Total	16.6%	38.6%	62.3%	46.8%

Table 4: Percentage of Subject-Round Observations with Full Contribution

or full contribution between the three Identifiers, aggregating over Attribution schemes.[37] On the other hand, there are significant differences between the different Attribution schemes, aggregating over Identifiers. Most striking is the difference in rate of full contribution: in the *Baseline*, subjects contributed their full endowment in 16.6% of rounds. That figure rises to 38.6% in *Partial Attribution* rounds ($p = 0.048$ when compared to *Baseline*), and further to 62.3% in *Full Attribution* rounds ($p = 0.005$ when compared to *Partial Attribution*, $p < 0.001$ when compared to *Baseline*).[38] This suggests that the increase in average contributions in the treatment conditions is primarily driven by increased rates of full contribution when individual contribution choices are made common knowledge.

Coming back to the theoretical framework presented in Section 3, the experimental data are not consistent with the punishment framework. Contributions in the *PhotoFull* condition are not significantly higher than in all other conditions; in fact, contributions in *RandomFull* and *AvatarFull* are just as high, even though subjects cannot be identified outside the lab with a number or avatar. The data are also not fully consistent with the image concerns framework either. While contributions are increasing in p as predicted, contributions are not increasing in b, nor do

[37] I conduct a Mann Whitney U test to compare the rate of zero contribution or full contribution, where each group is treated as one observation. For each group, I calculate the fraction of the 128 subject-round observations (4 subjects per group play 32 rounds each) that have zero contribution or full contribution.

[38] The differences across Identifiers in rate of zero contribution were not as stark or statistically significant. The decrease in zero contribution rate from *Baseline* to *Partial Attribution* was not significant ($p = 0.267$). The decrease in zero contribution rate from *Partial Attribution* to *Full Attribution* was significant at the 10% level ($p = 0.064$). The decrease in zero contribution rate from *Baseline* to *Full Attribution* was significant at the 5% level ($p = 0.022$).

they exhibit increasing differences in (b, p).

While one can interpret the data to say that image concerns are not at play, an alternative interpretation is that b_{Number}, b_{Avatar}, and b_{Photo} are very close, if not equal, to each other—but still greater than zero. Consider a modification of the image concerns framework where, instead of assuming $b_{\text{Number}} < b_{\text{Avatar}} < b_{\text{Photo}}$, the assumption was that there is some $b > 0$ as long as any identifier exists. While such an assumption is perhaps unintuitive, it would allow the image concerns framework to fit the data better. Taken in conjunction with the Andreoni & Petrie (2004) result, these results suggest that the role of an identifier in inducing higher contributions is not to identify the physical person, but rather to provide something that an action can be attributed to, whether with certainty or probabilistically.

4.3 Dynamics of Contributions

Given the results so far, what could be driving contributions? In this section, I focus my analysis on the difference in contributions across Attribution schemes. Note that when subjects are provided individual contribution choices, they are also given more information than if they only had the aggregate contribution. For example, when a subject sees the contribution choices of their group members, they know where their contribution ranks within the group; in most cases, they could not know this with only the aggregate contribution.[39] The analysis in this section asks: *Are subjects reacting to the information provided by the individual contribution choices under Partial and Full Attribution?* In particular, I consider a purely reactive model: I assume that subjects are backward-looking and respond simply to behavior in the previous round, without anticipating future reactions.

To address this question, I analyze the dynamics of contributions over the course of the experiment: I seek to identify the factors that explain how a subject's contribution varies from one period to the next, and whether these factors differ across Attribution schemes. I focus on the data aggregated at the Attribution level (*i.e.*, pooled across Identifiers). For each level of Attribution, I separately regress player i's contribution in round t on regressors from the previous round, $t-1$. The first four regressors correspond to pieces of aggregate information. *OwnContrib* is the

[39] There are a few edge case exceptions. For example, if the aggregate contribution is zero, then it can be known that every player chose to contribute zero.

subject's private contribution, and *SumOtherContribs* is the sum of the contributions made by the other members of the subject's group. *AboveMean* (*BelowMean*) is an indicator variable for whether the subject's contribution was above (below) the mean contribution by their group members.[40] The final three regressors, on the other hand, correspond to pieces of relative information that can only be known when the individual contribution choices are also revealed. *RankInGroup* is a variable that equals the number of group members who contributed strictly more than the subject. *SomeoneElseFreeRide* (*SomeoneElseFullContrib*) is an indicator variable for whether another member of the group contributed zero (contributed their full endowment).

I estimate a dynamic panel model using the generalized method of moments, specifically the Arellano & Bond (1991) estimator. Table 5 displays the results.

As expected, one's contribution in the previous period (*OwnContrib.lag1*) has a significant, positive impact on one's contribution in the current period, and there is a significant, negative time trend in all conditions. Additionally, I find a significant, positive coefficient *RankInGroup.lag1* and a significant, negative coefficient on *SomeoneElseFreeRide.lag1* for the Partial and Full Attribution subjects. Simultaneously, this reduces the size of the positive coefficient on *BelowMean.lag1* for Partial and Full Attribution subjects. This suggests that subjects are, in fact, conditioning their future contributions not only on information they glean from the aggregate contribution, but also from the relative information they get from the individual contribution choices. Subjects significantly reduce their contribution in the subsequent period when at least one of their group members contributed nothing; yet, they significantly increase their contributions in proportion to the number of group members that contributed strictly more than them. Moreover, note that the coefficient on the final three regressors is not significant for the Baseline data. This serves as a placebo test: Baseline subjects cannot know where their contribution ranks, or whether someone else in the group free rode or contributed the full endowment. Indeed, the data suggest that these regressors do not predict contributions by Baseline subjects.

The patterns revealed in this analysis are broadly consistent with a number of models that have been used to explain contribution behavior in public goods games. Croson (2007) finds that subjects in a finitely repeated public goods game tend to "match" the median contribution from the previous round, and that such behavior is most consistent with a model of reciprocity. In my experiment,

[40] Whether the mean includes or excludes the subject's own contribution is irrelevant.

	(1) Baseline	(2) Partial	(3) Full	(4) Baseline	(5) Partial	(6) Full	(7) Baseline	(8) Partial	(9) Full
OwnContrib.lag1	0.45***	0.60***	0.60***	0.50***	0.71***	0.72***	0.48***	0.73***	0.80***
	(0.11)	(0.05)	(0.06)	(0.12)	(0.06)	(0.07)	(0.12)	(0.06)	(0.08)
SumOtherContribs.lag1	0.03	0.08***	0.01	0.00	0.04	-0.03	0.03	0.01	-0.07**
	(0.03)	(0.02)	(0.02)	(0.03)	(0.02)	(0.02)	(0.04)	(0.02)	(0.03)
AboveMean.lag1	-1.48*	-1.93***	-1.94***	-1.55*	-2.26***	-2.46***	-1.36	-2.19***	-2.46***
	(0.75)	(0.43)	(0.45)	(0.78)	(0.43)	(0.49)	(0.70)	(0.43)	(0.48)
BelowMean.lag1	1.43*	4.72***	5.68***	0.68	2.10***	1.75**	0.93	1.82***	1.57*
	(0.65)	(0.58)	(0.64)	(0.66)	(0.56)	(0.62)	(0.60)	(0.53)	(0.64)
Round	-0.09**	-0.08***	-0.07***	-0.10***	-0.07***	-0.06***	-0.10**	-0.07***	-0.05***
	(0.03)	(0.01)	(0.01)	(0.03)	(0.01)	(0.01)	(0.03)	(0.02)	(0.01)
RankInGroup.lag1				0.73	1.82***	2.35***	0.70	1.94***	2.31***
				(0.40)	(0.28)	(0.29)	(0.40)	(0.28)	(0.31)
SomeoneElseFreeRide.lag1							0.30	-1.74***	-2.56***
							(0.45)	(0.48)	(0.61)
SomeoneElseFullContrib.lag1							-0.93	-0.13	1.04
							(0.56)	(0.46)	(0.57)
Constant	5.82***	2.45*	5.94***	5.60***	1.87	5.47***	5.03**	2.83*	6.25***
	(1.50)	(1.05)	(1.27)	(1.55)	(1.12)	(1.35)	(1.67)	(1.11)	(1.36)
Observations	1200	3720	4320	1200	3720	4320	1200	3720	4320
Number of Subjects	40	124	144	40	124	144	40	124	144

Robust standard errors in parentheses. Estimation performed using the *xtabond* command in STATA.

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 5: Arellano-Bond Estimation.

I find that subjects who contribute more than the mean tend to reduce their contribution in the next period, while subjects who contribute less than the mean tend to increase their contribution. This behavior could also be interpreted in two other ways. First, this behavior could be taken as evidence of backward-looking inequality aversion: if subject derive disutility from contributing more (less) than the average group member, they would adjust their subsequent contribution down (up). Such behavior is consistent *prima facie* with the model in Fehr & Schmidt (1999), though one key reversal in my data is that subjects who contribute less (earn more) than the mean respond much more strongly than subjects who contribute more (earn less) than the mean. The second interpretation of the contribution dynamics is that subjects use the feedback about individual contribution choices in order to learn and establish a norm for the group: this is particularly noticeable under *Partial Attribution* and *Full Attribution*, when subjects who contribute below the mean increase their contribution by a greater amount than by which subjects who contribute above the mean reduce their contribution.

Interestingly, the aforementioned effects are stronger under under *Full Attribution* than under *Partial Attribution*. Image concerns are one explanation of this: recall that, while the data are inconsistent with the image concerns framework for different Identifiers, the data are consistent with the image concerns framework regarding different Attribution schemes. One reading of the data, considering norm-setting and image concerns together, says that the effect observed under *Partial Attribution* captures the extent to which subjects use relative information to learn and establish a norm for the group. The additional effect observed moving from *Partial Attribution* to *Full Attribution* captures the image concerns associated with being seen as a relatively low contributor (*RankInGroup.lag1*) or as a "sucker" who contributed when someone else free rode (*SomeoneElseFreeRide.lag1*).

4.4 Avatars

In this section, I analyze two measures of subjects' connectedness to their avatars, and whether those measures are predictive of contribution behavior in the first round of the public goods game.[41] Additionally, I analyze whether subjects create avatars that resemble their physical persons. The

[41] Because contribution decisions in later rounds can be influenced by the contribution decisions in preceding rounds, I restrict the analysis to the first round only.

Figure 6: Avatars created by subjects.

first measure of connectedness is the amount of time the subject spent on customizing their avatar, and the second measure is the subject's self-reported intensity of feeling represented by their avatar. Figure 6 shows the avatars created by the 100 subjects randomly assigned to the Avatars condition.

While all subjects began customizing their avatars at the same time, each subject could decide for themselves when they were finished customizing their avatar by clicking a button at the bottom of the avatar customization screen; only when all subjects finished customizing their avatars did the public goods game begin. I calculate the amount of time spent by each subject on customizing their avatar as the difference (in seconds) between the time they clicked their button and the time they entered the avatar customization screen. Across the 100 subjects that were assigned to the Avatars condition, the mean amount of time spent on customizing an avatar was just over

3 minutes (189 seconds), with a minimum of 53 seconds and a maximum of 353 seconds.[42] The was no significant difference in customization time across Attribution: under Partial Attribution, average customization time was 188 seconds; under Full Attribution, average customization time was 190 seconds. However, the time spent on avatar customization does not have a significant effect on contributions: regressing own contribution in the first round on customization time and a Full Attribution dummy variable (and thus controlling for the Attribution scheme) yields a coefficient on customization time that is not significantly different from zero ($\beta = 0.005,\ p = 0.507$).

After the conclusion of the public goods game (but before being told their final earnings), subjects self-report their race, gender, and ethnicity. Additionally, they rate their agreement with the following statement: *I felt like my avatar represented me.*[43] The rating was done on a 7 point Likert scale, where 1 was *Strongly Disagree*, 4 was *Neutral*, and 7 was *Strongly Agree*.[44] On average, subjects in the *Partial Attribution* condition agreed more with the statement (average rating = 3.88) than the subjects in the *Full Attribution* condition (average rating = 3.38), but this difference is not significant ($p = 0.221$ from a Mann Whitney U test). There is also no correlation between the amount of time a subject spent on their avatar and their representation rating.

Does the reported level of representation correlate with first round contributions? The evidence is, at best, very weak. Under *Partial Attribution*, I can reject the null hypothesis of independence at the 10% level.[45] I cannot reject independence under *Full Attribution*.

Do subjects create avatars that resemble their physical persons? I analyze the similarity of avatar to physical person in two ways. First, I compare the subject's self-reported race and ethnicity to the skin tone of their self-designed avatar.[46] Second, I manually code each avatar with a gender (male, female, neutral) and compare the avatar's gender to the self-reported gender of the subject who created that avatar. On both dimensions, there is evidence that some subjects create avatars

[42] While subjects were instructed that they had five minutes to finish customizing their avatars, they were allowed a bit of flexibility: subjects were given a "2 minutes remaining" announcement, followed by a "1 minute remaining" announcement, followed by an announcement asking subjects still customizing their avatar to wrap up. This ensured that each subject entered the game with an avatar that they were satisfied with.

[43] Due to a server problem, the demographic data and agreement with representation statement was not collected for one session. The treatment in that session was *AvatarFull* and 16 subjects participated in that session.

[44] Subjects saw only the labels. They were not shown the numbers 1 through 7.

[45] The Spearman rank correlation coefficient between reported level of representation and level of contribution in the first round for Partial Attribution subjects is 0.28; the null hypothesis of Independence is rejected at the 10% level ($p = 0.064$). Because the reported level of representation is a categorical variable, I use the Spearman correlation coefficient instead of the Pearson correlation coefficient.

[46] The available skin tone options, from lightest to darkest, were: *Pale*, *Light*, *Yellow*, *Tanned*, *Brown*, *DarkBrown*, *Black*.

Figure 7: Avatars created by subjects who self-identified as Black women.

that don't resemble their physical person. Only one subject chose the *Black* skin tone for their avatar, yet that subject self-identified as non-Hispanic White. There is a similar disparity between genders of avatars and the subjects who made them. Of seventeen subjects who self-identified as male, five designed avatars that were coded as female. Of 64 subjects who self-identified as female, six designed avatars that were coded as male and thirteen designed avatars that were coded as neutral gender. At the intersection of race and gender, five subjects self-identified as Black women. Only three of those five subjects created avatars that coded as female, and none of those subjects chose the *Black* skin tone for their avatars: Figure 7 presents the five avatars created by these five subjects.

5 Conclusions

As our interactions shift increasingly to the digital world, it has become important to understand how its peculiarities affect our behavior. The digital world allows individuals to identify themselves with a wider range of identifiers than is available to individuals in the physical world. Moreover, the digital world allows interactions among individuals without identifiers at all, which further complicates how digital actions are attributed to the physical persons behind the screen.

The experimental data confirm that identifying individuals and providing information about individual actions work in conjunction to increase contributions in a public goods game. Surprisingly, however, the positive effect persists when actions are attributed to identifiers that are unable to identify the physical person, when actions can only be probabilistically attributed to the physical person, and even when actions can only be probabilistically attributed to identifiers that are unable to identify the physical person.

Pitting the two channels against each other, on the hand, the data show that variations in the Attribution scheme have a larger effect on contributions than variation in Identifier. This difference is driven by higher rates of subjects contributing their full endowment under *Full Attribution* than under *Partial Attribution* than under *Baseline*. Dynamic estimation of the contribution decisions suggests that subjects react to "relative" information when available: as subjects proceed through the game, they attempt to figure out, establish, and follow a norm. Indeed, many *Full Attribution* groups successfully coordinate on contributing their entire endowment, even without a coordinating device or central authority.[47]

Interestingly, much of the existing policies targeting anonymity in digital settings have revolved around users having identifiers that don't pin down their physical persons. Many online communities, from Facebook[48] to Statalist (the official Stata forum)[49], require users to use their "real" names as their identifiers. Yet, these policies have faced substantial resistance: for instance, Facebook has gotten pushback about its "real name" policy from users of various ethnicities, from transgender users, from users who needed their identities protected.[50] [51] [52] [53] However, the results from this experiment suggest that tech companies designing the digital world, and policymakers seeking to regulate it, may do better to emphasize a system where a user's history of actions is logged and attributed to some identifier, as opposed to a "real name" policy where users must use their real name.

The digital world is still young, and there remain many decisions to be made about what it will look like. Given the myriad anecdotes of troubling online behavior, inducing individuals to behave prosocially should be a priority in the design of digital spaces where many people gather and interact. Although much remains to be studied, the results of this paper suggests a path forward to ensuring that people do good in the digital world.

[47] This is particularly encouraging for proponents of decentralized governance in digital settings.

[48] "Facebook is a community where everyone uses the name they go by in everyday life." https://www.facebook.com/help/112146705538576.

[49] "You are asked to post on Statalist using your full real name [...] . Giving full names is one of the ways in which we show respect for others and is a long tradition on Statalist. It is also much easier to get to know people when real names are used." https://www.statalist.org/forums/help.

[50] https://www.telegraph.co.uk/news/newstopics/howaboutthat/2632170/Woman-called-Yoda-blocked-from-Facebook.html

[51] https://www.theguardian.com/technology/2015/feb/16/facebook-real-name-policy-suspends-native-americans

[52] https://www.cnn.com/2014/09/16/living/facebook-name-policy/index.html

[53] https://www.huffpost.com/entry/pakistans-religious-extremism_b_9577338

References

Abraham, Diya, Greiner, Ben, & Stephanides, Marianne. 2023. On the Internet You Can Be Anyone: An Experiment on Strategic Avatar Choice in Online Marketplaces. *Journal of Economic Behavior & Organization*, **206**(Feb.), 251–261.

Andreoni, James. 1989. Giving with Impure Altruism: Applications to Charity and Ricardian Equivalence. *Journal of Political Economy*, **97**(6), 1447–1458.

Andreoni, James. 1990. Impure Altruism and Donations to Public Goods: A Theory of Warm-Glow Giving. *The Economic Journal*, **100**(401), 464–477.

Andreoni, James, & Petrie, Ragan. 2004. Public Goods Experiments without Confidentiality: A Glimpse into Fund-Raising. *Journal of Public Economics*, **88**(7), 1605–1623.

Arellano, Manuel, & Bond, Stephen. 1991. Some Tests of Specification for Panel Data: Monte Carlo Evidence and an Application to Employment Equations. *The Review of Economic Studies*, **58**(2), 277–297.

Ariely, Dan, Bracha, Anat, & Meier, Stephan. 2009. Doing Good or Doing Well? Image Motivation and Monetary Incentives in Behaving Prosocially. *The American Economic Review*, **99**(1), 544–555.

Bénabou, Roland, & Tirole, Jean. 2006. Incentives and Prosocial Behavior. *American Economic Review*, **96**(5), 1652–1678.

Bente, Gary, Rüggenberg, Sabine, Krämer, Nicole C., & Eschenburg, Felix. 2008. Avatar-Mediated Networking: Increasing Social Presence and Interpersonal Trust in Net-Based Collaborations. *Human Communication Research*, **34**(2), 287–318.

Bente, Gary, Dratsch, Thomas, Kaspar, Kai, Hässler, Tabea, Bungard, Oliver, & Al-Issa, Ahmad. 2014. Cultures of Trust: Effects of Avatar Faces and Reputation Scores on German and Arab Players in an Online Trust-Game. *PLOS ONE*, **9**(6), e98297.

Charness, Gary, & Gneezy, Uri. 2008. What's in a Name? Anonymity and Social Distance in Dictator and Ultimatum Games. *Journal of Economic Behavior & Organization*, **68**(1), 29–35.

Charness, Gary, Cobo-Reyes, Ramón, Meraglia, Simone, & Sánchez, Ángela. 2020. Anticipated Discrimination, Choices, and Performance: Experimental Evidence. *European Economic Review*, **127**(Aug.), 103473.

Chaudhuri, Ananish. 2011. Sustaining Cooperation in Laboratory Public Goods Experiments: A Selective Survey of the Literature. *Experimental Economics*, **14**(1), 47–83.

Chen, Daniel L., Schonger, Martin, & Wickens, Chris. 2016. oTree—An Open-Source Platform for Laboratory, Online, and Field Experiments. *Journal of Behavioral and Experimental Finance*, **9**(Mar.), 88–97.

Croson, Rachel. 2001. Feedback in Voluntary Contribution Mechanisms: An Experiment in Team Production. *Research in Experimental Economics*, **8**(Jan.), 85–97.

Croson, Rachel T. A. 2007. Theories of Commitment, Altruism and Reciprocity: Evidence from Linear Public Goods Games. *Economic Inquiry*, **45**(2), 199–216.

Denant-Boemont, Laurent, Masclet, David, & Noussair, Charles N. 2007. Punishment, Counterpunishment and Sanction Enforcement in a Social Dilemma Experiment. *Economic Theory*, **33**(1), 145–167.

Dufwenberg, Martin, & Muren, Astri. 2006. Generosity, Anonymity, Gender. *Journal of Economic Behavior & Organization*, **61**(1), 42–49.

Falk, Armin. 2021. Facing Yourself – A Note on Self-Image. *Journal of Economic Behavior & Organization*, **186**(June), 724–734.

Fehr, Ernst, & Gächter, Simon. 2000. Cooperation and Punishment in Public Goods Experiments. *The American Economic Review*, **90**(4), 980–994.

Fehr, Ernst, & Gächter, Simon. 2002. Altruistic Punishment in Humans. *Nature*, **415**(6868), 137–140.

Fehr, Ernst, & Schmidt, Klaus M. 1999. A Theory of Fairness, Competition, and Cooperation. *The Quarterly Journal of Economics*, **114**(3), 817–868.

Fiedler, Marina, & Haruvy, Ernan. 2009. The Lab versus the Virtual Lab and Virtual Field—An Experimental Investigation of Trust Games with Communication. *Journal of Economic Behavior & Organization*, **72**(2), 716–724.

Frenkel, Sheera, & Browning, Kellen. 2021. The Metaverse's Dark Side: Here Come Harassment and Assaults. *The New York Times*, Dec.

Gächter, Simon, & Fehr, Ernst. 1999. Collective Action as a Social Exchange. *Journal of Economic Behavior & Organization*, **39**(4), 341–369.

Greiner, Ben. 2015. Subject Pool Recruitment Procedures: Organizing Experiments with ORSEE. *Journal of the Economic Science Association*, **1**(1), 114–125.

Groom, Victoria, Bailenson, Jeremy N., & Nass, Clifford. 2009. The Influence of Racial Embodiment on Racial Bias in Immersive Virtual Environments. *Social Influence*, **4**(3), 231–248.

Groves, Theodore, & Ledyard, John. 1977. Optimal Allocation of Public Goods: A Solution to the "Free Rider" Problem. *Econometrica*, **45**(4), 783–809.

Isaac, R. Mark, Walker, James M., & Thomas, Susan H. 1984. Divergent Evidence on Free Riding: An Experimental Examination of Possible Explanations. *Public Choice*, **43**(2), 113–149.

Jones, Michael, & McKee, Michael. 2004. Feedback Information and Contributions to Not-for-Profit Enterprises: Experimental Investigations and Implications for Large-Scale Fund-Raising. *Public Finance Review*, **32**(5), 512–527.

Karlan, Dean, & McConnell, Margaret A. 2014. Hey Look at Me: The Effect of Giving Circles on Giving. *Journal of Economic Behavior & Organization*, **106**(Oct.), 402–412.

Ledyard, John O. 1995. Public Goods: A Survey of Experimental Research. *In:* Kagel, John H., & Roth, Alvin E. (eds), *The Handbook of Experimental Economics*. Princeton, NJ: Princeton University Press.

Lim, Sohye, & Reeves, Byron. 2009. Being in the Game: Effects of Avatar Choice and Point of View on Psychophysiological Responses During Play. *Media Psychology*, **12**(4), 348–370.

Lugovskyy, Volodymyr, Puzzello, Daniela, Sorensen, Andrea, Walker, James, & Williams, Arlington. 2017. An Experimental Study of Finitely and Infinitely Repeated Linear Public Goods Games. *Games and Economic Behavior*, **102**(Mar.), 286–302.

Martey, Rosa Mikeal, & Consalvo, Mia. 2011. Performing the Looking-Glass Self: Avatar Appearance and Group Identity in Second Life. *Popular Communication*, **9**(3), 165–180.

Meier, Stephan. 2007. A Survey of Economic Theories and Field Evidence on Pro-Social Behavior. *Pages 51–87 of: Economics and Psychology: A Promising New Cross-Disciplinary Field.* CESifo Seminar Series. Cambridge, MA, US: MIT Press.

Nikiforakis, Nikos. 2010. Feedback, Punishment and Cooperation in Public Good Experiments. *Games and Economic Behavior*, **68**(2), 689–702.

Ostrom, Elinor, Walker, James, & Gardner, Roy. 1992. Covenants With and Without a Sword: Self-Governance Is Possible. *The American Political Science Review*, **86**(2), 404–417.

Pew Research Center. 2021. The State of Online Harassment.

Rege, Mari, & Telle, Kjetil. 2004. The Impact of Social Approval and Framing on Cooperation in Public Good Situations. *Journal of Public Economics*, **88**(7), 1625–1644.

Samek, Anya, & Sheremeta, Roman M. 2014. Recognizing Contributors: An Experiment on Public Goods. *Experimental Economics*, **17**(4), 673–690.

Sell, Jane, & Wilson, Rick K. 1991. Levels of Information and Contributions to Public Goods*. *Social Forces*, **70**(1), 107–124.

Shang, Jen, & Croson, Rachel. 2009. A Field Experiment in Charitable Contribution: The Impact of Social Information on the Voluntary Provision of Public Goods. *The Economic Journal*, **119**(540), 1422–1439.

Soetevent, Adriaan R. 2005. Anonymity in Giving in a Natural Context—a Field Experiment in 30 Churches. *Journal of Public Economics*, **89**(11), 2301–2323.

Vasalou, Asimina, & Joinson, Adam N. 2009. Me, Myself and I: The Role of Interactional Context on Self-Presentation through Avatars. *Computers in Human Behavior*, **25**(2), 510–520.

Weimann, Joachim. 1994. Individual Behaviour in a Free Riding Experiment. *Journal of Public Economics*, **54**(2), 185–200.

Wilson, Rick K., & Sell, Jane. 1997. "Liar, Liar...": Cheap Talk and Reputation in Repeated Public Goods Settings. *The Journal of Conflict Resolution*, **41**(5), 695–717.

Yee, Nick, & Bailenson, Jeremy. 2007. The Proteus Effect: The Effect of Transformed Self-Representation on Behavior. *Human Communication Research*, **33**(3), 271–290.

Appendix A

Image Concerns

Recall that all agents have a utility function of the following form: $U_i(g_i, g_{-i}) = \pi_i(g_i, g_{-i}) + (1 + b)W(g_i) + bpR(g_i)$. The following Proposition characterizes the Nash equilibrium contribution profile:

Proposition. The Nash equilibrium contribution profile of the one-shot game is given by $(g_1^*, ..., g_N^*)$, where for $i \in \{1, ..., N\}$:

$$g_i^* = \begin{cases} 0 & \text{if } 1 - \frac{m}{N} > \frac{\partial W}{\partial g_i} + b\frac{\partial W}{\partial g_i} + bp\frac{\partial R}{\partial g_i} \ \forall g_i \in [0, Y] \\ g_i & \text{such that } 1 - \frac{m}{N} = \frac{\partial W}{\partial g_i} + b\frac{\partial W}{\partial g_i} + bp\frac{\partial R}{\partial g_i} \\ Y & \text{if } 1 - \frac{m}{N} < \frac{\partial W}{\partial g_i} + b\frac{\partial W}{\partial g_i} + bp\frac{\partial R}{\partial g_i} \ \forall g_i \in [0, Y] \end{cases} \tag{3}$$

Additionally, if all agents are symmetric, then $g_1^* = ... = g_N^*$.

Proof. The condition in the expression for g_i^* comes from the first-order condition of the agent's utility maximization problem. On the left-hand side is the marginal payoff from not contributing a marginal unit to the public good: $1 - \frac{m}{N}$. On the right-hand side is the image utility gain from contributing a marginal unit to the public good: $\frac{\partial W}{\partial g_i} + b\frac{\partial W}{\partial g_i} + bp\frac{\partial R}{\partial g_i}$. If interior, g_i^* equates the two sides. Edge cases occur if one side is always larger than the other: if the marginal payoff from not contributing always exceeds the image gain from contributing, then $g_i^* = 0$ is the utility-maximizing contribution. A similar argument justifies when $g_i^* = Y$.

Punishment outside the Game

The punishment game is played with N players indexed by $i \in \{1, ..., N\}$. The action space is given by $A = \{P, X\}^N$. Each player's action profile is a N-dimensional vector $a_i = (a_{i1}, a_{i2}, ..., a_{iN})$, where $a_{ij} \in \{P, X\}$ denotes whether player i punishes (P) or does not punish (X) player j. There is a cost $c > 0$ to punish another player, as well as a cost $C >> c$ if punished by another player.

The payoff in the stage game is given by $\pi_i = -c\sum_{j=1}^{N}\mathbb{I}(a_{ij}=P) - C\sum_{j=1}^{N}\mathbb{I}(a_{ji}=P)$.[54]

The punishment game is repeated infinitely in discrete time, with discount factor $\delta \in (0,1)$. I assume there is perfect monitoring in the punishment game: at the end of each period, all players observe each other's action profiles. In other words, all players know who punishes who in each period. The following Proposition characterizes the relationship between the level of anonymity in a public goods game and the expected punishment from free riding in the subsequent punishment game. The intuition is straightforward: when a deviator's identity can be known, the other group members are able to coordinate punishment on the deviator. When the deviator's identity cannot be known, the deviator escapes punishment with strictly positive probability. Hence, the expected punishment from deviating is strictly lower.

Proposition. Let a finitely repeated public goods game be followed by an infinitely repeated punishment game. There exists some discount factor $\delta \in (0,1)$ that can sustain positive contributions in the public goods game (1) a deviation can be attributed with certainty to an identifier that (2) allows the deviator to be identified outside the game, but cannot sustain positive contributions otherwise. The reverse does not hold.

Proof. I begin by restricting focus to the scenario when it is common knowledge that there exists a "target" for punishment.[55] The proof then proceeds in two parts. First, I consider the case when the target's identity is also commonly known. For this case, I construct a subgame perfect Nash equilibrium in which the target is punished indefinitely by all other group members. Second, I consider the case where the target's identity cannot be commonly known. I show that in any equilibrium, the target escapes punishment with strictly positive probability.

I assume in the proof, without loss of generality, that the target is Player 1.

Part 1. Suppose that it is common knowledge among all players that Player 1 is the target: Player 1's actions were attributed with certainty to an identifier during the public goods game that allowed their group members to identify them during the punishment game. Consider the following

[54]The unique stage game Nash equilibrium is given by $a_{ij} = X$ for all i, j. Nobody punishes, and all players earn payoff $\pi = 0$. This is because punishment is costly to administer, and to incur.

[55]Given some strategy in the public goods game, players can deduce just from the aggregate contribution to the public good, whether a group member deviated. In all experimental conditions, the aggregate contribution is made known after every round.

set of strategies in which punishment is coordinated on the target:

$$\begin{cases} a_1 & = (X, X, ..., X) \\ a_i & = (P, X, ..., X), \ \forall i \in \{2, ..., N\} \end{cases}$$

This set of strategies can be sustained as a subgame perfect Nash equilibrium in the infinitely repeated game, so long as there is no profitable one-shot deviation.

It is straightforward to see that Player 1 (the target) has no profitable deviation from their strategy.

Without loss of generality, consider whether Player 2 has a profitable deviation. Suppose that Player 2 deviates from their strategy, and does not punish the target in a given period. In that period, Player 2 has a payoff of $0 > -c$. But starting in the following period, all other players $i \in \{1, 3, ..., N\}$ can punish Player 2 forever for deviating; in effect, Player 2 would become the new target. Therefore, Player 2 does not have a profitable deviation if: $-\frac{c}{1-\delta} \geq 0 - \frac{\delta}{1-\delta}(N-1)C$. Rearranging terms yields: $\delta \geq \frac{c}{(N-1)C}$. Because $\frac{c}{(N-1)C} < 1$, this set of strategies that coordinates punishment on the target can be sustained as a subgame perfect Nash equilibrium in the infinitely repeated game.

Note that when the target's identity is common knowledge, the target receives payoff $\pi_1 = -(N-1)C$ each period.

Part 2. Suppose the target's identity is not common knowledge. The only way in which the players can guarantee that the target receives payoff $-(N-1)C$ per period is if everyone punishes everyone else: $a_{ij} = P, \ \forall i \neq j$. However, this cannot be sustained in equilibrium: a player may profitably deviate by not punishing, and they cannot be punished for the deviation since they are already being punished by everyone else. Therefore, in any equilibrium of the game when the target's identity is not common knowledge, the target's expected payoff is strictly greater (*i.e.*, less negative) than $-(N-1)C$.

Denote the expected punishment from free riding as $-P_{FR}^{K} = -(N-1)C$ when the deviator's identity is commonly known and as $-P_{FR}^{UK}$ when the deviator's identity cannot be commonly known. The expected punishment from free riding in the public goods game is strictly smaller when there is sufficient anonymity to mask the deviator's identity: $-P_{FR}^{K} < -P_{FR}^{UK}$. Hence, the minimum

discount rate δ that can support positive contributions satisfies: $\delta^K < \delta^{UK}$. Therefore, there exists a δ that supports positive contributions when the deviator's identity can be commonly known that cannot support positive contributions when the deviator's identity cannot be commonly known.

Appendix B

Experimental Instructions for the *NumberPartial* Treatment

Thank you for participating in this experiment.

Please read and sign the consent form in front of you. Raise your hand if you have any questions.

In this experiment, you will make decisions in groups.

You will earn points throughout the course of the experiment.
How many points you earn depends both on your decisions and your group members' decisions.

For being here, you are guaranteed to earn at least $5.

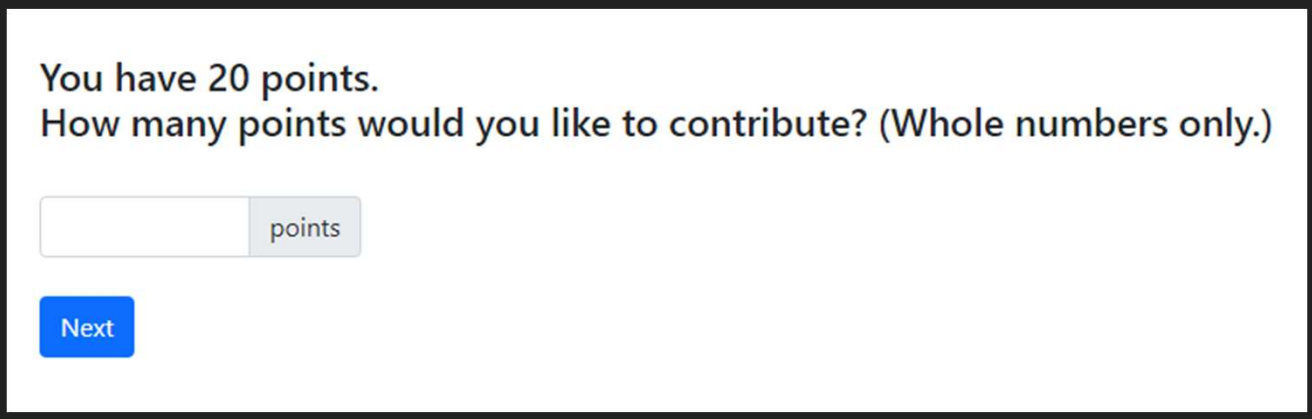
Your exact monetary earnings will depend on how many points you earn during the experiment.

At the start of this experiment, you will be randomly assigned to a group of 4 people (including yourself).

You will stay in this group for the entire experiment.

At the start of each round, you will be endowed with 20 points.

In each round, you must decide how many of these 20 points to Contribute to a group account. Any points you don't Contribute, you Keep for yourself.



All contributions to the group account are **doubled**.

For example: if you Contribute 1 point, it becomes 2 points in the group account.

At the end of each round, the group account is evenly split among all four members of the group.

The number of points you earn each round is equal to:

1. How many points you Keep

   PLUS

2. Your Share of the Group Account

In other words, the number of points you earn in a round equals:

(20 - # points you Contribute)

PLUS

¼ * 2 * (# points you Contribute

+ # points your group members Contribute)

If you, and everyone else in your group, chooses to **Contribute 10 points**:

You Keep: **10** points

Your Share of Group Contribution:

¼ * 2 * (10+10+10+10) = **20** points

You Earn: 10 + 20 = **30** points

If you, and everyone else in your group, chooses to **Contribute 0 points**:

You Keep: **20** points

Your Share of Group Contribution:

¼ * 2 * (0+0+0+0) = **0** points

You Earn: 20 + 0 = **<u>20</u>** points

If everyone else in your group chooses to **Contribute 10 points** and you choose to **Contribute 12 points**:

You Keep: **8** points

Your Share of Group Contribution:

¼ * 2 * (12+10+10+10) = **21** points

You Earn: 8 + 21 = **<u>29</u>** points

If everyone else in your group chooses to **Contribute 10 points** and you choose to **Contribute 12 points**:

You Keep: **8** points

Your Share of Group Contribution:

¼ * 2 * (12+10+10+10) = **21** points

You Earn: 8 + 21 = **29** points

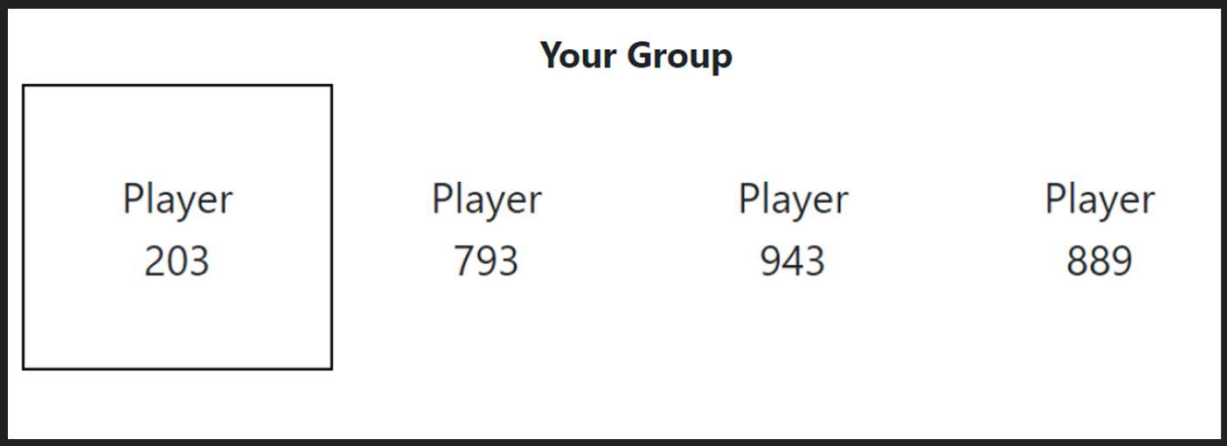
(Your Group Members Earn: 10 + 21 = 31 points)

At the start of the experiment, every player will be randomly assigned a three-digit number.

No two players will have the same number.

Only you will know what your number is.

Each round, you will see your number as well as your group members' numbers on your screen.



After each round, you will see a summary of your earnings from that round.

After each round, the amount each group member Contributes (including yourself) will appear from largest to smallest.

Contributions from Largest to Smallest

12 points, 11 points, 10 points, 9 points

The order of your group members is independent of the order of Contributions.

Player 203	Player 793	Player 943	Player 889

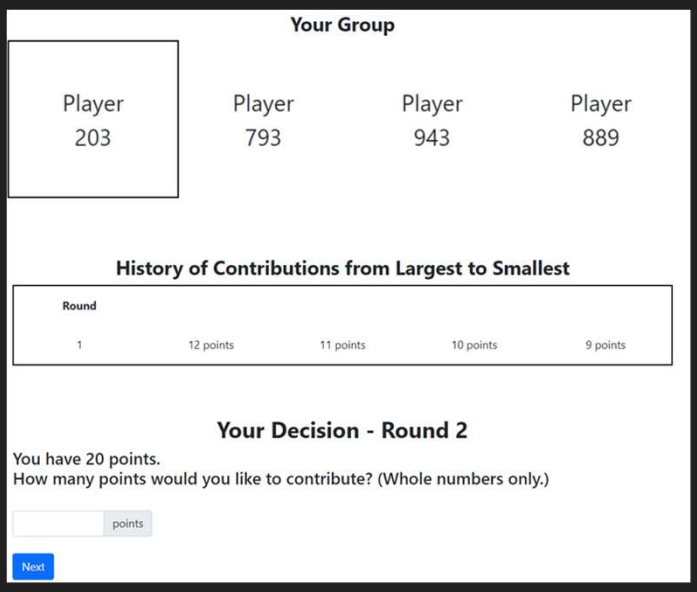
Contributions from Largest to Smallest

12 points, 11 points, 10 points, 9 points

You are not obligated to tell anybody what your number is, even after the experiment is over.

Nobody else, not even the experimenter, will be able to link your number back to you.

In later rounds, the computer will remind you of your group's Contributions in order from largest to smallest, from all previous rounds.



There will be **32 rounds** in total in this experiment, all of which proceed as previously described.

Your monetary earnings at the end of the experiment depend on the total number of points you earn in all 32 rounds.

Every 50 points you earn equals $1. (1 point = 2 cents)

SUMMARY

Same group of 4 people

Endowment: 20 points per round

Contributions are **doubled** in the group account

32 rounds in total

50 points = 1 dollar