

Moral Licensing: Prosocial Behaviors in Public and Private Spheres

Fuhai Hong* Jean Tirole† Chen Zhang‡

March 15, 2024

Abstract: This paper aims at understanding, theoretically and empirically, how image concerns in the public and private spheres link moral choices that are otherwise unrelated. The paper provides the first formalization of moral licensing: Moral behavior in the public sphere “motivates” more selfish behavior in the private sphere. Moreover, a larger public sphere leads to lower prosociality in both public and private spheres, due to more expensive signaling in the public sphere and increasing moral licensing in the private sphere, respectively. Overall, giving a socially-valued behavior more visibility does not necessarily make it more prevalent. To identify the moral-licensing mechanism, we match a laboratory experiment with the model. The experiment confirms the theoretical predictions.

Keywords: Prosocial behavior, multitask signaling, public and private spheres, transparency, moral licensing.

JEL numbers: D9, D64, D80, K38.

*Department of Economics, Lingnan University, Hong Kong; email: fuhaihong@ln.edu.hk.

†Corresponding author. Toulouse School of Economics and Institute for Advanced Study in Toulouse; e-mail: jean.tirole@tse-fr.eu.

‡Economics and Management School, Wuhan University, China; email: zhangchen34349@whu.edu.cn.

1 Introduction

AI, ratings, facial recognition, the recording of online and public-space interactions, and data externalities¹ make our life more and more exposed to public view, and behaviors that once belonged to our private sphere have become observable by a larger audience.² The experimental literature’s demonstration that we change our behavior when observed by others whose judgment we value suggests that the technological revolution will alter the nature of our social relationships.³ It is easy to predict that the higher visibility of some behaviors and the concomitant increase in social pressure in the corresponding interactions will raise reputational concerns. This paper offers theoretical predictions and experimental evidence on how relationships in our private sphere and the overall prosociality are likely to be affected.

To locate our contribution within the literature, reminisce the extensive lab-and-field evidence that increased visibility induces more moral behavior in a wide variety of contexts, from charitable contributions to public goods provision, voting, health and blood donations.⁴ This evidence can be summarized in the following assertion: “Giving a socially-valued behavior more visibility makes it more prevalent.” To grasp why this conclusion cannot be transported to a multitasking environment, suppose that the individual’s trait to be signaled (say, prosociality) is correlated or the same across activities. While increased visibility of behavior in one activity bolsters prosociality in that activity, the behavioral information educates the audience as to the underlying trait, reducing the scope for signaling in other activities. Thus, making an activity more visible generates both crowding-in and crowding-out in a multi-task environment.

Our objective in this paper is to understand, theoretically and empirically, the interdependency between moral choices. The paper shows that this interdependency hinges on two effects, the “cheap-signaling effect” and the “moral licensing effect”; and that a larger public sphere reduces ethical conduct in the private and public spheres, and may lead to a lower overall prosociality. It then provides model-based evidence to this effect.

¹The sharing by others of information about us on social media, blogs or e-mails.

²“Public and private spheres” and “transparency” in this paper will have their expected meanings. The *private sphere* will cover actions that are directly observed by a limited set of individuals. This limited observability often characterizes family, friendship and stable work relationships. The *public sphere* by contrast refers to actions observed by a much broader group of onlookers, through public-place behavior, ratings, facial recognition, AI analytics, word-of-mouth, or social networks. How public my behavior is hinges on the number of people who observe my actions and on how much I care about their opinion. *Transparency*, a reduction in privacy, is the policy through which a given behavior is made more visible to others.

³Vindicating Jaron Lanier (2010) in his assertion that “The most important thing about a technology is how it changes people.”

⁴See e.g., Ashraf-Bandiera (2018) and Bursztyn-Jensen (2017) for overviews of this literature. References include Freeman (1997), Ariely et al. (2009) for charitable contributions, Algan et al. (2016) for public goods provision, Funk (2010), DellaVigna et al. (2017), Perez-Truglia and Cruces (2017) for voting, Ashraf et al. (2014) for health, and Lacetera et al. (2012) for blood donations. There is also a large experimental literature that manipulates the subjects’ self-image concerns and leads to the same conclusion.

The theory

Section 2.1 describes a standard, single-task consensual-behavior model, in which reputational concerns help motivate an agent. The latter takes an action that exerts an externality onto others. The behavior is consensual in that the audience and the agent agree on what constitutes morally proper behavior. The agent is motivated by (a) her true empathy (or altruism, proclivity for doing good, internalization of others' well-being), (b) her extrinsic motivation, and (c) her image concerns, associated with the inferences that others will draw from her behavior. As is well-known,⁵ this benchmark model exhibits underprovision of prosocial behavior, provided that the externality is not too small.

Section 2.2 then analyzes a novel multi-task extension of the standard model, that exhibits a coexistence between private and public spheres. Behaviors in the private sphere are observed solely through direct interaction. Behaviors in the public sphere by contrast are the object of public disclosure. The individual cares about the sum of their reputations with audience members; such linear reputation concerns do not condition the insights but simplify the exposition. There is a two-way interaction between the private and public spheres. First, when behaving prosocially vis-à-vis a public-sphere partner, the agent receives a double dividend: she ingratiates herself with the partner, and she further earns brownie points from third-party onlookers who observe her behavior. Therefore, the presence of a private sphere makes behavior more moral in the public sphere. This “cheap-signaling effect” is but an extension of the familiar observation that prosociality is encouraged by a widening of the audience. Second, and a novel effect, prosocial behavior in the public sphere signals a minimal level of individual prosociality, and thereby makes it less costly to behave asocially in the private sphere. This is, to the best of our knowledge, the first formalization of the “moral licensing effect” that is prominent in psychology.⁶

Furthermore, as the public sphere expands, the number of widely observed decisions increases and signaling in the public sphere becomes less cheap. A prosocial behavior in the public sphere therefore brings more distinction. This increased distinction in turn inflates the image cushion the subject benefits from when not behaving well in the private sphere and thus raises the intensity of moral licensing.

Two main insights emerge, that will be the object of experimental testing. First, prosocial activities, regardless of their overall level, are misallocated, with too much attention

⁵Eg. Acquisti et al. (2016), Ali-Bénabou (2020), Bénabou-Tirole (2006) and Daughety-Reinganum (2010).

⁶Moral licensing is the phenomenon “whereby increased confidence and security in one’s self-image or self-concept tends to make that individual worry less about the consequences of subsequent immoral behavior and, therefore, more likely to make immoral choices and act immorally.” (Wikipedia). For example, Monin-Miller (2001) shows in their groundbreaking article that when people are made to behave initially in a moral way, they are more likely to display behaviors that are unethical. Efron et al. (2009) show that voicing support for Obama in 2008 may license people to make ambiguously racist comments. Relatedly, Zhong-Liljenquist (2006) and Gneezy et al (2014) develop a notion of “moral cleansing”, according to which people who first make an immoral choice are then more likely to try to redeem themselves through moral deeds. See also Merritt et al. (2012) and Efron et al. (2012), and the broader review in Section 4.

paid to the public sphere/too little to the private one. The agent behaves better in the public sphere than in the “all public” or “all private” benchmarks; the converse holds for behavior in the private sphere. Second, the public sphere crowds out the private one. Actually, an expansion in the public sphere (due, say, to technological change) reduces prosociality in both spheres and, despite the composition effect that favors prosociality, reduces overall prosociality over some range. The overall picture is one of public sphere dominance and disintegration of the social fabric in the private sphere.

Finally, we obtain a few complementary theoretical insights. Drivers that shift behavior in the private sphere (say, the price of good behavior in that sphere) make contributions in the two spheres covary negatively, while those that shift behavior in the public sphere (say, the price of good behavior in that sphere, or an expansion in the size of the public sphere that does not impact the size of the private sphere) make contributions in the two spheres covary positively. We also show that the moral-licensing (resp. moral-cleansing) experiments that have been conducted in psychology (a) can be formalized within our model and (b) in contrast with the public-and-private-spheres application, involve a moral-licensing (resp. moral-cleansing) effect, but no cheap-signaling effect.

The experiment

In our theory, image concerns are the key link between moral choices that are otherwise unrelated. In general, image payoffs may be linear or positional (prestige is relative, i.e., acquired at the expense of others), concave (as in Butera et al 2022), or – at least locally – convex (as in winner-take-all contests). A failure to control for the shape of the image reward makes it hard to identify the central feature of our study, the image spillovers across decisions. To be able to identify the moral-licensing mechanism, we therefore match the experiment with the model.

The identification imperative guides our choice between the two standard approaches for measuring image payoffs. One approach consists in having the subject confront peers informed (perhaps by the subjects themselves) about the subject’s prosocial behavior in the experiment. Let us call this the “public shaming or nonmonetary reward approach”. The other, the “monetary reward approach”, has observers rate the subject’s behavior, ultimately determining a reward. It captures the fact that reputations, besides bringing about affective benefits, can also be instrumental as they help be trusted, receive jobs, attract friends and partners. While the monetary-reward approach is a bit less natural than the public-shaming approach, the latter creates image payoffs that are likely to be non-linear. Firstly, how does the emotional impact of others’ judgment scale with the size of the audience? I.e., will I feel twice as embarrassed in front of six peers relative to three peers?⁷ Secondly, is the image cost of being generous with two peers and being selfish

⁷Similar issues arise outside the laboratory: Our desire to be thought well of by other people might be a highly non-linear function of the number of people. It may be that people care a whole lot if there is at least one person who has a positive attitude toward them, and then much less about successive admirers. Perhaps that is one benefit people get from having partners/significant others: You can’t be

vis-à-vis two other peers half of the image cost of being selfish vis-à-vis four peers?

Such non-linearities do not create problems for the theory and the key insights; but they generate a difficult identification problem, as we would need to estimate the shape of the image payoff function (e.g. the potential scale economies or diseconomies associated with multiple activities and/or peers) together with the image spillovers. For this reason, we argue that the monetary-reward approach enables a clean test of the theory; for, it minimizes non-linearity concerns in two ways. By construction, rewards are additive in the number of observers, and so image concerns are linear. Furthermore, we impose that the same behavior be chosen under identical circumstances (i.e. within a sphere), avoiding concerns about the warm glow from being nice to n recipients not being n times the warm glow from being nice to one recipient.

In our experiment, a dictator makes giving decisions to passive recipients via a charity; financially disinterested third-party observers, after observing (some of) the dictator's decisions, rate the dictator's generosity and this rating affects the dictator's final payoff. Arguably, the third-party observers' behavior is driven only by indirect reciprocity (i.e., the intrinsic motivation for rewarding good reputations). The dictator's image concern arising from the observers' indirect reciprocity is thus the only driver, other than self-image, of the image utility.

The dictator's decisions are made either in a public sphere or a private sphere. A giving decision made in the public sphere is observed by all observers, while a giving decision made in the private sphere is observed only by those observers who are also in the private sphere. We investigate how the differential observability in public and private spheres influences prosocial behavior.

In the experiment, subjects make decisions in six mutually exclusive worlds, Tx , where $x \in \{0, 1, 2, \dots, 5\}$. In world Tx , there are x recipients and x observers in the public sphere, and $5 - x$ recipients and $5 - x$ observers in the private sphere. Decisions are binary. A prosocial action in the public sphere of world Tx helps x recipients and is observed by all the five observers. A prosocial action in the private sphere of world Tx helps $5 - x$ recipients and is observed only by the $5 - x$ observers in the private sphere. For each action, the helping cost is proportional to the number of recipients being helped. After the subject makes all the decisions in the six worlds, one world is randomly drawn as the binding one and the subject's decisions for this world are sent to the observers accordingly for rating. In line with the theory's assumption of constant returns in image benefits (i.e. linear image benefits), the observers have an equal weight in determining the subject's income. The total image concern is thus equal to the sum of the image concerns vis-à-vis each observer.

The findings from the experiment support the key theoretical predictions of our model. (1) The cheap-signaling effect generates a higher prosociality in the public sphere than in the all-private treatment. (2) Prosociality in the public sphere decreases with the

all bad if there is at least one person who knows you and thinks highly of you.

size of the public sphere. (3) The moral-licensing effect generates a lower prosociality in the private sphere than in the all-private treatment. (4) Prosociality in the private sphere decreases with the size of the public sphere. (5) The subjects misallocate efforts by behaving more prosocially in the public sphere than in the private sphere. (6) Overall prosociality may not increase monotonically with an expansion of the public sphere.

The rest of this paper is organized as follows. Section 2 develops the theory. Section 3 is devoted to the experiment. It describes the experimental design and implementation and reports the findings. Section 4 reviews the related theoretical and experimental literature. The last section concludes.

2 Theory

2.1 Single-task benchmark

The model builds on the large theoretical and empirical literature that posits that an individual’s social behavior results from her intrinsic motivation to do good for others, her cost of doing so, and finally her desire to project a good image of herself.⁸ The benchmark model developed in this subsection is standard.

Drivers of social behavior. There is a continuum of agents with mass 1. Individual i selects an action $a_i \in \{0, 1\}$.⁹ Action $a_i = 1$ costs the agent $c > 0$ and is pro-social in that it creates an externality $e > 0$ onto the rest of society, while action $a_i = 0$ does not.¹⁰

Individuals are heterogenous with respect to their desire to do good. Namely, their intrinsic motivation to do good (exert a positive externality), v , is distributed according to smooth cumulative distribution $F(v)$ and density $f(v)$ on $[0, +\infty)$, with mean \bar{v} .¹¹ That the distribution F has support \mathbb{R}^+ captures the idea that the behavior is consensual: All agree that $a_i = 1$ is good for the rest of society, although they differ in the extent to which they are willing to incur a cost to contribute. Individual i ’s intrinsic motivation, v_i , is private information. Individual i cares about others’ posterior mean $\hat{v}_i(a_i) = E[v_i|a_i]$ about her type. For the moment, the agent has a single reputation, and their utility is linear in this reputation. Let μ denote the intensity of image concerns.

⁸The three motivations – intrinsic, extrinsic and image – model is borrowed from Bénabou-Tirole (2006, 2011a) and the broader signaling literature.

⁹Either there is a single action a_i or the agent plays the same action a_i with everyone.

¹⁰“Externalities” refer to the standard economic notion of inflicting physical harm, raising cost or creating nuisances.

¹¹One may argue that a realistic support for F is $[0, e)$, i.e., that agents never put more weight on others than they do on themselves. Note also that assuming that the intrinsic motivation grows with the magnitude of the externality (e.g., can be written ve) would not alter the results. In our experiment, we will take this externality as fixed anyway. These results would however hold under the more general description of intrinsic motivation.

Payoff functions. Agent i 's utility is¹²

$$u_i = (v_i - c)a_i + \mu\hat{v}_i(a_i).$$

Equilibrium. Because u_i is increasing in v_i , there exists a threshold v^* over which the individual behaves prosocially and under which she does not. Letting

$$\Delta(v^*) \equiv M^+(v^*) - M^-(v^*) \equiv E[v|v \geq v^*] - E[v|v < v^*]$$

denote the reputational gain from prosocial behavior, the cutoff $v^* \equiv v^*(\mu)$, if interior,¹³ solves

$$v^* - c + \mu\Delta(v^*) = 0. \quad (1)$$

We henceforth assume that $1 + \mu\Delta'(v^*(\mu)) > 0$ to preclude any multiplicity of equilibrium. This condition is always satisfied for an ‘‘anti-norm’’ ($\Delta' > 0$; Bénabou-Tirole 2006, 2011b); in the case of a ‘‘norm’’ ($\Delta' < 0$), it requires that the intensity μ of image concerns not be too large.

Proposition 1 (*prosocial behavior with uniform observability*)

When faced with an intensity μ of image concerns, individual i picks $a_i = 1$ if $v_i > v^$ and $a_i = 0$ if $v_i < v^*$, where¹⁴*

$$v^* - c + \mu\Delta(v^*) = 0. \quad (2)$$

When technology increases image concerns (μ increases), so prosocial behavior becomes more frequent, truly generous motives pale relative to personal score maximization (the ratio of average intrinsic motivation over image concerns decreases).¹⁵

While we focus in the text on the positive aspects of the theory, which will be tested in the experiment, part (a) of Appendix A derives social welfare. It shows that, in the single-task case, provided that prosocial behavior is socially desirable (the externality is

¹²Were a_i to be observed by only a fraction x of the population, u_i could be rewritten as $u_i = (v_i - c)a_i + \mu[x\hat{v}_i(a_i) + (1 - x)\bar{v}]$. So the same analysis holds, replacing μ by $\tilde{\mu} = \mu x$.

¹³For a uniform distribution of v on $[0, 1]$, $\Delta(v^*) = 1/2$ for all v^* . More generally, Jewitt (2004)'s lemma indicates that (a) if the density f is everywhere increasing, then $\Delta' < 0$; (b) if it is everywhere decreasing, $\Delta' > 0$; and (c) if f is single-peaked, Δ is first decreasing and then increasing in v^* . We will adopt the convention $v^* = 0$ if $-c + \mu\Delta(0) \geq 0$. Indeed, a corner solution at $v^* = 0$ exists if and only if $\mu\bar{v} \geq c$. Thus, the condition $c > \mu\bar{v}$ is sufficient for the existence of an interior equilibrium.

¹⁴See the conditions for interiority in Footnote 13.

¹⁵This is consistent with the second sentence in the following statement of Stuart Russell (2019, page 106): ‘‘[Under a system of intrusive monitoring and coercion] outward harmony masking inner misery is hardly an ideal state. Every act of kindness ceases to be an act of kindness and becomes instead an act of personal score maximization and is perceived as such by the recipient.’’ More specific to Section 2.2 is a tentative interpretation of the first sentence, which may be understood as a deterioration of behavior in the private sphere as technology expands the public sphere.

not too small, in that it exceeds “prestige stealing”), transparency increases welfare.¹⁶

2.2 Multitasking: actions with heterogeneous visibility

The benchmark model has a single moral choice. The audience may be small or large, but the size of the audience affects only the intensity of the agent’s image concerns. A richer picture emerges when the agent chooses multiple moral choices with different visibility. In particular, some behaviors are bound to remain in the private sphere because they are unobservable by third parties and furthermore cannot be reliably rated.¹⁷ Other behaviors, happening in public spaces, reported in social networks, or spread by word-of-mouth, are observed by a wider audience. This section focuses on the mutual interdependence through image concerns between the private and public spheres, and on how an expansion in the public sphere impacts overall behavior.

Agent i , with type v_i drawn from distribution $F(v_i)$ with support \mathbb{R}^+ , interacts with a mass 1 of other agents j (a finite number of interactions (even only two) would not affect the qualitative results). In each interaction, agent i decides to behave prosocially ($a_{ij} = 1$) or not ($a_{ij} = 0$). Behaving prosocially generates an externality e on agent j (the counterparty) or on society as a whole, and involves private cost c for individual i .

In the basic model (and in the experiment), we fix the number of relationships (see Section 2.3 for a relaxation of this assumption). Suppose that a fraction t of individual i ’s activities is transparent, while a fraction $s = 1 - t$ is private (“ t ” and “ s ” stand for “transparent” and “silo”). Individual i knows which activities are transparent or private. In practice, this fraction t may be affected by the technological evolution (cameras, social networks, cheap data storage, artificial intelligence, . . .), the social pressure for transparency as well as the government’s policy or firms’ algorithms.¹⁸ We will focus on *deterministic symmetric equilibrium behaviors* within each sphere ($a_{ij} = a_{ik} = a^t$ if j and k are in a public-sphere interaction with i and $a_{ij} = a_{ik} = a^s$ if j and k belong to i ’s private sphere, and $(a^t, a^s) \in \{0, 1\}^2$). Overall, in a deterministic symmetric equilibrium, agents in i ’s public sphere observe only a_i^t while agents in i ’s private sphere observe $\{a_i^t, a_i^s\}$. The single-task model of Section 2.1, which corresponds to $t = 0$ or $t = 1$, is nested in this

¹⁶Anticipating a bit, in the multi-task case, in contrast, even if prosocial behavior is socially desirable, an increase in transparency may lower both overall prosociality and welfare.

¹⁷The reliability of ratings by employers, friends or partners is usually questionable. Outsiders may be unable to ascertain whether a rating within a maintained relationship (or to the contrary following an acrimonious separation) is genuine.

¹⁸In 2014, the Chinese government launched a massive policy plan for building a “social credit system” that inter alia would score, publicize and even blacklist fraudulent behaviors or mispractices in marketplace, professions and everyday life (Dai 2020). This system’s narrative is the development of a unified and numerical record (social score) to evaluate trustworthiness for individuals and businesses. Scoring systems can be found in other countries (e.g., FICO in USA, Schufa in Germany), where most of them were developed by financial institutions to evaluate individuals’ creditworthiness. Concerns have been expressed about the social credit system’s scoring along divisive issues (e.g., its ability to generate societal control; see Tirole 2021); here we focus on a social credit system that, as the initial description claimed, aims at inducing better behavior along consensual issues.

broader multi-task one.

Assuming that agent i has the same image concern μ with respect to all members in the audience (whether public or private),¹⁹ individual i with type v_i has payoff function

$$u_i = \int_0^1 [(v_i - c)a_{ij} + \mu\hat{v}_{ij}]dj,$$

where \hat{v}_{ij} is agent i 's reputation with agent j .

Focusing on equilibria that involve deterministic symmetric behaviors with each sphere, we let $\hat{v}_i(a_i^t, a_i^s)$ and $\hat{v}_i(a_i^t)$ denote the posterior expectations of v_i conditional on the information in the private and public spheres, respectively,²⁰ and let v^t and v^s denote the cutoffs in the two spheres. Agent i chooses $(a^t, a^s) \in \{0, 1\}^2$ so as to solve:

$$\max_{(a^t, a^s) \in \{0, 1\}^2} \{(v_i - c)(ta^t + sa^s) + \mu[t\hat{v}_i(a^t) + s\hat{v}_i(a^t, a^s)]\}.$$

For expositional conciseness, we rule out corner solutions in the all-private or all-public spheres: For this, we assume that $c > \mu\bar{v}$, so that in the all-private ($t = 0, v^s = v^*$, where v^* is the cutoff in the single-task case) or all public ($t = 1, v^t = v^*$) cases, not all types contribute ($v^* > 0$). As we will show, there always exists an equilibrium in which agents behave more prosocially in the public sphere: $v^t < v^s$, as represented in Figure 1. The cutoff in the private sphere is then given by

$$v^s - c + \mu[M^+(v^s) - M(v^t, v^s)] = 0$$

or equivalently,

$$v^s - c + \mu[\Delta(v^s) - \underbrace{(M(v^t, v^s) - M^-(v^s))}_{\text{moral licensing}}] = 0 \quad (3)$$

¹⁹Our analysis can be generalized to different image intensities in the public and private spheres ($\mu^t \neq \mu^s$). Our two key effects, the cheap-signaling effect and the moral-licensing effect, are still present in the generalized model, generating the same misallocation of effort and the same crowding out of prosociality in the private sphere when the public sphere expands. Furthermore, our experiment is designed in such a way that image intensities are identical in the public and private spheres.

²⁰We abuse notation by letting \hat{v}_i denote both reputation functions (with one argument when the reputation is in the public sphere and two arguments when it is in the private one). Allowing for deviations from the equilibrium path, and letting \mathbf{a}_i^t denote the vector of agent i 's actions in the public sphere (that is, $\mathbf{a}_i^t = \{a_{ij}\}_{j \in T_i}$ where T_i denote i 's public sphere), agent i 's objective function can be more generally written as

$$u_i = \int_0^1 [(v_i - c)a_{ij} + \mu[\hat{v}_{ij}(\mathbf{a}_i^t)\mathbb{1}_{j \in T_i} + \hat{v}_{ij}(\mathbf{a}_i^t, a_{ij})\mathbb{1}_{j \notin T_i}]]dj$$

letting $\mathbb{1}_{j \in T_i} = 1 - \mathbb{1}_{j \notin T_i}$ denote the indicator function for i 's public sphere (equal to 1 if $j \in T_i$ and 0 otherwise).

As usual, one has substantial leeway in specifying off-path beliefs. One can for example take $\hat{v}_{ij} = \hat{v}_i(\min_{k \in T_i} a_{ik})$ in the public sphere and $\hat{v}_i(\min_{k \in T_i} a_{ik}, a_{ij})$ in the private sphere for the functions \hat{v}_i that emerge in the deterministic symmetric equilibrium.

where $M(v_0, v_1) \equiv [\int_{v_0}^{v_1} v dF(v)]/[F(v_1) - F(v_0)]$ is the expected type given that $v \in [v_0, v_1]$.

Condition (3) captures the *moral-licensing effect*: Because $M(v^t, v^s) \geq M^-(v^s)$, with strict inequality except when everyone behaves well in the public sphere ($v^t = 0$), condition (3) implies that $v^s \geq v^*$ (where, recall, v^* is the cutoff in the single-task case, given by (2)), with again a strict inequality whenever $v^t > 0$. Even if he does not contribute in the private sphere, he has already separated himself from the chaff if he has contributed in the public sphere. Moreover, when behavior in the public sphere improves (v^t decreases), it also improves in the private sphere (v^s decreases).

As for the public sphere, either $v^t = 0$ or $v^t > 0$ is given by the following equation:

$$v^t - c + \underbrace{\mu[\Delta(v^t) + \frac{s}{t}[M(v^t, v^s) - M^-(v^t)]]}_{\text{cheap signaling}} = 0 \quad (4)$$

Condition (4) captures the *cheap-signaling effect*, implying that $v^t \leq v^*$: When contributing, the individual perceives an extra reputational payoff relative to the all-public or all-private environments, proportional to $\frac{s}{t} = \frac{1-t}{t}$, per good deed in the public sphere. Thus signaling in the public sphere is particularly cheap when the public sphere is small. Besides the standard image benefit $\mu\Delta(v^t)$ per partner in the public sphere, the agent uses the public sphere to engage in damage control in the private sphere. When behavior in the private sphere improves (v^s decreases), it deteriorates in the public sphere (v^t increases).

The analysis of this equilibrium and of its uniqueness is developed in parts (b) and (c) of Appendix A. Proposition 2 summarizes the main conclusions for our experiment by looking at prosocial behavior in the two spheres when the fraction of activities in the private sphere varies.

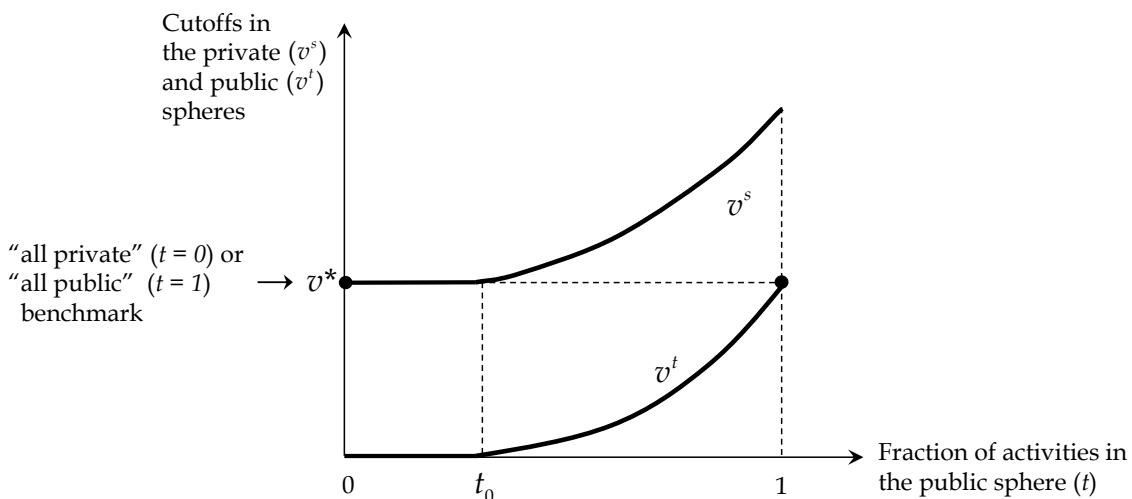


Figure 1: Equilibrium contributions (where $v^* - c + \mu\Delta(v^*) = 0$).

Proposition 2 (when private activities become public)

- (i) *Existence, uniqueness and monotonicity.* There exists an equilibrium satisfying $v^t < v^s$. The cutoffs v^t and v^s are given by conditions (3) and (4), are almost everywhere differentiable in t and satisfy

$$\frac{dv^t}{dt} \geq 0 \quad \text{and} \quad \frac{dv^s}{dt} \geq 0 \quad \text{a.e.}$$

When the density f is non-increasing, this equilibrium is the unique deterministic, symmetric-strategy equilibrium.²¹

- (ii) *Misallocation.* The co-existence of a public and a private spheres implies a misallocation of contributions between the two ($v^t < v^* < v^s$).
- (iii) *Crowding out by public sphere.* An expansion of the public sphere reduces prosociality in the two spheres. It increases the total contribution $\bar{a}(t)$ for $t < t_0$ for some $t_0 > 0$ and over some range reduces the total contribution ($\bar{a}(t)$ decreases with t). $\bar{a}(t)$ is hump-shaped in the case of a uniform distribution of v .

For a narrow public sphere, the individual acts infrequently in the public sphere and so the visibility/cost ratio is high: Signaling in the public sphere is particularly cheap and $v^t = 0$. Behavior in the public sphere is uninformative and so there is no moral licensing: The individual behaves in the private sphere as in the “all private” benchmark. As t grows, though, signaling in the public sphere becomes more expensive, and this cost effect (weakly) reduces contributions in the public sphere, generating a moral licensing effect.

The intuition behind Part (ii) of Proposition 2 can be explained by the cheap signaling effect and the moral licensing effect discussed above. To grasp the intuition behind Part (iii) of Proposition 2, note that the cost of controlling the damage associated with unethical behavior in the private sphere through ethical behavior in the public sphere increases with the size t of the public sphere and its benefit decreases with t . Put differently, damage control is relatively cheap when the public sphere is small, and its cost-benefit ratio increases when behavior becomes more transparent. Thus an expansion of the public sphere discourages contributions in that sphere. Furthermore, having contributed in the public sphere is more of a mark of distinction as the public sphere expands;²² and so, an expansion in the public sphere crowds out contributions in the private sphere.

The excessive attention to public behavior leads to a disintegration of the social fabric in the private sphere. But the split between private and public sphere also affects the

²¹When the density f is single-peaked, multiple equilibria may coexist for a small enough public sphere. The monotonicity of v^t and v^s in t however still applies to stable equilibria.

²²An increase in t reduces contributions in the public sphere. Therefore, an agent who contributes in the public sphere but not in the private sphere has a better image. This reduces the reputation gain of contributing in the private sphere as well (since $(M(v^t, v^s) - M^-(v^s))$ in (3) is larger, i.e. an increasing moral licensing effect).

total level of contributions:

$$\bar{a}(t) \equiv s[1 - F(v^s)] + t[1 - F(v^t)],$$

with $\bar{a}(0) = \bar{a}(1) = 1 - F(v^*)$. Hence, whenever differentiable,

$$\frac{d\bar{a}}{dt} = [F(v^s) - F(v^t)] - sf(v^s)\frac{dv^s}{dt} - tf(v^t)\frac{dv^t}{dt} \quad (5)$$

The first term in the RHS of (5) captures a composition effect: Contributions are higher in the public sphere and so an expansion of the public sphere raises the overall level of contributions. The other two terms capture the observation that contributions in both spheres decline with an expansion of the public sphere. The overall effect is in general ambiguous. Indeed, for $t \leq t_0$ where $t_0 = \sup\{t|v^t = 0\}$ (see Figure 1), $\bar{a}(t) = 1 - F(v^*) + tF(v^*)$ is linearly increasing in t ; and $\bar{a}(1) = \bar{a}(0) = 1 - F(v^*)$. So $\bar{a}(t)$ must be decreasing over some non-empty range.

Part (d) of Appendix A derives the function $\bar{a}(t)$ in the case of a uniform distribution of v . In that case $\bar{a}(t)$ is hump-shaped: (linearly) increasing on $(0, t_0)$ and decreasing on $(t_0, 1)$.

2.3 Extensions and applications

We consider a slightly more general model so as to accommodate two relevant extensions: (a) *Technology increases the number of interactions*. We assumed a fixed number of interactions ($s + t = 1$): The agent's number of interactions is given, and a fraction of them is made publicly observable. Alternatively, technology can augment the public sphere without decreasing the private sphere, increasing t by enlarging the audience without altering s . (b) *Price changes*. Second, we can address the complementarity/substitutability of moral choices by altering at the margin their "price", namely the cost of good behavior in each of the spheres. We develop these two extensions within a single framework.

Letting c^t and c^s denote the costs of the two behaviors, and not imposing $s + t = 1$, the equilibrium conditions are similar to conditions (3) and (4), provided that equilibrium

behavior is more prosocial in the public sphere (that is, c^t is not much larger than c^s)²³:

$$v^s - c^s + \mu[\Delta(v^s) - (M(v^t, v^s) - M^-(v^s))] = 0 \quad (3')$$

and

$$v^t - c^t + \mu \left[\Delta(v^t) + \frac{s}{t}[M(v^t, v^s) - M^-(v^t)] \right] = 0 \quad (4')$$

Condition (3') implies that, *ceteris paribus*, an increase in prosocial behavior in the public sphere triggers an increase in prosocial behavior in the private sphere. In contrast, condition (4') implies that, *ceteris paribus*, an increase in prosocial behavior in the private sphere triggers a decrease in prosocial behavior in the public sphere. For comparative statics, though, one must look at the driver of the equilibrium perturbation. Following the analysis in part (c) of Appendix A, we obtain:²⁴

Proposition 3 (*price changes, varying the number of interactions*) *Provided that equilibrium behavior is more prosocial in the public sphere (that is, c^t is not much larger than c^s):*

- (i) *The equilibrium is unique if the density is non-increasing ($f' \leq 0$) and satisfies $v^t < v^s$.*
- (ii) *A small increase in c^s makes behavior in the private sphere more selfish (v^s increases); this in turn increases the payoff $M(v^t, v^s)$ to behaving well in the public sphere, lowering v^t . So behaviors in the private and public spheres co-vary negatively when one manipulates the cost of good behavior in the private sphere.*
- (iii) *A small increase in c^t makes behavior in the public sphere more selfish (v^t increases); this increases moral licensing as $M(v^t, v^s)$ increases, leading to an increase in v^s . So*

²³From (3') and (4'), behavior is indeed more prosocial in the public sphere if $c^t \leq c^s$. Keeping c^s constant, let us increase c^t above c^s until $v^t = v^s = v$. (3') and (4') become

$$\begin{aligned} v - c^s + \mu [\Delta(v) - (v - M^-(v))] &= 0 \\ v - c^t + \mu \left[\Delta(v) + \frac{s}{t}(v - M^-(v)) \right] &= 0 \end{aligned}$$

Subtracting the latter from the former yields

$$c^t - c^s = \mu \left(\frac{s+t}{t} \right) [v - M^-(v)]$$

where $v \equiv v(c^s)$ is determined by

$$v - c^s + \mu [M^+(v) - v] = 0$$

For example, for $v \sim U[0, 1]$, $v - M^-(v) = \frac{v}{2}$ and $M^+(v) - v = \frac{1-v}{2}$. Assuming $\mu < 2$ and $\frac{\mu}{2} < c^s < 1$ for an interior solution, $v(c^s) = \frac{c^s - \frac{\mu}{2}}{1 - \frac{\mu}{2}}$. And $c^t \leq c^s + \mu \left(\frac{s+t}{t} \right) \frac{v(c^s)}{2}$ ensures prosocial behavior is more prevalent in the public sphere than in the private sphere (i.e. $v^t \leq v^s$).

²⁴The second and third points of Proposition 3 qualify the validity of the altruism-budget assumption posited by Gee and Meer (2019). See also the literature review in Section 4.1.

behaviors in the private and public spheres co-vary positively when one manipulates the cost of good behavior in the public sphere.

- (iv) A small increase in the size of the public sphere, keeping the private sphere one constant, operates as an increase in the price of good behavior in the public sphere. It reduces prosocial behaviors in both spheres. The general pattern of Proposition 2, obtained for a constant total activity, is unchanged.

Finally, our formalism helps us describe the moral-licensing and moral-cleansing experiments conducted in the psychology literature (see Footnote 6). These experiments compare behavior in one realm with and without prior behavior in another realm. For instance, formulating a racist joke or discriminating against a member of a minority is more widespread when the person has had a prior opportunity to vote for a minority politician. So these experiments exhibit moral licensing or moral cleansing in a context that does not emphasize the private-public distinction; the underlying image mechanism is however the same as in a private-public environment. Here the intensity of image concerns is the same for the two tasks, because there is usually a single audience (the inner self); what differs across tasks rather is the cost of prosocial behavior – as formalized below – or equivalently, the two distributions of the prosociality parameter are distinguished by a uniform shift.

The moral-licensing and moral-cleansing experiments involve two sequential moral tasks. Furthermore, the subjects are unaware of the incoming second task when responding to the first. To capture these experiments, let us label c_1 and c_2 the costs of prosocial behavior in tasks 1 and 2. In the first task, the cutoff is determined as if there were a single task:

$$S \equiv v_1^* - c_1 + \mu\Delta(v_1^*) = 0$$

Under moral licensing, prosociality in the first task ($v \geq v_1^*$) exonerates the subject from their moral duty in the second one, yielding cutoff $v_2^* \geq v_1^*$ (provided that c_2 is larger or not much smaller than c_1) given by:

$$L(v_2^*) \equiv v_2^* - c_2 + \mu[\Delta(v_2^*) - \underbrace{[M(v_1^*, v_2^*) - M^-(v_2^*)]}_{\text{moral licensing}}] = 0$$

Under moral cleansing, immoral behavior in the first task may induce the subject to wash their hands of the previous sin through prosocial behavior. Let $v_2^* \leq v_1^*$ be given by²⁵:

$$C(v_2^*) \equiv v_2^* - c_2 + \mu[\Delta(v_2^*) - \underbrace{[M^+(v_2^*) - M(v_2^*, v_1^*)]}_{\text{moral cleansing}}] = 0$$

Proposition 4 (*interpreting moral licensing and moral cleansing in psychology experiments*)

²⁵For this, c_2 must be smaller than a threshold, itself smaller than c_1 .

Psychology experiments involve two sequential moral tasks (with usually a single audience: the inner self). In the first task, the subject is unaware of the impending, second-stage signaling opportunity. There is therefore no cheap-signaling effect and only a moral-licensing or moral-cleansing one. The cutoffs are v_1^ in the first stage and v_2^* in the second stage, with $v_2^* \geq v_1^*$ (resp. $v_2^* \leq v_1^*$) under moral licensing (resp. moral cleansing). There exist thresholds $c_2^{ML}(c_1) < c_1$ and $c_2^{MC}(c_1) < c_1$ such that moral licensing is an equilibrium iff $c_2 > c_2^{ML}(c_1)$ and moral cleansing is an equilibrium iff $c_2 < c_2^{MC}(c_1)$.*

Remark 1 (*altruism fatigue*) Proposition 4 has implications for charities and other organizations that call upon the individuals' generosity. The first mover exerts a negative externality on the second mover if the subject engages in moral licensing; so the second mover would be better off preempting (asking first). Similarly, in case of moral cleansing, the second mover suffers from the fact that subjects who attempt to rebuild their reputation only partially succeed in doing so as previous behavior puts a ceiling on how good they are. There is thus a benefit to asking first.

Remark 2 (*complementarity or substitutability?*) Part (e) of Appendix A discusses the exact relation between Propositions 2(iii) and 3(iii) (drivers that reduce prosociality in the public sphere also reduce prosociality in the private sphere) on the one hand, and Proposition 4 (the existence of a prior behavior, the counterpart of behavior in the public sphere, may reduce prosociality in the subsequent behavior, the counterpart of behavior in the private sphere) on the other hand.

3 Experiment

3.1 Experimental design and hypotheses

In a basic interaction a dictator decides whether to forgo their endowment, 20 Chinese yuan, to benefit a recipient.²⁶ The recipient receives 30 yuan, and the dictator nothing, if the dictator contributes. If the dictator does not contribute, she keeps the endowment and the recipient receives nothing. There are five potential recipients, whose roles are passive, and five observers, who will observe some of the dictator's decisions.

We employ both a within-subject design and a between-subject design for the experiment. For the within-subject design, there are six mutually exclusive worlds, denoted by Tx , where $x \in \{0, 1, 2, \dots, 5\}$. The dictator makes simultaneous decisions for the six worlds. In world x , there are x recipients and x observers in the public sphere, and $5 - x$ recipients and $5 - x$ observers in the private sphere. Thus, the T0 world has no public sphere, the T5 world has no private sphere, while each of the other four worlds mixes public and private spheres in varying proportions. We thus call T0 and T5 "uniform worlds" and T1-T4 "mixed worlds". In order to be as close to the theory as possible, for

²⁶The experiment was conducted in China. 20 Chinese yuan \approx USD2.7.

each world, we let the dictator make binary decisions of contribution for its public and private spheres (if any) respectively. For all x , a prosocial action in the public sphere in world Tx costs the dictator $20x$ yuan, helps x recipients (each receiving 30 yuan), and is observed by all the five observers; a prosocial action in the private sphere in world Tx costs the dictator $20(5 - x)$ yuan, helps $5 - x$ recipients and is observed only by the $5 - x$ observers in the private sphere. After the dictator makes the ten decisions for the six parallel worlds, one world is randomly selected (with equal probability) and the dictator’s decisions for this world are sent to the five observers according to the above rule. The observers are then asked to simultaneously and independently evaluate the dictator’s generosity on a scale between 0 and 5 based on their own observations. The sum of the five observers’ evaluation is called the dictator’s “generosity score” and will influence the dictator’s income from the experiment.

The longstanding debate about the merits of the within-subject approach is nicely captured by List in the conclusion of his (2024) review: “*In general, but not always, the literature reveals that compared to behavior generated in between-subject designs, subjects in a within-subject design behave more rationally, behave more in line with neoclassical theory, and tend to conform to social norms more closely when they have a comparative context, or likewise an evaluability baseline.*” While a between-subject design may better capture a relatively stable environment, a within-subject approach may be a better fit for a new or fast-changing environment such as the one this paper is concerned with. Furthermore, a within-subject design provides us with larger experimental power for a given sample size, and even more importantly, sufficient information about the individual subject’s strategy (List 2024). We can thus examine whether the subjects’ strategies are consistent with the theoretical predictions (see Section 3.3.1).

The between-subject design varies in the stake of the “generosity score” for the dictator, which we call “High Stake” and “Low Stake”. With the High Stake, the dictator’s income is determined by

$$100 \text{ yuan endowment} - 20 \times \# \text{recipients helped} + \text{generosity score} \times 5,$$

where “# recipients helped” means the number of recipients helped by the dictator in the realized world. With the Low Stake, the dictator’s income is determined by

$$100 \text{ yuan endowment} - 20 \times \# \text{recipients helped} + \text{generosity score} \times 3.$$

Put differently, given the scale of observers’ evaluations in $[0, 5]$, the stake of the generosity score for the dictator’s income is either 125 yuan (“High Stake”), or 75 yuan (“Low Stake”), while the dictator’s total endowment is 100 yuan. Giving is not a dominated strategy even in the low-stake treatment, as these material payoffs ignore the intrinsic motivation. The Online Appendix presents the (translated) experimental instructions (for the Low Stake).

Let $a_{Tx}^s, a_{Tx}^t \in \{0, 1\}$ denote the action in the private sphere and in the public sphere

respectively in world Tx , for all $x \in \{0, 1, \dots, 5\}$. Figure 1 and Proposition 2 predict that,

$$\underbrace{a_{T1}^t \geq a_{T2}^t \geq \dots \geq a_{T5}^t}_{\text{cheap signaling}} = \underbrace{a_{T0}^s \geq a_{T1}^s \geq \dots \geq a_{T4}^s}_{\text{moral licensing}}. \quad (6)$$

Given the binary nature of the actions, the theory predicts that a dictator will either donate all ($a_{Tx}^s = a_{Tx}^t = 1$ for all x), or never donate ($a_{Tx}^s = a_{Tx}^t = 0$ for all x), or choose to donate ($a = 1$) for the beginning decision(s) listed in Equation (6) and switch to $a = 0$ for the rest of the decisions. If the dictator behaved according to the theory, there should not be much variation in a dictator’s choice pattern (which should exhibit at most one switch point).²⁷ Furthermore, if the subjects faced the same incentive scheme and their preference heterogeneity were small,²⁸ many subjects would behave similarly and the empirical results would hardly capture the potential diversity implied by Equation (6).²⁹

We thus adopt the above-mentioned between-subject design by randomly assigning subjects to one of the two stakes, high or low. Given the random assignment of subjects, the High Stake corresponds to a right-ward shift of the distribution of image concerns μ , relative to the Low Stake.

Appendix B presents a power analysis, which demonstrates that having two stakes in the “generosity score” reduces the required sample size needed to test our theory. Each stake could capture some aspects of the behavioral diversity predicted by the theory. A single stake would require a much larger sample size to test the theory.

Equation (6) gives us the following hypotheses.

Hypothesis 1 (cheap signaling in the public sphere): For all $x \in \{1, 2, 3, 4\}$, $a_{Tx}^t \geq a_{T0}^s = a_{T5}^t$. That is, players are more likely to contribute in the public sphere of the mixed worlds than in the uniform worlds.

Hypothesis 1 also predicts that players are equally likely to contribute in the two uniform worlds, T0 and T5.

Hypothesis 2 (increasingly expensive signaling in the public sphere): $a_{T1}^t \geq a_{T2}^t \geq a_{T3}^t \geq a_{T4}^t$. That is, players are less likely to contribute in the public sphere, as the public sphere expands.

²⁷In this sense, our design is similar to the Multiple Price List (MPL) that is commonly used in experimental economics. List (2024) argues that the MPL approach, being a within-subject design, may over-estimate the price elasticities; however, the subject’s preference (MPL’s ultimate interest), elicited from the entire pricing regime, is more reliable. In a similar vein, our ultimate interest is not the magnitude of the “price elasticities” when the public sphere increases. Rather, we are interested in the pattern of substitutability/complementarity emerging from the choice regime.

²⁸This is particularly the case when the subjects are from the same subject pool, e.g., students from the same university.

²⁹This turns out to be true when we analyse the data from a single stake, either high or low. For example, with the Low Stake, many subjects switch somewhere in the public sphere and so choose $a_{Tx}^s = 0$ for all x . Thus, it is hard to discern a decreasing trend in contribution in the private sphere. See Section 3.3 for details.

Hypothesis 3 (moral licensing in the private sphere): For all $x \in \{1, 2, 3, 4\}$, $a_{T_x}^s \leq a_{T_0}^s = a_{T_5}^t$. That is, players are less likely to contribute in the private sphere of the mixed worlds than in the uniform worlds.

Hypothesis 4 (increasing moral licensing in the private sphere): $a_{T_1}^s \geq a_{T_2}^s \geq a_{T_3}^s \geq a_{T_4}^s$. That is, players are less likely to contribute in the private sphere, as the public sphere expands.

Hypothesis 5 (misallocation in contributions): For all $x \in \{1, 2, 3, 4\}$, $a_{T_x}^t \geq a_{T_x}^s$. That is, in each world where the public sphere and private sphere coexist, players are more likely to contribute in the public sphere than in the private sphere.

The six worlds with varied t allow us to test the non-monotonic relationship between overall contributions and the size of the public sphere. Let the overall contribution \bar{a} be the sum of the numbers of the recipients helped in the public and private spheres, which ranges from 0 to 5.

Hypothesis 6 (non-monotonicity in total contributions): The overall contribution changes non-monotonically as the public sphere expands.

Finally, we expect that a larger stake in the image concerns will incentivize subjects to contribute more:

Hypothesis 7 (shift of image concerns): Subjects contribute more under the High Stake than under the Low Stake, other things equal.

3.2 Experimental implementation

Our experiment takes advantage of a real-world charity fund in China. This charity fund is run by the Hebei Charitable Joint Foundation on the Alipay charity platform to help sanitation workers, and is officially certified. Low-income sanitation workers, especially those who do cleaning work in central business districts in big cities, usually cannot afford to buy proper lunches near their working area. The charity uses collected funds to finance nutritious lunches for groups of sanitation workers in China. In our experiment, we tell the subjects that we will donate the contributions collected from the experiment to the charity in the name of the university where the experiment is run. To help one recipient (sanitation worker), a subject donates 20 yuan (or equivalently USD2.7 at the time when the experiment was implemented), and the recipient receives 30 yuan (USD4.1), which is roughly the price of a nutritious lunch in big cities in China.

After the subjects make contribution decisions, some student observers, who are not themselves dictators, evaluate the dictators' generosity (according to the above-mentioned rule for a randomly selected world). It is common knowledge that the observers receive a fixed payment, which is independent of their evaluations. The subjects' payments are determined after the evaluations are completed. They are paid in private via WeChat payment, which is a commonly used online payment method in China. An assistant,

who does not know about the experiment, handles the payments and anonymizes the data for our analysis. We then make a lump-sum donation to the charity through the Alipay platform, according to the subjects’ decisions in the randomly selected world. The subjects also receive a soft copy of the receipt of the lump-sum donation (if they made any donation in the selected world). The dictators (subjects), recipients (sanitation workers) and observers are mutually anonymous to each other.

We ran the experiment at the economics laboratory of Wuhan University, China, in October 2022. In total, 179 subjects, with various majors from the student subject pool of the university, participated. The subjects had no previous experience in our experiment or similar experiments. 108 of them were assigned to the High Stake and 71 in the Low Stake, depending on the sign-up. As mentioned above, we rely on each stake to capture some aspects of the behavioral diversity predicted by the theory. For this purpose, the sample size for each stake is large enough given the power analysis presented in Appendix B. An experimental session lasted for about one hour. On average a subject received 100.5 yuan (or equivalently USD13.8) and the total amount of donation we made to the charity was 13,230 yuan.

3.3 Experimental results

The results from the experiment support all the Hypotheses 1-7.

3.3.1 Summary statistics

(a) Behavioral patterns

Our within-subject design gives us full information about the subjects’ strategies, and thus, provides us with a unique opportunity to examine the proportion of subjects whose strategies are consistent with the theoretical predictions. Behavior that is consistent with the predictions in Equation (6) can be divided into the following categories: (i) “Donate All”: $a_{Tx}^t = a_{Tx}^s = 1$ for all $x \in \{0, 1, \dots, 5\}$; (ii) “Keep All”: $a_{Tx}^t = a_{Tx}^s = 0$ for all $x \in \{0, 1, \dots, 5\}$; (iii) “Switch in Pub”: There exists $1 \leq x^* \leq 4$ such that $a_{Tx}^t = 1$ for all $x \leq x^*$ and $a_{Tx}^s = 0$ otherwise, and $a_{Ty}^s = 0$ for all $y \in \{0, 1, 2, 3, 4\}$; (iv) “Switch in Pvt”: There exists $0 \leq x^* \leq 3$ such that $a_{Tx}^s = 1$ for all $x \leq x^*$ and $a_{Tx}^t = 0$ otherwise, and $a_{Ty}^t = 1$ for all $y \in \{1, 2, 3, 4, 5\}$. Any behavior that does not belong to any of these categories is inconsistent with the theory and is called “Irrational behavior”.

Table 1 shows the distribution of behavioral patterns for the whole sample and for High and Low Stakes, respectively. Overall, the behavior of 134 out of the 179 subjects (74.8%) is consistent with the theory (Equation (6)). The Fischer’s exact test shows that there is no significant difference in the likelihood of consistent-to-the-theory behavior between the High Stake sample and the Low Stake sample (p value = 0.29).

Due to the larger stake from the “generosity score”, the subjects under the High

Stake are more likely to choose $a_{T_x}^t = 1$ for all $x \in \{1, 2, 3, 4, 5\}$ (including “Donate All” and “Switch in Pvt”) than under the Low Stake (Fischer’s exact test, p value < 0.001); meanwhile, the subjects under the Low Stake are more likely to choose $a_{T_x}^s = 0$ for all $x \in \{0, 1, 2, 3, 4\}$ (including “Keep All” and “Switch in Pub”) than under the High Stake (Fischer’s exact test, p value < 0.001).³⁰ These findings will bear some implications on the significance of results when we test some of the hypotheses with subsamples from a single stake, as discussed below.

Table 1: Distribution of Behavioral Patterns

| Behavioral Patterns | Whole Sample | | High Stake | | Low Stake | |
|---------------------|-----------------|---------|-----------------|---------|-----------------|---------|
| | Num.of Subjects | Percent | Num.of Subjects | Percent | Num.of Subjects | Percent |
| Donate All | 35 | 19.55 | 28 | 25.93 | 7 | 9.86 |
| Switch in Pvt | 45 | 25.14 | 39 | 36.11 | 6 | 8.45 |
| Switch in Pub | 38 | 21.23 | 12 | 11.11 | 26 | 36.62 |
| Keep All | 16 | 8.94 | 5 | 4.63 | 11 | 15.49 |
| Irrational | 45 | 25.14 | 24 | 22.22 | 21 | 29.58 |
| Total | 179 | 100 | 108 | 100 | 71 | 100 |

(b) *Contribution frequencies under various scenarios*

Figure 2 plots the average contribution frequency in public and private spheres in different worlds, with the whole sample presented at Figure 2(a), the High Stake sample in Figure 2(b) and the Low Stake sample in Figure 2(c). We have the following observations for the whole sample and for the subsamples.

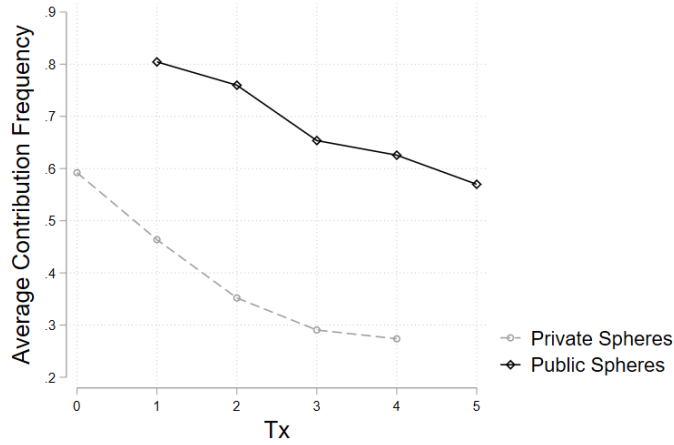
First, the average contribution frequency in world T0 is close to that in T5 ($\bar{a}_{T0}^s = 0.59 \approx \bar{a}_{T5}^t = 0.57$, and 171 out of the 179 subjects choose $a_{T0}^s = a_{T5}^t$).³¹

Second, in each of the mixed worlds, the average contribution frequency is higher in the public sphere than in the uniform world T5, and is lower in the private sphere than in the uniform world T0; that is, $\bar{a}_{T_x}^t > \bar{a}_{T5}^t$ and $\bar{a}_{T_x}^s < \bar{a}_{T0}^s$ for all $x \in \{1, 2, 3, 4\}$. If we look at the two stakes separately, the pattern remains, and is stronger in the public sphere under the Low Stake and in the private sphere under the High Stake, but is weaker in the private sphere under the Low Stake and in the public sphere under the High Stake.³²

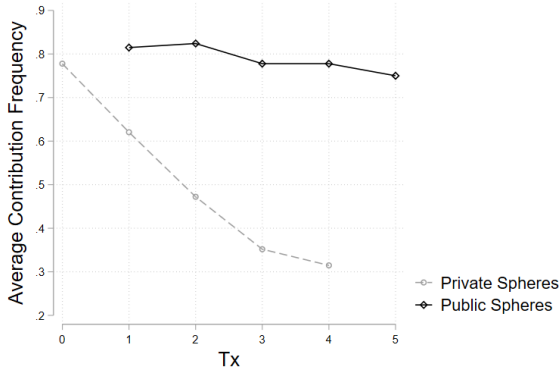
³⁰Under the High Stake, 62% of the subjects choose $a_{T_x}^t = 1$ for all $x \in \{1, 2, 3, 4, 5\}$, while under the Low Stake, only 18% of the subjects do so. Under the Low Stake, 52% of the subjects choose $a_{T_x}^s = 0$ for all $x \in \{0, 1, 2, 3, 4\}$, while under the High Stake, only 16% of the subjects do so.

³¹Under the High Stake, the average contribution frequency is 0.78 in world T0 and 0.75 in world T5. Under the Low Stake, the average contribution frequency is 0.31 in world T0 and 0.30 in world T5. The McNemar’s test shows that under each stake (High or Low), there is no statistically significant difference in terms of contribution likelihood between world T0 and T5 (p values > 0.17). Table C1 in Appendix C reports the p values of all the McNemar’s tests we conducted. Given the binary data structure and the within-subject design, the McNemar’s test is an appropriate non-parametric test, which is widely used in the literature (Moffatt 2015).

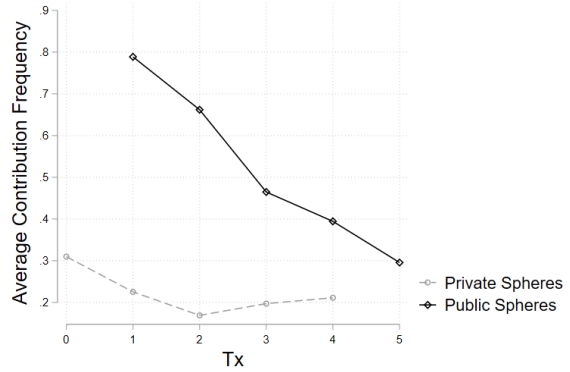
³²The weaker pattern there can be explained by the smaller behavioral variation in the public sphere under the High Stake and in the private sphere under the Low Stake. A majority (62%) of the subjects under the High Stake choose $a_{T_x}^t = 1$ for all $x \in \{1, 2, 3, 4, 5\}$, and a majority (52%) of the subjects under



(a) Whole Sample



(b) High Stake



(c) Low Stake

Figure 2: Average Contribution Frequency

Specifically, the McNemar's test shows that under the Low Stake, the contribution is significantly more likely for a_{Tx}^t , for all $x \in \{1, 2, 3, 4\}$, than for a_{T5}^t , and that under the High Stake, the contribution is significantly less likely for a_{Tx}^s , for all $x \in \{1, 2, 3, 4\}$, than for a_{T0}^s (p values < 0.08). Meanwhile, the test shows that under the Low Stake, the contribution is significantly less likely for a_{Tx}^s , for all $x \in \{1, \dots, 4\}$, than for a_{T0}^s (p values < 0.1 , except for a_{T4}^s where the p value is 0.108), and that under the High Stake, the contribution is significantly more likely for a_{T2}^t than for a_{T5}^t (p value < 0.1). Here, we compare a_{Tx}^s with a_{T0}^s and compare a_{Tx}^t with a_{T5}^t for all $x \in \{1, 2, 3, 4\}$. Since the average contribution frequency in world T0 is close to that in T5, the analysis comparing a_{Tx}^t with a_{T0}^s and comparing a_{Tx}^s with a_{T5}^t gives us qualitatively similar results. These findings, in general, lend support to Hypotheses 1 and 3.

Third, we observe that for the whole sample, the average contribution frequency decreases when the public sphere expands (moving from world T1 to T4), in both the public

the Low Stake choose $a_{Tx}^s = 0$ for all $x \in \{0, 1, 2, 3, 4\}$.

sphere and the private sphere of the mixed worlds. This is also the case for the private sphere under the High Stake and the public sphere under the Low Stake. These findings lend support to Hypotheses 2 and 4. The regression analysis below will test the monotonicity results in Hypotheses 2 and 4 statistically.

Fourth, in each of the mixed worlds, the average contribution frequency is higher in the public sphere than in the private sphere. That is, $\bar{a}_{T_x}^t > \bar{a}_{T_x}^s$ for all $x \in \{1, 2, 3, 4\}$. The McNemar’s test shows that under each stake (High or Low) and for each world $x \in \{1, 2, 3, 4\}$, players are significantly more likely to contribute in the public sphere than in the private sphere (p values < 0.01). These findings support Hypothesis 5.

Figure 3 plots the average overall contribution for the whole sample (solid line), the High Stake (dashed line), and the Low Stake (dotted line) respectively in different worlds. The relationship between the overall contribution and the magnitude of the public sphere, both varying between 0 and 5, is indeed non-monotonic. For example, under the Low Stake, the overall contribution in world T1 is significantly higher than that in world T0 and that in world T5; under the High Stake, the overall contribution in world T2 is significantly lower than that in world T0 and that in world T5 (Wilcoxon signed-rank test, p values < 0.01). These findings lend support to Hypothesis 6.³³

Finally, the shift of image concerns leads to more contributions under the High Stake than under the Low Stake. This can be observed from Figures 2 and 3. The Mann-Whitney test conducted at the individual level shows that the overall contribution is significantly higher under the High Stake than under the Low Stake in each world (p values < 0.001), confirming Hypothesis 7.

In this subsection, we tested multiple hypotheses simultaneously. In this case, the significance or the error rate of individual tests no longer represents the error rate of the combined set of tests. To address this concern, we apply the Holm-Bonferroni correction (Holm 1979) to adjust the p values of the nonparametric tests reported in this subsection. Following the correction, the significance levels remain unchanged for the Fisher’s exact tests reported in part (a) of Section 3.3.1 and the Wilcoxon signed-rank tests in part (b) of Section 3.3.1. The significance levels of most of the McNemar’s tests remain unchanged too.³⁴ Thus, our findings in this subsection stand firm after accounting for the issue of multiple hypothesis testing.

3.3.2 Regression results

(a) Determinants of contribution behavior

Table 2 reports odds ratios from logit regressions of subjects’ contribution decisions,

³³Moreover, the theory predicts that $\bar{a}(t)$ is inverse U-shaped for uniform distributions. The relation indeed looks inverse U-shaped under the Low Stake. The theory also predicts that $\bar{a}(t)$ is increasing when t is sufficiently low. However, the discrete nature of t in the experiment may not allow us to capture this property. We do not observe such a pattern from the High Stake sample.

³⁴The adjusted p values of the McNemar’s tests are reported in Table C1 in Appendix C.

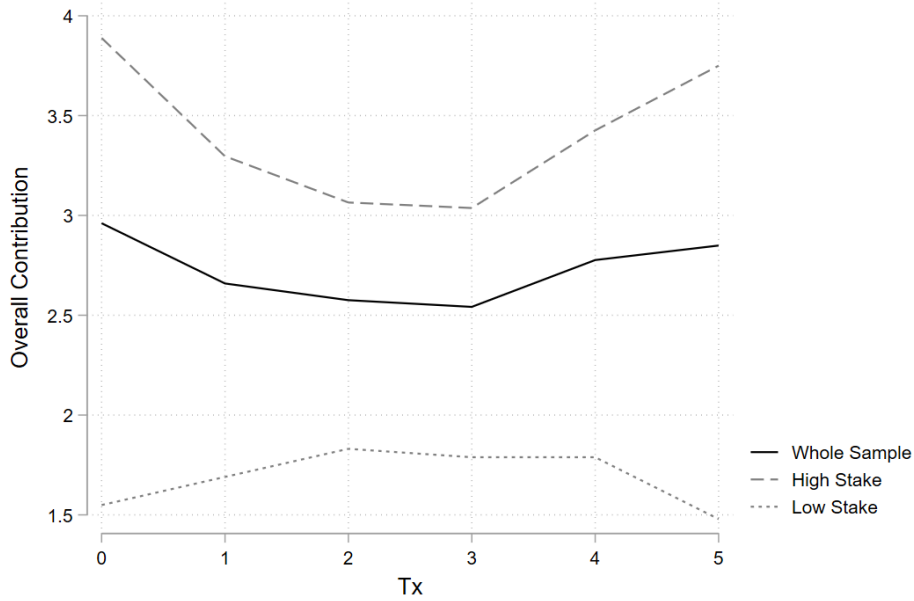


Figure 3: Overall Contribution

clustering standard errors at the individual level. Column (1) controls for the size of the public sphere, $x \in \{0, 1, \dots, 5\}$, which indicates the number of recipients in the public sphere, and a dummy variable *Public* indicating whether or not the decision is made in the public sphere. Column (2) further introduces the interaction term between x and *Public*. Columns (1) and (2) control for individual fixed effects. Column (3) modifies the regression in Column (2), by replacing the individual fixed effects with a dummy variable *Low Stake* (indicating the Low Stake) and demographic variables (gender, age, major, year of study, and a dummy variable indicating previous experience of participation in behavioral experiments).

In Column (1), the odds ratio of *Public* is significantly greater than one. In Columns (2)-(3), the odds ratios of $Public + Public \times x$ (by Wald test of linear restrictions) are also significantly greater than one. These findings support Hypothesis 5: Subjects are more likely to contribute in the public sphere than in the private sphere, other things equal. In both Columns (2)-(3), the odds ratios of x and $x + Public \times x$ are all significantly smaller than one, suggesting that in both the private sphere and the public sphere, the contribution is significantly decreasing in the magnitude of the public sphere $x \in \{0, 1, \dots, 5\}$. There are two implications of this finding. First, when there is a co-existence of public and private spheres, the contribution is significantly decreasing in the magnitude of the public sphere $x \in \{1, 2, 3, 4\}$, in both the private sphere and the public sphere, supporting Hypotheses 2 and 4. Second, the finding also implies that contribution $a_{T_x}^t$ for $x \in \{1, 2, 3, 4\}$ is higher than the baseline case $a_{T_5}^t$, and contribution $a_{T_x}^s$ for $x \in \{1, 2, 3, 4\}$ is lower than the baseline case $a_{T_0}^s$, lending some support to Hypotheses 1 and 3. Table 5 below, which uses the T0 world as the baseline, will more directly test Hypotheses 1 and 3. Finally,

Table 2: Determinants of contribution behavior

| | Whole Sample | | | Mixed Worlds T1–T4 | | |
|----------------------------------|-----------------------|-----------------------|---------------------|-----------------------|-----------------------|---------------------|
| | (1) | (2) | (3) | (4) | (5) | (6) |
| x | 0.443*** (0.043) | 0.373*** (0.062) | 0.671*** (0.038) | 0.439*** (0.063) | 0.424*** (0.089) | 0.731*** (0.050) |
| Public | 53.124*** (22.589) | 25.292*** (15.590) | 5.006*** (1.379) | 57.900*** (26.603) | 49.011*** (42.085) | 6.454*** (2.143) |
| Public \times x | | 1.371 (0.299) | 1.076 (0.100) | | 1.070 (0.322) | 0.960 (0.114) |
| Low Stake | | | 0.267*** (0.071) | | | 0.328*** (0.088) |
| Constant | 32.206*** (8.320) | 51.824*** (24.206) | 0.008** (0.018) | 29.983*** (12.389) | 33.318*** (20.669) | 0.005** (0.012) |
| Wald Test of linear restrictions | | | | | | |
| $x + \text{Public} \times x$ | | 0.512*** | 0.722*** | | 0.453*** | 0.702*** |
| Public + Public \times x | | 34.665*** | 5.385*** | | 52.450*** | 6.197*** |
| Demographics | No | No | Yes | No | No | Yes |
| Individual Fixed Effects | Yes | Yes | No | Yes | Yes | No |
| Pseudo R-squared | 0.617 | 0.620 | 0.183 | 0.651 | 0.651 | 0.187 |
| N | 1790 | 1790 | 1790 | 1432 | 1432 | 1432 |

Note: This table reports odds ratios from logit regressions, with standard errors clustered at the individual level. Variable $x \in \{0, 1, \dots, 5\}$ indicates the size of public sphere. In Columns (1)-(3), $x = 0$ and $Public = 0$ for world T0 and $x = 5$ and $Public = 1$ for world T5. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

in Column (3), the odds ratio of *Low Stake* is significantly smaller than one, which suggests lower contribution under the Low Stake than under the High Stake and confirms Hypothesis 7.

Columns (1)-(3) of Table 2 use the whole sample in the regressions, treating the uniform worlds differently: $x = 0$ and $Public = 0$ for world T0 and $x = 5$ and $Public = 1$ for world T5. Columns (4)-(6) repeat the regressions in Columns (1)-(3) but exclude the uniform worlds. The results are qualitatively the same.

Table 3 repeats the regressions in Columns (1)-(2) of Table 2 for the High Stake sample (in its Columns (1)-(2)) and the Low Stake sample (in Columns (3)-(4)) respectively. For both samples, the results are qualitatively the same as those reported in Table 2, except that in Column (2) the odd ratio of $x + Public \times x$ is smaller than one but statistically insignificant, while in Column (4) the odds ratio of x is smaller than one but statistically insignificant. These exceptions suggest that the decreasing contribution trend in x in the public sphere under the High Stake and in the private sphere under the Low Stake are not

statistically significant. This can be explained by the lack of behavioral variation in these two scenarios (as mentioned in Footnote 32, 62% of the subjects under the High Stake choose $a_{Tx}^t = 1$ for all $x \in \{1, 2, 3, 4, 5\}$ and 52% under the Low Stake choose $a_{Tx}^s = 0$ for all $x \in \{0, 1, 2, 3, 4\}$). Each stake captures some aspects of the behavioral diversity predicted by the theory.

Table 3: Determinants of contribution behavior (by stake)

| | High Stake | | Low Stake | |
|----------------------------------|-----------------------|-------------------------|-----------------------|-------------------------|
| | (1) | (2) | (3) | (4) |
| x | 0.445*** (0.060) | 0.223*** (0.065) | 0.441*** (0.060) | 0.735 (0.148) |
| Public | 53.312*** (31.923) | 2.884 (2.217) | 52.874*** (30.850) | 523.870*** (593.013) |
| Public \times x | | 3.538*** (1.160) | | 0.387*** (0.137) |
| Constant | 31.811*** (11.384) | 246.119*** (224.502) | 0.107*** (0.027) | 0.036*** (0.018) |
| Wald Test of linear restrictions | | | | |
| $x + \text{Public} \times x$ | | 0.790 | | 0.285*** |
| Public + Public \times x | | 10.204*** | | 202.915*** |
| Individual Fixed Effects | Yes | Yes | Yes | Yes |
| Pseudo R-squared | 0.604 | 0.640 | 0.583 | 0.606 |
| N | 1080 | 1080 | 710 | 710 |

Note: This table reports odds ratios from logit regressions, with standard errors clustered at the individual level. Variable $x \in \{0, 1, \dots, 5\}$ indicates the size of public sphere. $x = 0$ and $Public = 0$ for world T0 and $x = 5$ and $Public = 1$ for world T5. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

(b) Overall contributions

Table 4 reports OLS regression results when we regress subjects' overall contributions in a world on the size of the public sphere in the world, x , with robust standard errors clustered at the individual level. Column (1) controls for individual fixed effects, while Column (2) replaces the individual fixed effects with variable *Low Stake* and the demographic variables. Columns (3) and (5) repeat the regression in Column (1) for the High Stake sample and for the Low Stake sample, respectively. We find that, first, the coefficient of *Low Stake* in Column (2) is negative and statistically significant, confirming Hypothesis 7. Second, in all these columns, the coefficients of x are statistically insignificant and the magnitudes are small, which demonstrates that the overall contribution does not increase monotonically with an expansion of the public sphere.

To further investigate the non-monotonic relationship between the overall contribution

and the size of the public sphere, we adopt an interrupted regression (Nelson-Simonsohn 2014) in Columns (4) and (6), for the two samples respectively. We first determine extremum points, x^c , by adding the quadratic term of x to the regressions in Columns (3) and (5).³⁵ The extremum points, x^c , are 2.538 and 2.472 for the High Stake sample and Low Stake sample, respectively. Then, we regress the overall contributions on variables $x(low)$, $x(high)$ and $D^{x>x^c}$ for the High Stake sample and Low Stake sample respectively, reported in Columns (4) and (6), where $x(low)$ equals $x - x^c$ if $x < x^c$ and zero otherwise, $x(high)$ equals $x - x^c$ if $x > x^c$ and zero otherwise, and $D^{x>x^c}$ is a dummy variable equal to one if $x > x^c$, which allows a discontinuity around $x = x^c$. Both the coefficients of $x(high)$ and $x(low)$ are statistically significant and have opposite signs in Column (4), which demonstrates opposite monotonicities for $x \leq 2$ and for $x \geq 3$ for the High Stake sample. The coefficients of $x(high)$ and $x(low)$ have opposite signs, but are statistically insignificant for the Low Stake sample (Column (6)).

Table 4: Determinants of overall contributions

| | Whole Sample | | High Stake | | Low Stake | |
|--------------------------|---------------------|----------------------|---------------------|----------------------|---------------------|---------------------|
| | (1) | (2) | (3) | (4) | (5) | (6) |
| x | -0.007 (0.017) | -0.007 (0.016) | -0.010 (0.022) | | -0.003 (0.028) | |
| Low Stake | | -1.662*** (0.274) | | | | |
| x (low) | | | | -0.412*** (0.102) | | 0.141 (0.140) |
| x (high) | | | | 0.356*** (0.100) | | -0.155 (0.142) |
| $D^{x>x^c}$ | | | | 0.100 (0.167) | | 0.025 (0.215) |
| Constant | 4.850*** (0.043) | -1.925 (2.297) | 4.857*** (0.054) | 4.206*** (0.171) | 0.507*** (0.071) | 0.710*** (0.224) |
| Demographics | No | Yes | No | No | No | No |
| Individual Fixed Effects | Yes | No | Yes | Yes | Yes | Yes |
| Adjusted R-squared | 0.685 | 0.181 | 0.644 | 0.672 | 0.607 | 0.609 |
| N | 1074 | 1074 | 648 | 648 | 426 | 426 |

Note: This table reports OLS regression results of subjects' overall contributions (varying between 0 and 5), with robust standard errors clustered at individual level. Variable x indicates the size of public sphere. Variable $x(low)$ equals $x - x^c$ if $x < x^c$ and zero otherwise, $x(high)$ equals $x - x^c$ if $x > x^c$ and zero otherwise, and $D^{x>x^c}$ is a dummy variable equal to one if $x > x^c$, where x^c , in Columns (4) and (6), is the extremum determined by adding the quadratic term of x to the regression in Columns (3) and (5), respectively. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

(c) *Pairwise comparisons*

³⁵The regression results with the quadratic term are not presented here.

The above regressions look at the effect of public sphere expansion by controlling for the variable x which indicates the size of the public sphere. Table 5, which also reports odds ratios from logit regressions of subjects' contribution decisions, takes a different approach. In the experiment, a subject makes binary decisions in 10 scenarios in total, which differ in whether the decision is made in the public or private sphere and in the size of the public sphere. Table 5 uses the uniform world T0 as the baseline and introduces dichotomous variables that indicate the other nine scenarios $T1Pub, \dots, T4Pub, T1Pvt, \dots, T4Pvt$, and $T5$. Such an approach facilitates direct pairwise comparison between any two of the scenarios and allows us to test Hypotheses 1 and 3 more directly. Column (1) controls for individual fixed effects, column (2) replaces the individual fixed effects by variable *Low Stake* and the demographic variables, while columns (3) and (4) repeat the regression in column (1) with the High Stake sample and Low Stake sample respectively. In all the columns, standard errors are clustered at the individual level.

In Columns (1)-(2) of Table 5, the odds ratios of $T5$ are statistically insignificant, confirming that there is no significant difference in terms of contribution frequency between the two uniform worlds, T0 and T5. The odds ratios of $T1Pub$, $T2Pub$, and $T3Pub$ are statistically significant and greater than one (the odds ratios of $T4Pub$ are greater than one but statistically insignificant). These results lend support to Hypothesis 1. The odds ratios of $T1Pvt$, $T2Pvt$, $T3Pvt$ and $T4Pvt$ are all statistically significant and smaller than one, which supports Hypothesis 3. By Wald test of linear restrictions, the odds ratios of $T2Pub - T1Pub$, $T3Pub - T2Pub$, $T4Pub - T3Pub$, $T5 - T4Pub$ are smaller than one and in particular, those of $T3Pub - T2Pub$ and $T5 - T4Pub$ are statistically significant; the odds ratios of $T1Pvt$, $T2Pvt - T1Pvt$, $T3Pvt - T2Pvt$, $T4Pvt - T3Pvt$ are statistically significant and smaller than one (except for $T4Pvt - T3Pvt$ whose odds ratios are smaller than one but statistically insignificant). These results lend support to Hypotheses 2 and 4. Finally, the odds ratios of $T1Pub - T1Pvt$, $T2Pub - T2Pvt$, $T3Pub - T3Pvt$, and $T4Pub - T4Pvt$ are all statistically significant and greater than one, supporting Hypothesis 5. In Columns (3)-(4) with only one of the stakes, the results are qualitatively similar but some become less statistically significant, due to the fact that a majority of the subjects under the High Stake choose $a_{Tx}^t = 1$ for all $x \in \{1, 2, 3, 4, 5\}$, and a majority of the subjects under the Low Stake choose $a_{Tx}^s = 0$ for all $x \in \{0, 1, 2, 3, 4\}$.

Finally, to address the issue of multiple hypothesis testing, we adjust the p values for Tables 2, 3, and 5, by calculating the Romano-Wolf stepdown adjusted p -values (Romano-Wolf 2005a, b, and 2016). This method controls the family-wise error rate and considers dependencies among p values through bootstrap resampling. We choose this approach because it allows the correction of p values *within* the regression framework. After the correction, the significance levels of the coefficients of interest remain unchanged for Tables 2 and 3, except for that of $Public \times x$ in Column (4) of Table 3, which decreases from 1% to 5%. For Table 5, after the correction, the results remain qualitatively the same, except that the coefficients of $T3Pub$ in Columns (1) and (2), and of $T1Pvt$ and $T3Pvt$ in Column (4) are no longer statistically significant. Overall, the correction of p values does not change the findings qualitatively.

Table 5: Determinants of contribution behavior (Dichotomous scenario variables)

| | Whole Sample | | High Stake | Low Stake |
|----------------------------------|-----------------------|---------------------|-------------------------|-------------------------|
| | (1) | (2) | (3) | (4) |
| T1Pub | 11.557*** (6.496) | 3.276*** (0.830) | 1.668 (1.189) | 188.546*** (225.026) |
| T2Pub | 6.330*** (2.847) | 2.439*** (0.519) | 1.911 (1.150) | 29.499*** (23.460) |
| T3Pub | 1.908* (0.678) | 1.353* (0.226) | 1.000 (0.522) | 4.457*** (2.551) |
| T4Pub | 1.420 (0.420) | 1.177 (0.163) | 1.000 (0.426) | 2.312 (1.193) |
| T5 | 0.794 (0.128) | 0.899 (0.067) | 0.700 (0.181) | 0.861 (0.226) |
| T1Pvt | 0.269*** (0.085) | 0.549*** (0.078) | 0.163*** (0.076) | 0.374* (0.207) |
| T2Pvt | 0.079*** (0.031) | 0.323*** (0.055) | 0.035*** (0.020) | 0.154** (0.114) |
| T3Pvt | 0.034*** (0.018) | 0.235*** (0.048) | 0.007*** (0.006) | 0.250* (0.187) |
| T4Pvt | 0.025*** (0.016) | 0.214*** (0.047) | 0.003*** (0.004) | 0.308 (0.228) |
| Low Stake | | 0.265*** (0.071) | | |
| Constant | 61.589*** (26.475) | 0.009** (0.020) | 269.920*** (226.179) | 0.055*** (0.028) |
| Wald Test of linear restrictions | | | | |
| T1Pub - T1Pvt | 42.982*** | 5.963*** | 10.224*** | 503.788*** |
| T2Pub - T2Pvt | 79.652*** | 7.550*** | 54.148*** | 191.665*** |
| T3Pub - T3Pvt | 56.227*** | 5.755*** | 139.063*** | 17.816*** |
| T4Pub - T4Pvt | 55.897*** | 5.488*** | 295.093*** | 7.498*** |
| T2Pub - T1Pub | 0.548 | 0.744 | 1.146 | 0.156** |
| T3Pub - T2Pub | 0.301*** | 0.555*** | 0.523 | 0.151*** |
| T4Pub - T3Pub | 0.744 | 0.870 | 1.000 | 0.519* |
| T5 - T4Pub | 0.559** | 0.764** | 0.700 | 0.372* |
| T2Pvt - T1Pvt | 0.296*** | 0.588*** | 0.216*** | 0.411 |
| T3Pvt - T2Pvt | 0.427** | 0.728** | 0.204*** | 1.625 |
| T4Pvt - T3Pvt | 0.748 | 0.912 | 0.471 | 1.232 |
| Demographics | No | Yes | No | No |
| Individual FE | Yes | No | Yes | Yes |
| Pseudo R-squared | 0.623 | 0.186 | 0.642 | 0.617 |
| N | 1790 | 1790 | 1080 | 710 |

Note: This table reports odds ratios from logit regressions, with standard errors clustered at the individual level. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

4 Relationship with the literature

4.1 Moral licensing

To the best of our knowledge, this paper presents the first formalization of the moral licensing effect and, accordingly, model-based experimental evidence. We do so in the two-moral-choice context of public and private behaviors. The essence of moral licensing is the existence of (at least) two moral choices, one used to exonerate the individual from their moral duty in the other. As discussed below, most experiments on moral licensing have used a self-signaling paradigm rather than a social-signaling one (and therefore studied individual moral licensing rather than social moral licensing). Yet, in the tradition of Adam Smith’s “impartial spectator” and the recent treatments of self- and social- signaling in the economics literature (e.g., Bénabou-Tirole (2004) and the *Journal of Economic Perspectives* (2016) symposium on motivated beliefs), there is no essential difference between the two forms of signaling. Our main model, developed in Section 2.2, is a social signaling one, as it relies on the heterogeneous visibility to different audiences. But its generalization in Section 2.3 accommodates as a special case self-signaling (in which the audience in the different tasks is the same – the inner self, and the heterogeneity among tasks is driven by other factors such as the cost or frequency of acting)³⁶. On the theory front, we are not aware of any modeling of moral licensing that takes on board a multiplicity of agent choices and analyzes whether these choices are substitutes (one good deed “motivating” a bad one) or complements (good deeds being self-reinforcing). The specific context – with public and private spheres – is also novel.

(a) Comparison with alternative explanations for moral licensing

- *Moral wiggle room.* Moral wiggle room experiments have repeatedly demonstrated the important role of “feeble excuses” or “plausible deniability” in reducing prosocial behavior. Whether based on information avoidance (as in the seminal experiment of Dana et al 2007), on the possibility of delegating dirty work to a like-minded third party (as in Hamman et al 2010), or on the avoidance of the ask (as in Andreoni et al 2017), the inner spectator deludes himself as to the reality of his selfishness thanks to the slight haze around the mapping from actions to consequences. Because the

³⁶Another reason not to restrict the study of moral licensing to self-signaling is that the moral licensing literature itself does not exclude social signaling. Noting the literature’s focus on individual moral licensing, Lasarov and Hoffmann (2020) advocate future research into social moral licensing, a call for research that our paper echoes. Merritt et al (2010)’s definition of moral licensing captures individuals’ fear of appearing immoral to others. Likewise, the Wikipedia’s definition of moral licensing refers to individuals’ worry about the consequences of immoral behavior and so involves social signaling: moral licensing is used to avoid public shaming and thus has a social component. List and Momeni (2020) argue that “*CSR can increase worker misbehavior through moral licensing. Prosocial behavior is promoted in part by self- and social-image motives: People act prosocially, in part, to signal to themselves (and to others) that they are good and moral individuals... Relatedly, moral licensing has been raised as a potential dark side of CSR’s appeal to image concerns because people who have recently ‘done good’ in one dimension may feel immunized against negative (social or self) inferences, and thus later on act less morally constrained*”.

“excuse” vocabulary is often used in the context of moral licensing, it is important to note that neither our theory nor our experiment is driven by moral wiggle room, due to the absence of the “haze” in our settings. Here moral licensing occurs under a good understanding and a clear conscience of the nature of the game being played, and not from a feeble excuse. This is not to say that moral licensing and moral wiggle room are incompatible, but rather that they are two separate conceptual objects.

- *Altruism budget.* On the basis of experimental evidence, Gee and Meer (2019) posit the existence of an “altruism budget”. Our moral licensing theory provides some foundation for their altruism budget by explaining how reputational concerns can create a substitutability between moral acts. It further qualifies the validity of the altruism-budget assumption by relating the tightness of the altruism budget to the structure of information and by pointing out that the substitutability/complementarity property is price contingent and that both patterns can emerge (see Section 2.3 and part (e) of Appendix A). Furthermore, the “altruism budget” posits that the total prosociality is fixed for a give c , while our study presents a richer picture in which the overall prosociality changes non-monotonically when transparency evolves. These theoretical predictions are validated by the experimental results.

(b) Comparison with other experiments

On the empirical front, moral licensing has been studied in a large literature in psychology (see Footnote 6) and more recently managerial economics (see e.g., List-Momeni 2020 for references), social economics (e.g., Schmitz 2019), environmental economics (e.g., Dorner 2019), consumer behavior (e.g. Kouchaki-Jami 2018, Engel-Szech 2020), and legal studies (e.g., Cain-Loewenstein-Moore 2005, 2011). Interestingly, List and Momeni’s experiment finds that corporate social responsibility may backfire and induce worker misbehavior. The worker has a single action (good or bad production behavior). In the corporate social responsibility (CSR) treatment, the firm gives money to a charity, while no such money is announced in the control treatment. The CSR treatment itself is decomposed into two frames: one in which CSR is presented as a choice of the firm (which it always is, as the agent is informed that her compensation scheme is unrelated to CSR), and one in which it is added that the donation is “on behalf of the worker”.³⁷ The latter framing treatment is the most prone to misbehavior. Our work differs from this line of research by considering multiple subject actions (making it closer to the standard effect documented in psychology), by not relying on framing, and finally by formally rationalizing why agent’s moral actions are substitutes.

Exley’s (2018) experiment on image concerns reports empirical findings that are similar to a moral licensing effect: Disclosing donors’ previous giving behavior to third-party

³⁷A similar trick is used in Kouchaki-Jami (2018), in which workers are replaced by consumers.

observers reduces donors’ current contribution, while providing such information mitigates the crowding-out effect of monetary incentives on contributions. Our experiment includes varied mixtures of public and private spheres, which differ in observability but are otherwise identical. With such a design, we are able to demonstrate some novel effects that are different from Exley (2018), List-Momeni (2020), and the psychology literature on moral licensing, including the misallocation between public and private spheres, the cheap signaling effect in the public sphere, and the reduced contributions in both spheres when the public sphere expands.

In Cain et al (2005, 2011), an advisor/sender has a conflict of interest when making a recommendation to a decision-maker/receiver. Comparing mandatory disclosure of the conflict of interest to the absence of disclosure (either way, the sender has no choice regarding disclosure or its absence), they find that making the sender disclose the conflict of interest may impair the integrity of her recommendation to the receiver; the disclosure makes the sender more prone to select selfish choices, that is to recommend her materially beneficial option at the expense of the receiver. The disclosure of a conflict of interest serves as a moral license. However, their moral licensing effect has a different nature from most of the other moral licensing studies.³⁸

4.2 Other related literatures

The model expositied in Section 2.1 follows the theoretical and empirical paradigm of behavior driven by explicit, implicit and image motivations (Bénabou-Tirole 2006). While this literature emphasizes the role of transparency in incentivizing socially desirable actions, Section 2.2 qualifies this conventional wisdom. Through its theme, the paper also fits within the broader privacy literature.³⁹

Section 2.2, which emphasizes the difference in behavior in the public and private spheres, speaks to the multitasking literature (Holmström-Milgrom 1991); in our paper,

³⁸In the absence of disclosure choice, moral licensing reflects the idea that a competitive behavior may be licit (or at least less objectionable) if there is a level-playing field (including in the informational domain), but not otherwise. This leaves open the question of the formalization of this effect. One potential (and admittedly ad-hoc) rationale for the Cain et al experiments is that the level-playing field imperative shifts (uniformly, say) the motivation of agents: If θ is an index of absence of level-playing field, agent v ’s intrinsic motivation is $v + \theta$. Then as θ increases (the level-playing field is tilted), the fraction of prosocial behavior (truthful recommendation) increases.

³⁹The case against transparency in the economics literature has several branches. The first branch focuses on abuses by the receiver of the information. Sellers may capture too much of the consumer surplus as they acquire much information about individual tastes (Acquisti et al. 2016). They may exploit the consumer’s impulsiveness or her incomplete information (people are rarely aware of privacy threats). Information collection may as well destroy insurance (Hirshleifer 1971), most prominently in the realm of health insurance, but also by amplifying the impact of behavioral or information-collection mistakes: subjective profiling (“lazy”, “alcoholic”...) may deprive the individual from a job, data dissemination may make a person into a social pariah, etc. Other concerns arising on the receiver side include surveillance by the state and platforms (Tirole 2021) and the violation of the right to be forgotten (the loss of a second chance).

though, different tasks do not compete for resources (the cost of accomplishing them is additive). Relative to the multitask career concerns model (Dewatripont et al. 1999), in which performance in a task may reveal information about a trait, reducing signaling incentives on other tasks, our framework puts more structure and accordingly delivers new results, such as the existence of moral licensing, the impact of a change in the visibility of activities, and the pattern of substitutability/complementarity of moral choices. Bernheim and Bodoh-Creed (2023) provide a bound on signaling distortions as a function of the number of interactions an agent is engaged in. There are two major differences in focus between their paper and the analysis of the expansion of the public sphere in Section 2.2. First, in the latter the public sphere inflates at the expense of the private one and the emphasis is on the impact on behavior in the private sphere. Second, the results hinge on the existence of multiple audiences with different information structures.

It is well documented that individuals tend to change their behavior when their actions are observed by others (see, e.g., the studies cited in Footnote 4). Thus, a classic approach to creating image concerns in the laboratory is to rely on pure social esteem concerns: Subjects make decisions in front of others and/or with their identity being disclosed, or have to tell others about their choices, and are thus subjected to social approval or opprobrium. This approach is used for instance by Andreoni-Petrie (2004), Rege-Telle (2004), Andreoni-Bernheim (2009), Ariely et al. (2009), Jones-Linardi (2014), and Butera et al. (2022). A different approach to creating image concerns in the lab goes beyond pure esteem concerns and allows observers to take a subsequent action that influences the subject’s payoff. The observers’ reciprocal behavior is thus the driver of the subject’s image concerns.

Our experiment follows the second approach and relies on third-party observers’ indirect reciprocity to generate image concerns. Nowak and Sigmund (1998a, b, 2005) demonstrated the role of indirect reciprocity in enabling cooperation among strangers, a hypothesis confirmed by the experimental literature (e.g., Bolton et al. 2005, Seinen-Schram 2006, and Engelmann-Fischbacher 2009). These experimental studies typically focus on repeated-interaction settings, where agents have both an intrinsic motivation to reward others with good reputations (indirect reciprocity) and incentives to build good reputations for themselves. To test our (static) theory cleanly, our experiment creates image concerns in a static setting, in line with the indirect reciprocity approach of Coffman (2011), Exley (2018) and Bolton et al. (2021).

Importantly, the focus of our study is the two-way interaction between public and private spheres, not the exact image-concern-generating mechanism. We are agnostic as to which of the “public shaming or nonmonetary reward approach” (the subject discloses their behavior in front of third-party observers, and is thus subjected to social approval or opprobrium, but without cash transfer implications) and the “monetary reward approach” (third-party observers directly reward the subject, as in our experiment and in Coffman (2011), Exley (2018) and Bolton et al. (2021)) is more appropriate in general to create image concerns. Our choice of the latter is pragmatic and is guided by the desire to

cleanly test the theory. While the cheap signaling and moral licensing effects are robust to non-additive image payoffs, our theoretical framework presumes additive/linear image payoffs, i.e., the total image concern is equal to the sum of the image concerns vis-à-vis each observer. The monetary reward approach allows us to stick as closely as possible to the theoretical framework. Under the public-shaming approach, however, the image concern may well be non-linear both in the number of observers and in the number of good deeds, confounding the identification of the central feature of our study, the spillovers across decisions.

With this central feature, a major difference between our study and the above-mentioned experimental studies lies in our distinction between public and private spheres. The literature on indirect reciprocity focuses on, in our terminology, the single-task, “all public” setting where all behaviors enter image scoring. Our experiment, with co-existence of public and private spheres, allows us to investigate individuals’ reallocation of contributive efforts between the two spheres as well as the effect of an expansion in the public sphere.

5 Conclusion

With tech giants’ and governments’ rapid deployment of data technologies, individual behavior becomes more transparent in many aspects of life. Optimists argue that transparency will promote socially valued behavior. Our paper challenges this “conventional wisdom” both theoretically and empirically. Agents’ social interactions involve multi-tasking between their public and private spheres. An increase in transparency generates crowding out in both, public and private, spheres: When the public sphere expands, public sphere signaling is no longer cheap, reducing prosociality in the public sphere. This reduced prosociality in turn augments moral licensing in the private sphere, which crowds out prosociality in that sphere; so agents behave less prosocially in both spheres. The composition effect and the level effect work in opposite directions, making the aggregate impact of public sphere expansion on prosociality ambiguous; this impact is always negative over some range, though.

We designed an experiment to test the theory. In the experiment, indirect reciprocity from third-party observers drives image concerns. The findings from the experiment support the key theoretical predictions of our model. (1) The cheap-signaling effect generates a higher prosociality in the public sphere than in the all-private treatment. (2) Prosociality in the public sphere decreases with the size of the public sphere. (3) The moral-licensing effect generates a lower prosociality in the private sphere than in the all-private treatment. (4) Prosociality in the private sphere decreases with the size of the public sphere. (5) The subjects misallocate efforts by behaving more prosocially in the public sphere than in the private sphere. (6) Overall prosociality may not increase monotonically with an expansion of the public sphere. While the experiment, which is closely matched with the theory,

validates the theoretical mechanisms, in future research, field experiments conducted in more natural settings would be valuable in confirming the generalizability of the empirical findings.

Our description of behavior in the public and private spheres does not reflect the real world's richness of image concerns. For example, the visibility of one's behavior is in part a choice of agents, who may emphasize private interactions or use technologies that encrypt data and preserve some privacy. Interestingly, the choice of activities and visibility itself would be a signal of prosociality. Second, the dichotomy between privately and publicly observed behaviors is simplistic. We all belong to overlapping social networks of very heterogeneous sizes. There is no doubt that the crowding-in and crowding-out effects associated with cheap signaling and moral licensing would remain relevant, but the overall pattern would again be richer. Third, one might extend the literature on moral wiggle room to multi-tasking environments and ask whether the reputation linkage across tasks intensifies or dulls the search for feeble excuses to "justify" immoral behavior. We leave these and other theoretical extensions and their testing for future research.

Acknowledgement

This paper was previously circulated with the title “Prosocial Behavior in Public and Private Spheres: Theory and Evidence.” The authors are grateful to Roland Bénabou, Armin Falk, John List, George Loewenstein, Karine Van der Straeten, and participants in seminars and conferences at the Chinese University of Hong Kong, Princeton University, Toulouse School of Economics (TSE), the Durham-Wuhan Workshop on Experimental Economics, the 2023 Econometric Society European Meetings (Barcelona), the 2023 ESA World Meeting (Lyon), and the 2023 China BEEF for helpful comments, and to Amirreza Ahmadzadeh, Bin Cheng and Paul-Henri Moisson for able research assistance. Pak Hung Lam participated in useful discussions on the experiment.

Funding and conflict of interest

This project received funding from the European Research Council (ERC) under the European Union’s Horizon 2020 research and innovation programme (grant agreement no. 669217 - ERC MARKLIM). Jean Tirole acknowledges funding from the French National Research Agency (ANR) under the Investments for the Future (Investissements d’Avenir) program, grant ANR-17-EURE-0010. He gratefully acknowledges the financial support of the TSE Digital Center (the list of sponsors is available at <https://www.tse-fr.eu/digital>). Fuhai Hong thanks the Lingnan University ERSS Fund for financial support. Declaration: No conflict of interest.

References

- Acquisti, A., Taylor, C., and L. Wagman (2016), “The Economics of Privacy”, *Journal of Economic Literature*, 54: 442–492.
- Algan, Y., Benkler, Y., Fuster-Morell, M. and J. Hergueux (2016), “Cooperation in a Peer Production Economy: Experimental Evidence from Wikipedia”, Sciences Po Working Paper, November.
- Ali, S. N., and R. Bénabou (2020), “Image Versus Information: Changing Societal Norms and Optimal Privacy”, *American Economic Journal: Microeconomics*, 12(3): 116–164.
- Andreoni, J., and R. Petrie (2004), “Public Goods Experiments without Confidentiality: A Glimpse into Fund-Raising”, *Journal of Public Economics*, 88(7): 1605-1623.
- Andreoni, J., Rao, J. and H. Trachtman (2017), “Avoiding the Ask: A Field Experiment on Altruism, Empathy, and Charitable Giving”, *Journal of Political Economy*, 125(3): 625-653.
- Andreoni, J., and D. Bernheim (2009), “Social Image and the 50–50 Norm: A Theoretical and Experimental Analysis of Audience Effects”, *Econometrica*, 77(5): 1607-1636.
- Ariely, D., Bracha, A., and S. Meier (2009), “Doing Good or Doing Well? Image Motivation and Monetary Incentives in Behaving Prosocially”, *American Economic Review*, 99(1): 544–555.
- Ashraf, N., and O. Bandiera (2018), “Social Incentives in Organizations?” *Annual Review of Economics*, 10: 439–463.
- Ashraf, N., Bandiera, O., and J. Kelsey (2014), “No Margin, No Mission? A Field Experiment on Incentives for Public Services Delivery”, *Journal of Public Economics*, 120: 1–17.
- Bénabou, R., and J. Tirole (2004), “Willpower and Personal Rules”, *Journal of Political Economy*, 112(4): 848-886.
- Bénabou, R., and J. Tirole (2006), “Incentives and Prosocial Behavior”, *American Economic Review*, 96(5): 1652–1678.
- Bénabou, R., and J. Tirole (2011a), “Identity, Morals and Taboos: Beliefs as Assets”, *Quarterly Journal of Economics*, 126(2): 805–855.
- Bénabou, R., and J. Tirole (2011b), “Laws and Norms”, Technical report, National Bureau of Economic Research.
- Bernheim, B. D., and A. L. Bodoh-Creed (2023), “Pervasive signaling”, *Theoretical Economics*, 18(1): 163-196.
- Bolton, G. E., Katok, E., and A. Ockenfels (2005), “Cooperation among Strangers with Limited Information about Reputation”, *Journal of Public Economics*, 89(8): 1457–1468.
- Bolton, G., Dimant, E., and U. Schmidt (2021), “Observability and Social Image: On the Robustness and Fragility of Reciprocity”, *Journal of Economic Behavior and Organization*, 191: 946-964.
- Bursztyn, L., and R. Jensen (2017), “Social Image and Economic Behavior in the Field: Identifying, Understanding and Shaping Social Pressure”, *Annual Review of Economics*, 9: 131–153.

- Butera, L., Metcalfe, R., Morrison, W., and D. Taubinsky (2022), “Measuring the Welfare Effects of Shame and Pride”, *American Economic Review*, 112(1): 122-68.
- Cain, D. M., Loewenstein, G., and D. A. Moore (2005), “The Dirt on Coming Clean: Perverse Effects of Disclosing Conflicts of Interest”, *The Journal of Legal Studies*, 34(1): 1-25.
- Cain, D. M., Loewenstein, G., and D. A. Moore (2011), “When Sunlight Fails to Disinfect: Understanding the Perverse Effects of Disclosing Conflicts of Interest”, *Journal of Consumer Research*, 37(5): 836-857.
- Coffman, L. C. (2011), “Intermediation Reduces Punishment (and Reward)”, *American Economic Journal: Microeconomics* 3(4): 77-106.
- Dai, X. (2020), “Toward A Reputation State: A Comprehensive View of China’s Social Credit System Project”, In *Social Credit Rating*, Springer Fachmedien Wiesbaden.
- Dana, J., Weber, R., and J. Kuang (2007) “Exploiting Moral Wriggle Room: Experiments Demonstrating an Illusory Preference for Fairness”, *Economic Theory* 33(1): 67–80.
- Daughety, A., and J. Reinganum (2010), “Public Goods, Social Pressure, and the Choice between Privacy and Publicity”, *American Economic Journal: Microeconomics*, 2(2): 191–221.
- DellaVigna, S., List, J., Malmendier, U., and G. Rao (2017), “Voting to Tell Others”, *Review of Economic Studies*, 84: 143–181.
- Dewatripont, M., Jewitt, I., and J. Tirole (1999), “The Economics of Career Concerns, I: Comparing Information Structure”, *Review of Economic Studies*, 66(1): 183–198.
- Dorner, Z. (2019), “A Behavioral Rebound Effect”, *Journal of Environmental Economics and Management*, 98: 102257.
- Effron, D., Cameron, J., and B. Monin (2009), “Endorsing Obama Licenses Favoring Whites”, *Journal of Experimental Social Psychology*, 45: 590–593.
- Effron, D., Miller, D., and B. Monin (2012), “Inventing Racist Roads Not Taken: The Licensing Effect of Immoral Counterfactual Behaviors”, *Journal of Personality and Social Psychology*, 103(6): 916–932.
- Engel, J., and N. Szech (2020) “A Little Good is Good Enough: Ethical Consumption, Cheap Excuses, and Moral Self-Licensing”, *PLOS ONE*, January 15, pages 1-9.
- Engelmann, D., and U. Fischbacher (2009), “Indirect Reciprocity and Strategic Reputation Building in an Experimental Helping Game”, *Games and Economic Behavior*, 67(2): 399–407.
- Exley, C. (2018), “Incentives for Prosocial Behavior: The Role of Reputations”, *Management Science*, 64 (5): 2460-2471.
- Freeman, R. (1997), “Working for Nothing: The Supply of Volunteer Labor”, *Journal of Labor Economics*, 15(1): S140–66.
- Funk, P. (2010), “Social Incentives and Voter Turnout: Evidence from the Swiss Mail Ballot System”, *Journal of the European Economic Association*, 8(5): 1077–1103.
- Gee, L. K., and J. Meer (2019), “The Altruism Budget: Measuring and Encouraging Charitable Giving”, National Bureau of Economic Research.

- Gneezy, U., Imas, A., and K. Madarász (2014) “Conscience Accounting: Emotion Dynamics and Social Behavior”, *Management Science*, 60 (11): 2645-2658.
- Hamman, J., Loewenstein, G. and R. Weber (2010), “Self-Interest through Delegation: An Additional Rationale for the Principal-Agent Relationship”, *American Economic Review*, 100 (4): 1826-46.
- Holm, S. (1979), “A Simple Sequentially Rejective Multiple Test Procedure”, *Scandinavian Journal of Statistics*, 6(2), 65–70.
- Holmström, B., and P. Milgrom (1991), “Multitask Principal-Agent Analyses: Incentive Contracts, Asset Ownership, and Job Design”, *Journal of Law, Economics, and Organization*, 7: 24–52.
- Hirshleifer, J. (1971), “The Private and Social Value of Information and the Reward to Inventive Activity”, *American Economic Review*, 61: 561–574.
- Jewitt, I. (2004), “Notes on the Shape of Distributions”, unpublished.
- Jones, D., and S. Linardi (2014), “Wallflowers: Experimental Evidence of an Aversion to Standing Out”, *Management Science*, 60 (7): 1757-1771.
- Journal of Economic Perspectives* (2016), “Symposium: Motivated Beliefs”, 30(3): 133-212.
- Kouchaki M., and A. Jami (2018) “Everything We Do, You Do: The Licensing Effect of Prosocial Marketing Messages on Consumer Behavior”, *Management Science*, 64(1):102–111.
- Lacetera, N., Macis, M., and R. Slonim (2012), “Will There Be Blood? Incentives and Displacement Effects in Pro-social Behavior”, *American Economic Journal: Economic Policy*, 4(1): 186–223.
- Lanier, J. (2010), “Missing Persons”, In *You Are Not a Gadget*, London: Penguin.
- Lasarov, W., and S. Hoffmann (2020) “Social Moral Licensing”, *Journal of Business Ethics*, 165:45–66.
- List, J. (2024), “Within-Subject Experimental Designs”, mimeo.
- List, J., and F. Momeni (2020) “When Corporate Social Responsibility Backfires: Evidence from a Natural Field Experiment”, *Management Science*, 67(1): 8-21.
- Merritt, A., Effron, D., and B. Monin (2010) “Moral Self-Licensing: When Being Good Frees Us to Be Bad”, *Social and Personality Psychology Compass*, 4(5), 344–357.
- Merritt, A., Effron, D., Fein, S., Savitsky, K., Tuller, D., and B. Monin (2012), “The Strategic Pursuit of Moral Credentials”, *Journal of Experimental Social Psychology*, 48: 774–777.
- Moffatt P. (2015), *Experiments: Econometrics for experimental economics*, Palgrave Macmillan.
- Monin, B., and D. Miller (2001), “Moral Credentials and the Expression of Prejudice,” *Journal of Personality and Social Psychology*, 81(1): 33–43.
- Nelson, L., and U. Simonsohn (2014), “Thirty-Somethings Are Shrinking and Other U-shaped Challenges”, unpublished.

- Nowak, M. A., and K. Sigmund (1998a), “The Dynamics of Indirect Reciprocity”, *Journal of Theoretical Biology*, 194(4): 561–574.
- Nowak, M. A., and K. Sigmund (1998b), “Evolution of Indirect Reciprocity by Image Scoring”, *Nature*, 393(6685): 573–577.
- Nowak, M. A., and K. Sigmund (2005), “Evolution of Indirect Reciprocity”, *Nature*, 437(7063): 1291–1298.
- Perez-Truglia, R., and G. Cruces (2017), “Partisan Interactions: Evidence from a Field Experiment in the United States”, *Journal of Political Economy*, 125(4): 1208–1243.
- Rege, M., and K. Telle (2004), “The Impact of Social Approval and Framing on Cooperation in Public Good Situations”, *Journal of Public Economics*, 88(7): 1625–1644.
- Romano, J.P., and M. Wolf (2005a), “Exact and Approximate Stepdown Methods for Multiple Hypothesis Testing”, *Journal of the American Statistical Association*, 100(469): 94–108.
- Romano, J.P., and M. Wolf (2005b), “Stepwise Multiple Testing as Formalized Data Snooping”, *Econometrica*, 73(4): 1237–1282.
- Romano, J.P., and M. Wolf (2016), “Efficient Computation of Adjusted P-values for Resampling-based Stepdown Multiple Testing”, *Statistics and Probability Letters*, 113: 38–40.
- Russell, S. (2019), *Human Compatible: Artificial Intelligence and the Problem of Control*, Viking, Penguin Random House.
- Schmitz, J. (2019), “Temporal Dynamics of Pro-social Behavior: An Experimental Analysis”, *Experimental Economics*, 22: 1–23.
- Seinen, I., and A. Schram (2006), “Social Status and Group Norms: Indirect Reciprocity in a Repeated Helping Experiment”, *European Economic Review*, 50(3): 581–602.
- Tirole, J. (2021), “Digital Dystopia”, *American Economic Review*, 111(6): 2007–2048.
- Zhong, C., and K. Liljenquist (2006) “Washing Away your Sins: Threatened Morality and Physical Cleansing”, *Science*, 313(5792): 1451–1452.

Appendix A: Theory

(a) Welfare. *Single task.* Because reputation is a “positional good”, the agent’s reputational payoff has no social value: an agent’s gain is another agent’s loss. And indeed, our assumptions imply that the average reputation is a constant, \bar{v} , and so we ignore reputational concerns in W :

$$W = (v_i - c + e)a_i,$$

where e is the social value of the induced externality. Changing the expression of W would affect the regions for over- and under-provision of prosocial behavior, but the qualitative insights would remain the same.⁴⁰

Let us index the socially optimal behavior with superscript “ SO ”. From the expression of W , we see that agent i should choose $a_i = 1$ for all j if $v_i \geq v^{SO}$ (and $a_i = 0$ for all j otherwise), where

$$v^{SO} - c + e = 0.$$

In the single-task benchmark case, there is underprovision (resp. overprovision) if $v^{SO} < v^*(\mu)$ (resp. $v^{SO} > v^*(\mu)$). Underprovision therefore corresponds to $e > \mu\Delta(v^*(\mu))$. Thus, more transparency, which marginally increases μ and so raises prosociality, increases welfare when $e > \mu\Delta(v^*(\mu))$.

Multiple tasks. In the deterministic symmetric equilibrium with public and private spheres, welfare is given by

$$W(t) = t \int_{v^t}^{+\infty} (v - c + e)dF(v) + s \int_{v^s}^{+\infty} (v - c + e)dF(v).$$

And so

$$\frac{dW}{dt} = \left[\int_{v^t}^{v^s} (v - c + e)dF(v) \right] - f(v^t) \frac{dv^t}{dt} t(v^t - c + e) - f(v^s) \frac{dv^s}{dt} s(v^s - c + e).$$

Part (d) in this Appendix shows for a uniform distribution that, even if prosocial behavior is socially desirable, $d\bar{a}/dt < 0$ and $dW/dt < 0$ over some interval of t .

(b) Off-the-equilibrium-path beliefs. Let $\bar{a}^t \in [0, 1]$ denote the average contribution in the public sphere. Under a deterministic and symmetric behavior, $\bar{a}^t \in \{0, 1\}$. Suppose for conciseness that $v^s < \sup v$, so both $a_{ij} = 0$ and $a_{ij} = 1$ are on the equilibrium path for (ij) in the private sphere. If $v^t = 0$, specify that $\hat{v}_{ji} = 0$ if $\bar{a}^t < 1$ where \hat{v}_{ji} is j ’s posterior estimate of v_i . If $v^t > 0$, then both $\bar{a}^t = 0$ and $\bar{a}^t = 1$ are on-path behaviors. For $\bar{a}^t < 1$, set $\hat{v}_{ji} = M^-(v^t)$ (which, incidentally, covers the case $v^t = 0$) no matter whether the (ij) relationship is in the public or private sphere. These beliefs sustain the deterministic and

⁴⁰The expression of W could be modified in at least two ways. First, reputation might not be positional (the reputation-stealing game might not be a zero-sum game). Second, the intrinsic motivation v may or may not be part of welfare.

symmetric behavior described by (3) and (4) as an equilibrium.

(c) Uniqueness of the deterministic symmetric equilibrium.

Lemma 1 *A sufficient condition for there to always be more contributions in the public sphere ($v^s \geq v^t$) in any equilibrium is that the density of the type distribution be non-increasing (e.g., uniform).*

Proof of Lemma 1. Suppose to the contrary that $v^s < v^t$ and let $M(v_0, v_1)$ denote the mean of v over the interval $[v_0, v_1]$.

Behavior in the private sphere, being unobservable except to the counterparty, does not impact the reputation in the public sphere. So, for any $v_i \in [v^s, v^t]$,

$$s [(v_i - c) + \mu[M(v^s, v^t) - M^-(v^s)]] \geq 0.$$

Similarly the fact that in this interval, agents do not want to contribute publicly implies that:

$$t(v_i - c) + \mu [s[M^+(v^t) - M(v^s, v^t)] + t[M^+(v^t) - M^-(v^t)]] \leq 0.$$

These two inequalities are inconsistent if

$$M(v^s, v^t) - M^-(v^s) < [M^+(v^t) - M^-(v^t)] + \frac{s}{t}[M^+(v^t) - M(v^s, v^t)].$$

The latter condition is satisfied in particular (for $s > 0$) if for $v^s < v^t$

$$M^+(v^t) - M^-(v^t) \geq M(v^s, v^t) - M^-(v^s). \tag{7}$$

Inequality (7) is satisfied at $v^s = v^t$ (since $M^+(v^t) \geq v^t$). Furthermore, applying Jewitt (2004)'s lemma on $[0, v^t]$, $M(v^s, v^t) - M^-(v^s)$ is non-decreasing in v^s if the density f is non-increasing. ■

Finally, one cannot guarantee that (4) has a unique solution, unless $f' \leq 0$ and so the additional term is non-increasing with prosocial behavior in the public sphere. Nonetheless, it can be shown that the prosocial behavior in the public and private spheres is decreasing in t at stable equilibria.

(d) Uniform distribution. Suppose that $v \sim U[0, 1]$. In the single-task case, letting $0 < c - \mu/2 < 1$ (to ensure an interior solution),

$$v^* = c - \frac{\mu}{2}.$$

Then $\bar{a}(0) = \bar{a}(1) = 1 - (c - \frac{\mu}{2})$. More prosocial behavior is desirable provided that $e > \mu/2$. We will actually make the slightly stronger assumption that $e > (c + \mu/2)/2$.

In the multi-task case and in the interior region ($0 < v^t < v^s < 1$), (3) and (4) become

$$v^s - c + \frac{\mu}{2}(1 - v^t) = 0 \quad (3'')$$

$$v^t - c + \frac{\mu}{2}\left(1 + \frac{s}{t}v^s\right) = 0. \quad (4'')$$

Let

$$0 < t_0 \equiv \frac{1}{1 + \frac{2}{\mu}} < t_1 \equiv \min \left\{ \frac{\mu^2}{4(c-1)\left(\frac{\mu}{2} + 1\right)}, 1 \right\}.$$

Then for all $t \leq t_0$, $v^t = 0$, and $v^s = v^*$. The overall contribution

$$\bar{a}(t) = \begin{cases} 1 - (1-t)\left(c - \frac{\mu}{2}\right) & \text{for } t \leq t_0 \\ 1 + \frac{1-2c}{\frac{2}{\mu} - \frac{\mu}{2} + \frac{\mu}{2t}} & \text{for } t_0 \leq t \leq t_1 \\ \frac{\mu}{2} - t(c-1) & \text{for } t \geq t_1 \text{ (if } t_1 < 1\text{)}. \end{cases}$$

So $\bar{a}(t)$ increases linearly with t in the first region, decreases in the second region (is convex if $\mu < 2$, concave if $\mu > 2$), and decreases linearly in the third region (if it exists).

Let $X^t \equiv f(v^t)\frac{dv^t}{dt}t > 0$ and $X^s \equiv f(v^s)\frac{dv^s}{dt}s > 0$. For $t_0 \leq t \leq t_1$,

$$\begin{aligned} \frac{dW}{dt} &= [F(v^s) - F(v^t)] [M(v^t, v^s) - c + e] - X^t(v^t - c + e) - X^s(v^s - c + e) \\ &= \underbrace{[M(v^t, v^s) - c + e]}_+ \frac{d\bar{a}}{dt} + [(X^t + X^s)M(v^t, v^s) - X^t v^t - X^s v^s], \end{aligned}$$

where the second line uses $\frac{d\bar{a}}{dt} = [F(v^s) - F(v^t)] - X^t - X^s$ for $t_0 \leq t \leq t_1$. That $M(v^t, v^s) - c + e > 0$ results from the fact that $M(v^t, v^s) \geq M(0, v^*) = (c - \mu/2)/2$ and the assumption that $e > (c + \mu/2)/2$. Note that

$$\begin{aligned} &(X^t + X^s) M(v^t, v^s) - X^t v^t - X^s v^s \\ &= \frac{X^t v^s + X^s v^t - X^t v^t - X^s v^s}{2} \\ &= \frac{v^s - v^t}{2} \left(\frac{dv^t}{dt} t - \frac{dv^s}{dt} s \right) \\ &= \underbrace{\frac{v^s - v^t}{2}}_+ \underbrace{\frac{dv^t}{dt}}_+ \left[\left(1 + \frac{\mu}{2}\right) t - \frac{\mu}{2} \right], \end{aligned}$$

where the third line uses $X^t = f(v^t)\frac{dv^t}{dt}t = \frac{dv^t}{dt}t$ and $X^s = f(v^s)\frac{dv^s}{dt}s = \frac{dv^s}{dt}s$ given

$v \sim U[0, 1]$, and the fourth line uses $\frac{dv^s}{dt} = \frac{\mu}{2} \frac{dv^t}{dt}$ (by (3')). Thus,

$$\lim_{t \rightarrow t_0^+} [(X^t + X^s)M(v^t, v^s) - X^t v^t - X^s v^s] = 0.$$

Note that $\lim_{t \rightarrow t_0^+} [M(v^t, v^s) - c + e] > 0$, and $\lim_{t \rightarrow t_0^+} \frac{d\bar{a}}{dt} < 0$. Therefore,

$$\lim_{t \rightarrow t_0^+} \frac{dW}{dt} = \lim_{t \rightarrow t_0^+} \underbrace{[M(v^t, v^s) - c + e]}_{-} \frac{d\bar{a}}{dt} + \lim_{t \rightarrow t_0^+} \underbrace{[(X^t + X^s)M(v^t, v^s) - X^t v^t - X^s v^s]}_{=0} < 0$$

To sum up, even though prosocial behavior is socially desirable, $\frac{d\bar{a}}{dt} < 0$ for $t \geq t_0$ and $\lim_{t \rightarrow t_0^+} \frac{dW}{dt} < 0$: Transparency reduces both the overall contribution and welfare to the right of t_0 .

(e) Complementarity or substitutability? Equipped with Propositions 3 and 4, we can identify when moral choices are substitutes (one good deed “motivating” a bad one) or complements (good deeds being self-reinforcing). Substitutability and complementarity are usually defined through the reaction of the vector of “consumptions” (here of good deeds) to price movements: For example, when one changes the price of good behavior in the public sphere, $\partial \bar{a}^s / \partial c^t < 0$ (resp. $\partial \bar{a}^s / \partial c^t > 0$) indicates a complementarity (resp. a substitutability). Proposition 3(iii) shows that the moral behaviors in the two spheres co-move when the price c^t varies and behavior is more moral in the public sphere ($\bar{a}^s < \bar{a}^t$). In contrast, Proposition 4 and the moral-licensing terminology suggest a pattern of substitutability.

To reconcile these apparently contradictory observations, observe that while $\partial \bar{a}^s / \partial c^t < 0$ when behavior is more moral in the public sphere ($\bar{a}^s < \bar{a}^t$), in contrast $\partial \bar{a}^s / \partial c^t > 0$ when behavior is more moral in the private sphere ($\bar{a}^s > \bar{a}^t$): the pattern of complementarity/substitutability is *price contingent*. The cheap-signaling effect implies that, *ceteris paribus*, behavior is more moral in the public sphere; however, for a sufficiently high price c^t , moral behavior in the public sphere is rare, and its becoming more frequent (when c^t decreases) acts as a moral restraint on moral behavior in the private sphere (there is less glory to contributing in the private sphere).⁴¹ The control group in the moral-licensing psychology experiments has $c^t = +\infty$ (the subject has no prior opportunity to signal goodness) while the treatment group has a prior opportunity to demonstrate goodness. This is why a pattern of moral substitutability can emerge, and so there is no contradiction between Propositions 2(iii) and 3(iii) on the one hand, and Proposition 4 on the other hand.

⁴¹Note that (3') and (4') are obtained under the assumption that c^t is not much larger than c^s and so behavior is more moral in the public sphere. Here, c^t is large enough and so $v^s < v^t$ (behavior is more moral in the private sphere). Thus, (3') becomes $v^s - c^s + \mu[\Delta(v^s) - (M^+(v^s) - M(v^s, v^t))] = 0$. We have $c^t \downarrow \Rightarrow v^t \downarrow \Rightarrow v^s \uparrow$.

Appendix B: Power analysis

This appendix uses the data (i.e. with hindsight) to show that the two-stake approach economizes on sample size.

The theory predicts that $\bar{a}_{T_x}^k > \bar{a}_{T_y}^k$ for any $k \in \{s, t\}$ and any $x < y$. Take the test of $\bar{a}_{T_1}^t > \bar{a}_{T_3}^t$ for example. In the data, the High Stake sample does not discriminate well in the high contributions range: 68% of the subjects choose $a_{T_1}^t = a_{T_3}^t = 1$ (a lack of behavioral variation in the public sphere due to the large stake), and only 13.9% of the subjects choose $a_{T_1}^t = 1 > a_{T_3}^t = 0$ as the theory predicts and 10.2% of the subjects behave “irrationally” by choosing $a_{T_1}^t = 0 < a_{T_3}^t = 1$. The McNemar’s test (see Footnote 31) determines that a sample size of 1380 is needed to achieve 80% statistical power at a significance level of 5% to detect $\bar{a}_{T_1}^t > \bar{a}_{T_3}^t$. However, with the Low Stake, 29.6% of the subjects choose $a_{T_1}^t = 1 > a_{T_3}^t = 0$, while 2.8% of the subjects behave “irrationally” by choosing $a_{T_1}^t = 0 < a_{T_3}^t = 1$, and the McNemar’s test determines that a sample size of only 33 is needed to achieve 80% power at a significance level of 5%. Thus, to test $\bar{a}_{T_1}^t > \bar{a}_{T_3}^t$, the Low Stake requires a much smaller sample size than the High Stake, due to its larger behavioral variation in the public sphere.

Take the test of $\bar{a}_{T_1}^s > \bar{a}_{T_3}^s$ for example. In the data, the Low Stake sample does not discriminate well in the low contributions range: 47% of the subjects choose $a_{T_1}^s = a_{T_3}^s = 0$ (a lack of behavioral variation in the private sphere due to the small stake), while only 8.5% of the subjects choose $a_{T_1}^s = 1 > a_{T_3}^s = 0$, and 5.6% of the subjects behave “irrationally” by choosing $a_{T_1}^s = 0 < a_{T_3}^s = 1$. The McNemar’s test determines that a sample size of 1314 is needed to achieve 80% power at a significance level of 5% to detect $\bar{a}_{T_1}^s > \bar{a}_{T_3}^s$. However, with the High Stake, 36.6% of the subjects choose $a_{T_1}^s = 1 > a_{T_3}^s = 0$, and 4.2% of the subjects behave “irrationally” by choosing $a_{T_1}^s = 0 < a_{T_3}^s = 1$. The McNemar’s test determines that a sample size of only 29 is needed to achieve 80% power at a significance level of 5%. Thus, to test $\bar{a}_{T_1}^s > \bar{a}_{T_3}^s$, the High Stake requires a much smaller sample size than the Low Stake, due to its larger behavioral variation in the private sphere.

These two examples illustrate the benefits of using two stake levels in our test of the theory. Each stake, with relatively small sample size, could capture some aspects of the behavioral diversity predicted by the theory.

In general, with a certain stake of the “generosity score”, to test any relation $\bar{a}_{T_x}^k > \bar{a}_{T_y}^k$ for $x < y$ and for $k \in \{s, t\}$, if we assume that 5% of the subjects behave “irrationally” by choosing $a_{T_x}^k = 0 < a_{T_y}^k = 1$ and 30% of the subjects choose $a_{T_x}^k = 1 > a_{T_y}^k = 0$,⁴² the McNemar’s test determines that a sample size of 42 is needed to achieve 80% statistical power at a significance level of 5%. Our sample size for each of the stakes is higher than the bar.

⁴²Here, the numbers are chosen given the above two examples that use the Low Stake to capture $\bar{a}_{T_1}^t > \bar{a}_{T_3}^t$, and the High Stake to capture $\bar{a}_{T_1}^s > \bar{a}_{T_3}^s$.

Appendix C: Tables

Table C1: McNemar's tests on contribution (by stake)

| | High Stake | | Low Stake | |
|----------------|-------------------------------|-----------------------------|-------------------------------|-----------------------------|
| | Unadjusted <i>p</i> values | Adjusted <i>p</i> values | Unadjusted <i>p</i> values | Adjusted <i>p</i> values |
| T1Pub vs T1Pvt | 0.0006 | 0.0011 | 0.0000 | 0.0000 |
| T2Pub vs T2Pvt | 0.0000 | 0.0000 | 0.0000 | 0.0000 |
| T3Pub vs T3Pvt | 0.0000 | 0.0000 | 0.0001 | 0.0003 |
| T4Pub vs T4Pvt | 0.0000 | 0.0000 | 0.0067 | 0.0170 |
| T0 vs T5 | 0.1797 | 0.2516 | 0.5637 | 0.5637 |
| T1Pub vs T5 | 0.1779 | 0.2516 | 0.0000 | 0.0000 |
| T2Pub vs T5 | 0.0881 | 0.1423 | 0.0000 | 0.0000 |
| T3Pub vs T5 | 0.4669 | 0.5160 | 0.0073 | 0.0170 |
| T4Pub vs T5 | 0.3657 | 0.4517 | 0.0707 | 0.1061 |
| T1Pvt vs T0 | 0.0002 | 0.0004 | 0.0833 | 0.1166 |
| T2Pvt vs T0 | 0.0000 | 0.0000 | 0.0124 | 0.0237 |
| T3Pvt vs T0 | 0.0000 | 0.0000 | 0.0593 | 0.0958 |
| T4Pvt vs T0 | 0.0000 | 0.0000 | 0.1083 | 0.1269 |
| T1Pub vs T0 | 0.4652 | 0.5160 | 0.0000 | 0.0000 |
| T2Pub vs T0 | 0.2752 | 0.3612 | 0.0000 | 0.0000 |
| T3Pub vs T0 | 1.0000 | 1.0000 | 0.0116 | 0.0237 |
| T4Pub vs T0 | 1.0000 | 1.0000 | 0.1088 | 0.1269 |
| T1Pvt vs T5 | 0.0010 | 0.0018 | 0.1655 | 0.1738 |
| T2Pvt vs T5 | 0.0000 | 0.0000 | 0.0389 | 0.0681 |
| T3Pvt vs T5 | 0.0000 | 0.0000 | 0.1083 | 0.1269 |
| T4Pvt vs T5 | 0.0000 | 0.0000 | 0.1573 | 0.1738 |

Note: This table reports the *p* values of the McNemar's tests and the adjusted *p* values using the Holm-Bonferroni correction (Holm 1979).

Online Appendix: English translation of experimental instructions (Low Stake)

INSTRUCTIONS

Welcome to the experiment. Please read the instructions carefully; the payment you will obtain after the experiment depends both on your decisions and the decisions made by others. The experiment contains two stages. Stage I will be conducted today. It will take no longer than one hour. After Stage II is finished, we will pay you privately through WeChat transfer.

During the experiment, your identity will not be disclosed, and your decisions will not be associated with your identity. In order to ensure smooth implementation of the experiment, please do not leave the laboratory until the end of Stage I. During the experiment, please turn off your electronic devices.

Stage I

A Charity Fund

The following information is important for the decisions you will make. Our experiment is related to a charity fund on the Alipay charity platform that aims to help sanitation workers.



Description of Charity Fund

The city’s “beauticians”, early morning runners. A group of people are wearing orange cleaning uniforms, holding a broom in their hands, or riding a cleaning tricycle. Regardless of season and weather, they appear in the streets and alleys every morning on time, silently cleaning the environment for us.

Auntie Wang is 65 years old. Her son touched a high-voltage line at a very young age, resulting in incapacitation. The 40 years old son has been taken care of by Auntie Wang since that. She said: “People in our village know that we are poor, and introduced this job to me. I have been doing this job for over 20 years. Every morning I start to go to the working area around three o’clock. Although it is very hard, I treasure this job since my family relies on my salary. So I save as much as possible, and my lunch is also brought from home.” The volunteers saw that Auntie Wang’s lunchbox was filled with a cold steamed bun and leftovers from the last night.

There are many sanitation workers like Auntie Wang, who often eat cold food brought from home because they can’t afford to spend money on lunch. But in the summer, the food may spoil as the weather gets hot. Thus, having lunch becomes a big problem for them.

“Going out with the sun and dew, wearing the stars and moon to return.” With hardworking hands, the sanitation workers do their best to clean the environment and beautify the city. To thank them for their efforts, the “One Lunch Warms a City” charity fund is launched to provide sanitation workers with nutritious and delicious lunches.

Your task

You now have RMB100 as your endowment.

You will decide, under different circumstances, whether to help some sanitation workers by providing a nutritious lunch for them. If you choose to help, your cost of helping each worker is RMB20, and the worker will receive RMB30. You can also choose not to help.

There are 5 potential “**recipients**” (sanitation workers). Their role is passive. We will transfer the donations generated from the experiment in the name of Wuhan University after the experiment. Meanwhile, there are 5 “**observers**” who will observe some of your choices and will make decisions that influence your final income based on their observations in Stage II.

The 5 “observers” are anonymous students at Wuhan University.

The recipients are in either an “X” or a “Y” scenario: If the recipient is in the “X” scenario, your choice towards him/her will be observed by all observers. If the recipient is in the “Y” scenario, your choice towards him/her will only be observed by some observers (explained below).

Meanwhile, there are 6 possible parallel worlds: T0: All 5 recipients are in the Y-scenario.

T1: There is 1 recipient in the X-scenario and 4 recipients in the Y-scenario.

T2: There are 2 recipients in the X-scenario and 3 recipients in the Y-scenario.

T3: There are 3 recipients in the X-scenario and 2 recipients in the Y-scenario.

T4: There are 4 recipients in the X-scenario and 1 recipient in the Y-scenario.

T5: All 5 recipients are in the X-scenario.

You don't need to specify which recipient you are going to help; you only need to tell us about your decisions under different circumstances, i.e., in a parallel world T0, T1, T2, T3, T4, or T5, 1) in the X-scenario, whether you choose to help or not; 2) in the Y-scenario, whether you choose to help or not.

In the parallel world T0: all 5 recipients are in the Y-scenario.

If you choose to help, you will pay 100 yuan (to help 5 recipients), and your choice of helping will be observed by the 5 observers; if you choose not to help, you do not need to pay the 100 yuan, and your choice of not helping will also be observed by the 5 observers.

In the parallel world T1: There is 1 recipient in the X-scenario and 4 recipients in the Y-scenario.

T1 (X-scenario): If you choose to help, you will pay 20 yuan (to help 1 recipient), and your choice of helping will be observed by the 5 observers; if you choose not to help, you do not need to pay the 20 yuan, and your choice of not helping will also be observed by the 5 observers.

T1 (Y-scenario): If you choose to help, you will pay 80 yuan (to help 4 recipients), and your choice of helping will be observed by 4 observers; if you choose not to help, you do not need to pay the 80 yuan, and your choice of not helping will also be observed by these 4 observers.

In the parallel world T2: There are 2 recipients in the X-scenario and 3 recipients in the Y-scenario.

T2 (X-scenario): If you choose to help, you will pay 40 yuan (to help 2 recipients), and your choice of helping will be observed by the 5 observers; if you choose not to help, you do not need to pay the 40 yuan, and your choice of not helping will also be observed by the 5 observers.

T2 (Y-scenario): If you choose to help, you will pay 60 yuan (to help 3 recipients), and your choice of helping will be observed by 3 observers; if you choose not to help, you do not need to pay the 60 yuan, and your choice of not helping will also be observed by these 3 observers.

In the parallel world T3: There are 3 recipients in the X-scenario and 2 recipients in the Y-scenario.

T3 (X-scenario): If you choose to help, you will pay 60 yuan (to help 3 recipients), and your choice of helping will be observed by the 5 observers; if you choose not to help, you do not need to pay the 60 yuan, and your choice of not helping will also be observed by

the 5 observers.

T3 (Y-scenario): If you choose to help, you will pay 40 yuan (to help 2 recipients), and your choice of helping will be observed by 2 observers; if you choose not to help, you do not need to pay the 40 yuan, and your choice of not helping will also be observed by these 2 observers.

In the parallel world T4: There are 4 recipients in the X-scenario and 1 recipient in the Y-scenario.

T4 (X-scenario): If you choose to help, you will pay 80 yuan (to help 4 recipients), and your choice of helping will be observed by the 5 observers; if you choose not to help, you do not need to pay the 80 yuan, and your choice of not helping will also be observed by the 5 observers.

T4 (Y-scenario): If you choose to help, you will pay 20 yuan (to help 1 recipient), and your choice of helping will be observed by 1 observer; if you choose not to help, you do not need to pay the 20 yuan, and your choice of not helping will also be observed by the 1 observer.

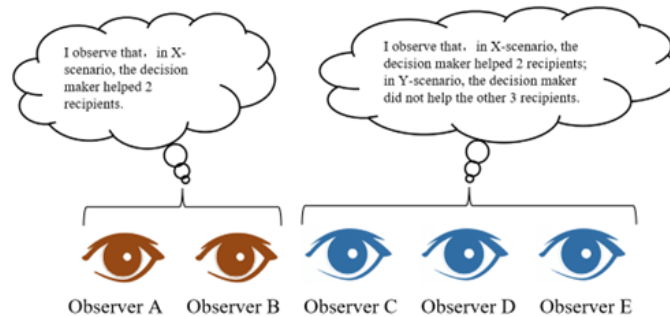
In the parallel world T5: all 5 recipients are in the X-scenario.

If you choose to help, you will pay 100 yuan (to help 5 recipients), and your choice of helping will be observed by the 5 observers; if you choose not to help, you do not need to pay the 100 yuan, and your choice of not helping will also be observed by the 5 observers.

Stage II

We will randomly select a parallel world with equal probability from T0/T1/T2/T3/T4/T5 and send your choices in this world to the 5 observers according to the rule described above. The decisions you make in the other parallel worlds will not be known to the observers.

Taking the T2 world as an example, if you choose to help in the X-scenario and not to help in the Y-scenario, the figure below shows the information observed by the 5 observers: 2 observers (brown eyes) only know that you help the two recipients in the X-scenario;



meanwhile, the other 3 observers (blue eyes) not only know that you help two recipients in X-scenario, but they also know that you do not help 3 other recipients in Y-scenario.

After the observers observe (some of) your choices, each observer will rate your generosity in a scale of 0 to 5. Note: The observers' ratings on you do not affect their own income. Their income is fixed. Each of the 5 observers makes their decisions independently. There will be no communication between them.

Your Earning

Your payment = 100 yuan (endowment) - 20 yuan \times number of recipients you help + (sum of the ratings from 5 observers) \times 3

Stage II will be completed within 10 days. After that, we will make the donation generated to the charity program. We will also pay you for the experiment via WeChat transfer and inform you of the ratings from the observers within 10 days.

The observers will not see your name or any other identification information. Your decisions, income and ratings from the observers will not be disclosed to anyone else.

An Illustration

An illustration is provided below.

Suppose your decisions are as follows:

| | T0 | T1 | T2 | T3 | T4 | T5 |
|------------|-------------------------------------|-------------------------------------|-------------------------------------|-------------------------------------|-------------------------------------|-------------------------------------|
| X-scenario | | <input checked="" type="checkbox"/> | <input checked="" type="checkbox"/> | <input checked="" type="checkbox"/> | <input checked="" type="checkbox"/> | <input checked="" type="checkbox"/> |
| Y-scenario | <input checked="" type="checkbox"/> | <input checked="" type="checkbox"/> | <input type="checkbox"/> | <input type="checkbox"/> | <input type="checkbox"/> | |

denotes “to help”, denotes “not to help”

Suppose the parallel world T3 is drawn, and your decisions in this world are to help in the X-scenario, but not in the Y-scenario.

Among the 5 observers, 3 of them only observe that you choose to help 3 recipients in the X-scenario. The remaining 2 observers not only observe that you choose to help in the X-scenario, but they can also see that you choose not to help two other recipients in the Y-scenario.

Suppose that the 3 observers who can only see your choice in the X-scenario give you ratings of U, V, and W respectively, and the 2 observers who can see your choices in both X-scenario and Y-scenario give you ratings M and N, respectively. U, V, W, M and N all lie between 0 and 5. Your total payment will be: $100 - 60 + (U+V+W+M+N) \times 3$ yuan.

A recap

You will have a donation decision to make vis-a-vis each of 5 potential recipients. In parallel world Tz, where $z \in \{0, 1, 2, 3, 4, 5\}$, your choice (the same) for z recipients in the X-scenario will be observed by all 5 observers; your choice (the same) for the 5 - z other recipients in the Y-scenario will be observed by only 5 - z observers.

Please tick your choice in each graph:

| | | |
|--|--|--|
| <p>The Parallel World: T0</p> <p>T0 (X)</p> <p>If you choose to help, you will pay 100 \$1000 (to help 2 recipients), and your choice of helping will be observed by the 5 observers.</p> <p>If you choose not to help, you do not need to pay the 100 \$1000, and your choice of not helping will also be observed by the 5 observers.</p> <p>(You)</p> <p>Observe</p> <p>(Observer A) (Observer B) (Observer C) (Observer D) (Observer E)</p> <p>Observe</p> <p>Observe</p> <p>Observe</p> <p>(Recipients in the X-Scenario)</p> <p>Observe</p> <p>Observe</p> <p>Observe</p> <p>(Recipients in the Y-Scenario)</p> | <p>The Parallel World: T1</p> <p>T1 (X)</p> <p>If you choose to help, you will pay 20 \$200 (to help 1 recipient), and your choice of helping will be observed by the 5 observers.</p> <p>If you choose not to help, you do not need to pay the 20 \$200, and your choice of not helping will also be observed by the 5 observers.</p> <p>(You)</p> <p>Observe</p> <p>(Observer A) (Observer B) (Observer C) (Observer D) (Observer E)</p> <p>Observe</p> <p>Observe</p> <p>Observe</p> <p>(Recipients in the X-Scenario)</p> <p>Observe</p> <p>Observe</p> <p>Observe</p> <p>(Recipients in the Y-Scenario)</p> | <p>The Parallel World: T2</p> <p>T2 (X)</p> <p>If you choose to help, you will pay 40 \$400 (to help 2 recipients), and your choice of helping will be observed by the 5 observers.</p> <p>If you choose not to help, you do not need to pay the 40 \$400, and your choice of not helping will also be observed by the 5 observers.</p> <p>(You)</p> <p>Observe</p> <p>Observe</p> <p>Observe</p> <p>(Observer A) (Observer B) (Observer C) (Observer D) (Observer E)</p> <p>Observe</p> <p>Observe</p> <p>Observe</p> <p>(Recipients in the X-Scenario)</p> <p>Observe</p> <p>Observe</p> <p>Observe</p> <p>(Recipients in the Y-Scenario)</p> |
| <p>The Parallel World: T3</p> <p>T3 (X)</p> <p>If you choose to help, you will pay 60 \$600 (to help 3 recipients), and your choice of helping will be observed by the 5 observers.</p> <p>If you choose not to help, you do not need to pay the 60 \$600, and your choice of not helping will also be observed by the 5 observers.</p> <p>(You)</p> <p>Observe</p> <p>(Observer A) (Observer B) (Observer C) (Observer D) (Observer E)</p> <p>Observe</p> <p>Observe</p> <p>Observe</p> <p>(Recipients in the X-Scenario)</p> <p>Observe</p> <p>Observe</p> <p>Observe</p> <p>(Recipients in the Y-Scenario)</p> | <p>The Parallel World: T4</p> <p>T4 (X)</p> <p>If you choose to help, you will pay 80 \$800 (to help 4 recipients), and your choice of helping will be observed by the 5 observers.</p> <p>If you choose not to help, you do not need to pay the 80 \$800, and your choice of not helping will also be observed by the 5 observers.</p> <p>(You)</p> <p>Observe</p> <p>(Observer A) (Observer B) (Observer C) (Observer D) (Observer E)</p> <p>Observe</p> <p>Observe</p> <p>Observe</p> <p>(Recipients in the X-Scenario)</p> <p>Observe</p> <p>Observe</p> <p>Observe</p> <p>(Recipients in the Y-Scenario)</p> | <p>The Parallel World: T5</p> <p>T5</p> <p>If you choose to help, you will pay 100 \$1000 (to help 5 recipients), and your choice of helping will be observed by the 5 observers.</p> <p>If you choose not to help, you do not need to pay the 100 \$1000, and your choice of not helping will also be observed by the 5 observers.</p> <p>(You)</p> <p>Observe</p> <p>(Observer A) (Observer B) (Observer C) (Observer D) (Observer E)</p> <p>Observe</p> <p>(Recipients in the X-Scenario)</p> <p>(Nobody is in the Y-Scenario)</p> |

We will randomly select one of the above six parallel worlds with equal probability and reveal your choices in this parallel world to the observers according to the rule described above; the observers will then decide on ratings, and your payment will be finalized. You will receive your payment via WeChat transfer together with the ratings from the anonymous observers and a copy of the electronic receipt on the donations to the charity (if any).