Eliciting Moral Preferences under Image Concerns: Theory and Experiment

Roland Bénabou Armin Falk Luca Henkel Jean Tirole

June 24, 2023

Abstract

We analyze how the impact of image motives on behavior varies with two key features of the choice mechanism: single versus multiple decisions, and certainty versus uncertainty of consequences. Using direct elicitation (DE) versus multiple-price-list (MPL) or equivalently Becker-DeGroot-Marschak (BDM) schemes as exemplars, we characterize how image-seeking inflates prosocial giving. The signaling bias (relative to true preferences) is shown to depend on the interaction between elicitation method and visibility level: it is greater under DE for low image concerns, and greater under MPL/BDM for high ones. We experimentally test the model's predictions and find the predicted crossing effect.

JEL codes: C91, D01, D62, D64, D78.

Keywords: Moral behavior, deontology, utilitarianism, consequentialism, social image, self-image, norms, preference elicitation, multiple price list, experiments.

Affiliations: Bénabou: Princeton University, NBER, CEPR, briq, IZA, BREAD, and THRED. Falk: Institute on Behavior and Inequality (briq) and University of Bonn. Henkel: University of Bonn. Tirole: Toulouse School of Economics (TSE) and Institute for Advanced Study in Toulouse (IAST).

Acknowledgments: We are grateful to Ingela Alger, Jean-François Bonnefon, Gary Charness, Franz Ostrizek, Pëllumb Reshidi, Marie-Claire Villeval and Joël van der Weele for valuable comments. Ana Luisa Dutra, Juliette Fournier, Pierre-Luc Vautrey, Ben S. Young, Youpeng Zhang and Egshiglen Batbayar provided superb research assistance.

Funding: Bénabou gratefully acknowledges financial support from the Canadian Institute for Advanced Study, Tirole and Falk from the European Research Council (European Community's Seventh Framework Program Grant Agreement no. 249429 and no. 340950, the European Union's Horizon 2020 research and innovation program, Grant Agreement no. 669217) as well as the German Research Foundation (DFG) through CRC TR 224 (Project A01).

Ethics approval: The study was approved by the ethical committees of the University of Bonn (no. 2019-01), Toulouse School of Economics and Princeton University (no. 11818).

1 Introduction

Individuals' desire to signal to others and maintain to themselves that they are generous, caring, or generally "morally good," is a powerful driver of behavior. People act more responsibly when knowing their choices will be observed and less so when given the opportunity to remain ignorant of potential harms they might cause.

The previous literature on image motives (see, e.g, Bursztyn and Jensen (2017) for an overview) has extensively documented this *level* effect on the prosociality of choices. We explore here a new channel, namely the *interaction* of image with different choice mechanisms. We focus on two key features of the latter: single versus multiple simultaneous decisions, and certainty versus uncertainty of the consequences. Both vary across charitable-contribution schemes, and they critically distinguish the two methods most commonly used to elicit preferences: direct elicitation (*DE*) and Becker-DeGroot-Marschak (*BDM*), for instance in its multiple-price list (*MPL*) format. The former features a single choice implemented with certainty, the latter multiple decisions (at different prices), of which one is randomly chosen and implemented.

Taking *DE* and *MPL* (or *BDM*) as exemplars of choice sets' interactions with signaling, we present a simple model and experiment in which agents incur a cost to do good, or forfeit a "bribe" for causing harm. The model identifies three effects that make the mechanisms *differentially* image sensitive and, when combined, generate a "crossing" pattern: when image concerns are low (but positive) *DE* will yield more contributions than *MPL*, and when they are high the ordering reverses. Relatedly, image-minded consequentialists will display Kantian-like behavior –choosing the morally right action "at any price"– much more readily under *MPL* than under *DE*.

To understand the effects at work, consider first a (DE-type) situation in which individuals may contribute to a cause (generate an externality $e > 0$) at some opportunity cost c, in time or money. In the relevant population there are two types, represented by Alex and Bob, who intrinsically value the cause at $v_H e$ and $v_L e < v_H e$. When social or self image concerns are present but not very strong, there is a range of prices $c > v_L e$ for which Bob will contribute in order to look as good as Alex, whereas for c' closer to $v_H e$ he will decline. In an *MPL/BDM* format, by contrast, the richer choice set and information thus generated make pooling more difficult, as Bob would have to state a willingness to pay of at least $v_H e$; this is too high for him, so he will decline to contribute at *any* list price $c > v_L e$. This *discouragement effect* underlies the result that *MPL/BDM* yields less giving than *DE* when image concerns are positive but relatively weak.

Working in the other direction are two effects arising from the contingent nature of *MPL/BDM* bids, which effectively lower the purchase price of image. First, the randomly drawn list price could exceed one's bid, making the latter partly *cheap talk*. This is related to random implementation, but more closely to the ability of participants in a public auc-

tion to "posture" with a high bid, while hoping that someone else will outbid them. Second is what we term the *cheap-act* effect: conditional on a bid c being binding ex-post, the average price paid is only $E[\tilde{c}|\tilde{c} \leq c]$. As image concerns intensify, Bob's desire to pool and Alex's desire to separate lead to increasingly high bids, so the cheap-talk effect weakens (implementation becomes more certain). In contrast, the cheap-act effect strengthens (for standard distributions the "discount" $c - E[c|\tilde{c} \leq \tilde{c}]$ grows), causing *MPL* contributions to rise above those under *DE*.

We test the model's predictions using an experiment in which about 700 participants face a choice between: (i) directing a 350€ donation to a charity in India that will use the money to treat five tuberculosis patients, resulting statistically in the expected saving of one human life; or (ii) taking money for themselves, where the amount is either a fixed 100€ under *DE*, or determined by the subjects' cutoff on an *MPL* where prices range from 0 to 200€. These two elicitation conditions are crossed with low and high moral-image treatments. Comparing the fractions of subjects choosing the "saving a life" contribution over taking 100€, we find a sizeable reversal between *DE* and *MPL* as image concerns go from weak to strong, as predicted by the theory. In the *Low Image* treatments, the fraction opting to save a life is 48% under MPL versus 59% under DE, while in the *High Image* condition it is 63% under DE versus 72% under *MPL*.[1] On the cautionary side, statistical significance is only at the 6-7 percent level, so our simple experiment should be seen as proof-of-concept for the mechanisms brought to light by the model, opening them up to more systematic exploration.

1.1 Related Literature

Previous research on social and self image has primarily focused on how they spur prosocial behaviors, and how this signaling incentive is affected by the presence of rewards (Bénabou and Tirole, 2006, 2011a,b; Ariely et al., 2009; Ashraf et al., 2014; Grossman and van der Weele, 2017; Falk, 2021) or excuses (Dana et al., 2007; Exley, 2016; DellaVigna et al., 2012). Our analysis highlights instead their interaction with the mechanism through which choices are made. Not only are schemes such as *DE* vs *MPL/BDM* differentially sensitive to image concerns, but their effectiveness at measuring intrinsic preferences, or on the contrary spurring higher contributions, can even reverse as reputational motives intensify.

Another strand of work focuses on decision makers' probability of being pivotal (Feddersen et al., 2009; Grossman, 2015; Falk et al., 2020; Bartling et al., 2022). We show how, in mechanisms such as *MPL*, the probability of having one's choice implemented varies systematically with the intensity of image concerns, as does the expected cost at which the choice will be implemented, and we analyze how both effects shape equilibrium behavior. This re-

[1] We also conduct a placebo experiment with 366 additional subjects, keeping all aspects unchanged except that choices are now over a non-moral good, for which no image concerns arise. As expected, we find no significant difference between the two elicitation methods.

lates the paper to work on auctions with signaling, in which bidders seek to demonstrate goodness, wealth, or a strong aftermarket position (Goeree, 2003; Giovannoni and Makris, 2014; Bos and Pollrich, 2020; Bos and Truyts, 2022). In our setting, an agents' distribution of potential outcomes depends only on his own choices, and this lower strategic complexity allows us to identify intuitive effects and testable predictions.

With respect to experimental methodology, we contribute to the study of alternative elicitation mechanisms. Substantial research has compared how *DE*, *BDM*, *MPL* or random implementation (Selten, 1967) affects behavior in one-shot, anonymous games such as dictator or public-goods (Brandts and Charness, 2011; Chen and Schonger, 2016).[2] There is also a large body of research on elicitation methods for risk, time and ambiguity preferences (Charness et al., 2013; Cox et al., 2015; Cohen et al., 2020; Baillon et al., 2022). To our knowledge, no such study has explored reputationally sensitive decisions like those analyzed here. For choices in the moral domain, self-image (at least) is almost inevitably at play, and can create differences between elicitation methods.[3]

Finally, the paper relates to the debate between consequentialist and deontological principles. The evidence on how people behave in practice is mixed: the literature on public-goods contributions and charitable giving finds that choices are generally sensitive to the implied consequences (Ledyard, 1995; Goeree et al., 2002), including the risk of having no impact (Brock et al., 2013) and overhead costs (Gneezy et al., 2014). At the same time, there is evidence of "warm glow" altruism, in which utility is derived from the act as such (Andreoni, 1989, 1990). Experiments that directly focus on consequentialist versus deontological or expressive choices (Van Leeuwen and Alger, 2021; Chen and Schonger, 2022; Falk et al., 2020; Bénabou et al., 2022) also suggest a mix of preferences. Our paper shows that, when image concerns are important, a mechanism like *MPL* or *BDM* can easily lead consequentialist agents to adopt deontological-looking behaviors.

2 Model

2.1 Preferences

Agents are risk-neutral, with a two-period horizon, $t = 1, 2$. At date 1, an individual can engage in prosocial behavior $(a = 1)$ or act selfishly $(a = 0)$. Choosing $a = 1$ involves a personal cost $c > 0$ but generates a public good or externality $e \geq 0$. Agents differ in their intrinsic motivation to act morally: given e, it is either $v_H e$ (high type) or $v_L e$ (low

[2]Concerning DE with deterministic versus random implementation (an intermediate case relative to MPL), the overview by Charness et al. (2016) reports generally ambiguous effects. As the model will make clear, it is only in the presence of sufficient signaling concerns that probabilistic implementation will matter. In contrast, risk attitudes play no role in the effects that we identify, which directly affect expected returns.

[3]In the non-moral domain, in contrast, the literature tends to find no difference between *DE* and *BDM* (Miller et al., 2011; Berry et al., 2020; Cole et al., 2020).

type), with probabilities ρ and $1-\rho$, $v_H > v_L \geq 0$, and average $\bar{v} = \rho v_H + (1-\rho)v_L$.[4] Besides the externality, the second feature of action $a = 1$ tying it to the moral domain is that it can be reputationally valuable, conferring a social or self-image benefit at date 2. In the social context, the agent knows his type but the audience (peer group, firms, potential partners) does not. In the self-signaling context, he has an immediate, "intuitive" sense of his deep preferences at the moment of action – for instance, how much empathy or spite he experiences – but later on the intensity of that feeling is imperfectly accessible ("forgotten"), and only the deed itself, $a = 0$ or 1, can be reliably recalled to assess his own moral identity.

Under either interpretation, an agent of type $v = v_H, v_L$ has expected utility

$$(ve - c)\, a + \mu \hat{v}(a), \tag{1}$$

where $\hat{v}(a)$ is the expected type conditional on the action $a \in \{0, 1\}$ and the circumstances under which it took place (deterministic cost, random draw from a list, etc.), while μ is the strength of self or social-image concerns, common to all agents. This utility may be additively augmented by any externalities generated by others, but since that term is independent of the agent's action we omit it here. Note that these preferences are consequentialist: an agent's desire to behave prosocially trades off the externality he expects his actions to have, the personal costs involved, and the reputational consequences.

As common in signaling models, multiple equilibria may coexist: when

$$\max\{v_L e - c + \mu(v_H - v_L),\ v_H e - c + \mu(v_H - \bar{v})\} \leq 0 \leq v_H e - c + \mu(v_H - v_L),$$

there is both a pooling equilibrium at $a = 0$ and a separating one in which the v_H type contributes, with a mixed-strategy one in-between; see the Appendix, which gathers all the paper's proofs. In case of multiplicity we choose the equilibrium that is best for both types, namely the no-contribution pooling equilibrium. Indeed, separation yields lower payoffs for both, since $\mu v_L < \mu \bar{v}$ and $v_H e - c + \mu v_H \leq \mu \bar{v}$.

This simple framework readily implies that an agent is more likely to act morally the higher the externality e, his preference $v \in \{v_H, v_L\}$, and/or his image concern μ.

2.2 Direct Elicitation

Under *DE,* the individual faces a take-it-or-leave-it opportunity to incur a given cost (or forfeit a given prize) c to create an external benefit e. As illustrated in Panel A of Figure 1 (for $\rho < 1/2$), equilibrium behavior is characterized by three cost thresholds, increasing in the reputational concern μ, that delineate regions of separation, semi-separation, and

[4]The Appendix discusses how the paper's mechanisms and results translate in richer type spaces.

pooling:

$$v_H e - c_H^{DE}(\mu) + \mu\,(v_H - \bar{v}) \equiv 0, \tag{2}$$

$$v_L e - \bar{c}_L^{DE}(\mu) + \mu\,(v_H - v_L) \equiv 0, \tag{3}$$

$$v_L e - \underline{c}_L^{DE}(\mu) + \mu\,(\bar{v} - v_L) \equiv 0. \tag{4}$$

Denoting $a_H^{DE}(c,\mu)$ and $a_L^{DE}(c,\mu)$, or a_H and a_L for short, the two types' probabilities of choosing $a = 1$, we show

Proposition 1. *The outcome of direct elicitation is as follows:*

1. *For low costs,* $c < \min\{\underline{c}_L^{DE}, c_H^{DE}\}$, *everyone behaves morally,* $a_H = a_L = 1$.
2. *For intermediate costs,* $c \in (\underline{c}_L^{DE}, c_H^{DE})$, *the high type behaves morally* $(a_H = 1)$, *but the low type's probability* $a_L(c)$ *of doing so decreases with* c, *and then equals* 0 *for* $c \geq \min\{\bar{c}_L^{DE}, c_H^{DE}\}$.
3. *For high costs,* $c \geq c_H^{DE}$, *both types behave immorally,* $a_H = a_L = 0$.

Relative to "pure" (intrinsic) moral preferences ve, decision thresholds are inflated due to reputational concerns; see (2)-(4). In particular, the range of costs $[\bar{c}_L^{DE}, c_H^{DE}]$ where full separation occurs shrinks with μ, becoming empty for $\mu > e/\rho$.

2.3 Multiple-Price List

Under *BDM*, the individual "names his price" by stating what maximum cost $c \in [0, c_{\max}]$ he is willing to incur for taking action $a = 1$, where $0 \leq v_L e < v_H e < c_{\max}$. Equivalently, c represents his willingness to accept a "bribe" to make the immoral choice, $a = 0$. This elicitation is made incentive-compatible by drawing some $\tilde{c} \in [0, c_{\max}]$ according to a pre-announced distribution $G(\tilde{c})$, and implementing $a = 1$ at cost $\tilde{c}$ only when $\tilde{c} \leq c$. With *MPL*, the price range is discretized and subjects state contingent choices at each level. Both schemes generate identical incentives, so we gather them under the label of *MPL*, since that is the format we implement experimentally.

In experiments, G is typically uniform, but we allow any other case, including $c_{\max} = +\infty$. Let $L(c)$ denote the low type's net loss from selecting a cutoff $c \geq v_L e$:

$$L(c) \equiv \int_{v_L e}^{c} (\tilde{c} - v_L e)\, dG(\tilde{c}) = \underbrace{\mathbb{P}(\tilde{c} \in [v_L e, c])}_{\text{cheap-talk effect}} \underbrace{(\mathbb{E}(\tilde{c} | \tilde{c} \in [v_L e, c]) - v_L e)}_{\text{cheap-act effect}} \tag{5}$$

and assume $L(c_{\max}) < \infty$, for which it suffices that $E_G[\tilde{c}] < \infty$. We will say that a subject is *observationally deontological* if he turns down all prices on the proposed list (with distribution G): given the available data, he behaves as someone who would not act immorally "at any price."

We now solve for both types' willingness to accept (WTA) under the multiple-price list, denoted c_H^{MPL} and c_L^{MPL} respectively. Note first that, *absent* reputation concerns ($\mu = 0$), MPL and DE are equivalent, and reveal true preferences: $c_H^{DE} = c_H^{MPL} = v_H e$, $\underline{c}_L^{DE} = \bar{c}_L^{DE} = c_L^{MPL} = v_L e$. For $\mu > 0$, comparing $L(c)$ to the reputational stakes $\mu(v_H - v_L)$ and $\mu(v_H - \bar{v})$ yields both types' equilibrium strategies, illustrated in Panel B of Figure 1 and characterized again by critical thresholds between separating, semi-separating and pooling regions:

$$\underline{\mu} \equiv \frac{L(v_H e)}{v_H - v_L} < \mu^* \equiv \frac{L(c_{\max})}{v_H - v_L} < \frac{L(c_{\max})}{\rho(v_H - v_L)} \equiv \bar{\mu}. \tag{6}$$

Proposition 2. *The outcome of the MPL mechanism is as follows:*

1. *When the (self) reputational concern μ is low, $\mu < \mu^*$, the high type's WTA for behaving immorally is $c_H^{MPL} = \max\{v_H e,\ L^{-1}(\mu(v_H - v_L))\}$, while the low type finds it too costly to pool and accepts $c_L^{MPL} = v_L e$.*

   Initially, for $\mu \leq \underline{\mu}$, separation is costless for the high type, then as μ rises he has to raise his reservation price to separate from the low type.

2. *When μ is intermediate, $\mu \in [\mu^*, \bar{\mu}]$, the high type can no longer separate and becomes observationally deontological, $c_H^{MPL} = c_{\max}$. The low type randomizes, with probability $a_L(\mu)$ increasing in μ, between that same "virtuousness" ($c_L^{MPL} = c_{\max}$) and revealing himself (accepting $c_L^{MPL} = v_L e$).*

3. *When $\mu > \bar{\mu}$, (self) image concerns are strong enough that both types' behavior is observationally deontological: $c_H^{MPL} = c_L^{MPL} = c_{\max}$.*

2.4 Comparison of DE vs. MPL

Under both elicitation schemes, image concerns naturally raise contributions, as seen in Figure 1. More novel are the following questions:

1. Is one elicitation scheme *more image-sensitive* than the other?
2. Which one yields *more expected contributions?*

Formally, at a given cost $c \in [0, c_{\max}]$, what fraction of people $\bar{a}^{DE}(c, \mu)$ accept forfeiting c to implement $a = 1$ under *DE,* versus what fraction $\bar{a}^{MPL}(c, \mu)$ state a willingness to pay of at least c under *MPL*? And how does $\bar{a}^{DE}(c, \mu) - \bar{a}^{MPL}(c, \mu)$ depend on μ?

While the answers generally depend on the specific value of c, the cases of sufficiently low and high image concerns yield clear predictions. We will denote as μ^{**} the solution to $\underline{c}_L^{DE}(\mu) = c_{\max}$, or

$$\mu^{**} \equiv \frac{c_{\max} - v_L e}{\bar{v} - v_L} > \frac{L(c_{\max})}{\bar{v} - v_L} = \bar{\mu}. \tag{7}$$

Putting together Propositions 1 and 2, we have:

Proposition 3. *For each type $\tau = H, L$,*

1. *Visibility raises contributions: for any $c \in [0, c_{\max}]$, $a_\tau^{DE}(c,\mu)$ and $a_\tau^{MPL}(c,\mu)$ coincide at $\mu = 0$, then both increase (weakly) as μ rises, reaching 1 for μ large enough.*

2. *Under low image concerns, DE yields more contributions: for all $\mu \in (0, \underline{\mu})$, $a_\tau^{DE}(c,\mu) \geq a_\tau^{MPL}(c,\mu)$, with strict inequality for $c \in (v_L e, \bar{c}_L^{DE}(\mu))$ and $c \in (v_H e, c_H^{DE}(\mu))$, both nonempty.*

3. *Under high image concerns, MPL yields more contributions: for all $\mu \geq \bar{\mu}$, $a_\tau^{DE}(c,\mu) \leq a_\tau^{MPL}(c,\mu) = 1$, with strict inequality for $\tau = L$ and $c \in (\underline{c}_L^{DE}(\mu), c_{\max})$, which is nonempty whenever $\mu \in (\bar{\mu}, \mu^{**})$.*

4. *The average behavior over types, $\bar{a}^m(c,\mu) \equiv \rho a_H^m(c,\mu) + (1-\rho) a_L^m(c,\mu)$, $m = DE, MPL$, inherits these same properties.*

The first result is standard, while the others stem from the interplay of three effects.

Weak image concerns: discouragement effect dominates. When $\mu > 0$ is low enough that separation under *MPL* is costless, we have $c_H^{MPL}(\mu) = v_H e < c_H^{DE}(\mu)$ and $c_L^{MPL}(\mu) = v_L e < \underline{c}_L^{DE}(\mu)$, hence the second result. Intuitively, *MPL* raises the cost to the low type of mimicking the high one, since to do so he must forego up to $v_H e$, and for low reputational gain such a discrete cost is not worth it. Under *DE*, in contrast, he pays only in proportion to the gain. This intuition is reflected in the fact that the lower boundary of the separating region is linear in Panel A of Figure 1, whereas it is initially flat in Panel B.

Strong image concerns: cheap-act effect dominates. At high values of μ, reputational concerns become paramount, and the cost of signaling is lower under *MPL* than under *DE,* since high values of c must only be paid with a probability less than 1: the effective cost of stating a cutoff c is only $E[\tilde{c}|\tilde{c} \leq c] < c$. It is even bounded by $L(c_{\max}) + v_L e < \infty$, which limits the extent to which the high type can separate, so that for $\mu > \bar{\mu}$ full pooling occurs: $c_H^{MPL} = c_L^{MPL} = c_{\max}$, so $a^{MPL}(c,\mu) = 1$, whereas $\bar{a}_L^{DE}(c,\mu) < 1$ as long as $\mu < \mu^{**}$. Most importantly:

Property 1. *For any distribution satisfying the monotone hazard rate property $(g/(1-G)$ increasing), the "discount" $c - E[\tilde{c}|\tilde{c} \leq c]$ is increasing in c. Therefore, as μ rises and with it each type's cutoff, the* cheap-act *effect becomes stronger, which increases MPL contributions relative to DE.*

Intermediate image concerns. Inside $(\underline{\mu}, \bar{\mu})$, a third *"cheap-talk"* effect is also important. Under *MPL,* an agent who states a cutoff $c < c_{\max}$ has only a probability $G(c) < 1$ of being called upon to actually "deliver": if $\tilde{c} > c$ is drawn, he neither incurs a cost *nor* generates the externality e. This makes it safer to state high cutoffs, thus adding to the *cheap-act* effect.

The latter is not as strong in this range as for high values of μ, and conversely the *cheap-talk* effect weakens as μ rises, pushing $G(c^{MPL})$ closer to 1. The net balance of the three effects is generally ambiguous in this intermediate range, and consequently so is the sign of $a^{DE} - a^{MPL}$.

Implications. Three main predictions emerge from the model. First, as usual, greater visibility increases contributions. Second, at low but positive levels of visibility, *DE* leads to more prosocial outcomes, as the *discouragement effect* dominates. Third, at high levels (but not so high as to push everyone to $a = 1$ under *DE*), this ordering reverses: *MPL* induces more moral decisions, due to the now dominating *cheap-act* effect.

The inequalities in Proposition 3 can be weak or strong, depending on the region of the parameter space. This is a standard feature of models with discrete types and action spaces, which typically disappears when there is sufficient heterogeneity to span all cases. For this reason, when confronting the model with data, we will tighten the predicted inequalities to be strict ones.

3 Experimental Design

3.1 Saving a Life

We adopt the *Saving a Life* paradigm from Falk and Graeber (2020), in which subjects can either take money for themselves or implement a fixed, life-saving donation to a charity dedicated to the treatment of tuberculosis in India. According to the World Health Organization, tuberculosis is one of the ten leading causes of death worldwide, even though there are highly effective antibiotic treatments available. Together with the Indian non-profit organization *Operation ASHA*, we calculated a specific monetary amount sufficient to identify, treat, and cure a number of patients such that – in expectation – one patient will be saved from death by tuberculosis due to the donation. Combining public information on the charity's operations with estimates from peer-reviewed studies on mortality due to tuberculosis and treatment effectiveness for the specific location considered (Straetemans et al., 2011; Tiemersma et al., 2011; Kolappan et al., 2008), we determined that level to be 350€: by allowing for the treatment of five patients, such a donation allows the (expected) saving of one human life.

This paradigm contrasts the option of saving a life (major positive externality e) by triggering a donation of 350€ versus that of taking money for oneself (opportunity cost c), inducing a clear tradeoff between morality and self-interest.

3.2 Treatments

We use a 2×2 between-subjects design, varying the elicitation method (DE vs. MPL) as well as the visibility and moral salience of choices (*Low Image* vs. *High Image*) at the payment stage.

Under DE, subjects faced the binary choice between receiving $c = 100$€ (≈$110) as payment, or saving a human life in expectation. As part of the experimental design, we predetermined this single value of $c = 100$€ as a compromise between two practical concerns: (i) c must be high enough to generate choices of both types; (ii) in contrast to *MPL*, each implemented decision has a sure cost to the experimental budget of either c or the full 350€ donation, which quickly adds up.

For the MPL conditions, we used a price-list design: starting with $c = 0$€ and proceeding in 10€ increments up to $c = 200$€, subjects could indicate in each of the 21 contingent choices whether they wanted to save a life or take c for themselves. Each price was then equally likely to be drawn for implementation (uniform G).[5] Figures B.1 and B.2 in the Online Appendix B display the corresponding decision screens.

Turning to visibility, recall that the two key forces underlying Proposition 3, namely the *discouragement* and the *cheap-act* effects, both require a non-zero level of image concerns. To ensure a minimal level of image concern in both treatments, we notified subjects at the start that: (i) they were anonymously paired with another participant in the same session; (ii) they would see, at the end of the experiment, their own and their partner's choices displayed alongside on their screens, as would their partner. Apart from observing the partner's choices, subjects received no information about them, so that no other aspect of the dyad would influence decisions.

To keep image concerns minimal in the *Low Image* treatment ($\mu = \mu_L$), we made the payment procedure double-blind, so that not even the experimenter could link subjects' decisions to their identity. Following Barmettler et al. (2012), at the start of each session one subject was randomly designated to carry out all payments: they did not participate in the regular experiment, and thus had no knowledge about the choices. At the end, payments were stuffed into envelopes and the selected subject handed them out, in an adjacent room, to those who had actively participated.

The *High Image* treatment ($\mu = \mu_H$), in contrast, was designed to induce strong image concerns. Subjects were informed that upon receiving payment: (i) their choice would be compared to that of their matched partner by a committee of three persons, sitting in the room where payments would take place; (ii) both partners' choices would be projected onto a wall, and they would have to read both aloud, using two predetermined sentences.

[5] To avoid inconsistent answers, we enforced a single-switching rule.

3.3 Procedure

697 subjects (405 female, mean age = 24.01, SD = 6.21) participated in 36 sessions at the BonnEconLab of the University of Bonn: 178 in the *MPL-Low Image* treatment, 178 in *MPL-High Image*, 165 in *DE-Low Image*, and 176 in *DE-High Image*. Subjects were recruited using Hroot (Bock et al., 2014), and the experiment was conducted using oTree (Chen et al., 2016). Sessions lasted about 60 minutes, with a show-up fee of 12€. For each session, one matched pair of subjects was randomly drawn, and their choices implemented. Thus, in the DE treatments, each of the two either received 100€, or triggered a life-saving 350€ donation. In the *MPL* treatments, one price from the list was randomly drawn (uniformly), and the pre-stated choices of both partners for this price were implemented. Therefore, each one either triggered the donation or received up to 200€.[6]

At the beginning of each session, subjects received a verbal introduction to the experiment. In the *Low Image* treatments, the procedure ensuring anonymity was explained and demonstrated. In the *High Image* treatments, the committee setup was shown. Subsequently, all subjects received detailed information about tuberculosis, its effects, and treatment. The instructions also linked to a website where they were invited to confirm the validity of the information. We then introduced the charity and its working procedure, and explained our calculations regarding the life-saving effect of the 350€ donation. Subjects then learned about their choice options and, after answering a couple of comprehension questions, made their decisions. Finally, they completed a short questionnaire and were paid in a separate room, with payment procedures depending on treatment status, as explained above. For further details on the procedure and instruction, see Online Appendix D.

4 Hypotheses and Results

Our outcome variable is the fraction $\bar{a}^m(c,\mu)$ of subjects who choose to save a life over receiving c, given an elicitation method $m \in \{DE,\ MPL\}$ and a level of visibility $\mu \in \{\mu_L, \mu_H\}$. For brevity, we will refer to $\bar{a}^m(c,\mu)$ as "total contributions".

4.1 Hypotheses

Based on Proposition 3, we state:

Hypothesis 1. *For both* DE *and* MPL, *total contributions are higher under* High Image *than under* Low Image: $\bar{a}^{DE}(c,\mu_H) > \bar{a}^{DE}(c,\mu_L)$, $\bar{a}^{MPL}(c,\mu_H) > \bar{a}^{MPL}(c,\mu_L)$.

[6]This random implementation adds another layer of the cheap-talk effect, but one that affects *DE* and *MPL* in exactly the same way (formally equivalent to dividing μ by the probability of implementation), and thus leaves all comparisons between the two unaffected.

Hypothesis 2. *Under* Low Image*, total contributions are higher under* DE *than under* MPL*:* $\bar{a}^{DE}(c, \mu_L) > \bar{a}^{MPL}(c, \mu_L)$.

Hypothesis 3. *Under* High Image*, total contributions are higher under* MPL *than under* DE*:* $\bar{a}^{DE}(c, \mu_H) < \bar{a}^{MPL}(c, \mu_H)$.

Hypothesis 1 captures the standard effect of signaling concerns. The novel ones are Hypotheses 2 and 3**,** reflecting the dominance of the *discouragement effect* at μ_L and the *cheap-act effect* at μ_H. Together, they constitute the model's distinctive crossing prediction, which we will test at $c = 100$€, as explained earlier.

4.2 Results

Hypothesis 1. Under both elicitation methods, increased visibility led to a rise in total contributions, but the magnitude was markedly different. Under *DE*, 58.8% of subjects chose to save a life in *Low Image* and 62.5% in *High Image* – a relatively small and insignificant increase ($p = 0.51$, Fisher's exact test). Under *MPL*, increased visibility had a much larger effect. At almost all payment levels, the fraction of subjects choosing to save a life is at least 15 pp. higher under *MPL-High Image* than under *MPL-Low Image*, resulting in significantly different distributions ($p < 0.001$, Kolmogorov–Smirnov test); see Panel A of Figure 2. At 100 €, contributions are 23.6 pp. and significantly higher under *High Image* than under *Low Image* ($p < 0.001$).

Hypotheses 2 and 3. Panel B of Figure 2 shows that the fractions $\bar{a}^m(100, \mu)$ choosing to save a life over 100€ clearly differ by elicitation method, with the ranking reversing between μ_L and high μ_H. Under *Low Image*, we observe $\bar{a}^{MPL}(\mu_L) < \bar{a}^{DE}(\mu_L)$, as predicted by Hypothesis 2, and consistent with the dominance of the *discouragement effect*. The difference is large, with the fraction saving a life rising from 48.3% to 58.8% between *MPL* and *DE*, though significance is slightly below the conventional level ($p = 0.065$, Fisher's exact test). Conversely, under *High Image* we observe $\bar{a}^{MPL}(\mu_H) > \bar{a}^{DE}(\mu_H)$, in line with the *cheap-act effect* dominating, as predicted by Hypothesis 3. The difference is again about 10 percentage points, but now in the opposite direction, rising from 62.5% under DE to 71.9% under MPL, albeit again with significance slightly short of 5% ($p = 0.070$).

Table 1, Panel A regresses the probability of choosing to save a life (instead of taking 100€) on a dummy for the type of elicitation (1 for *MPL*), which yields a positive coefficient for *Low Image* in Column (1), and a negative one for *High Image* in Column (3).[7] Columns (2) and (4) show that these effects remain largely unaffected by controls for age, gender, high-school graduation grade, highest educational degree obtained so far, self-reported monthly income, and a measure of religiousness (Likert scale).

[7]The results remain qualitatively unchanged with Probit or Logit regressions.

Hypotheses 2-3 represent the strictest possible test of the model – a particular ordering of four variables– which may explain the marginal significance of those results. A more standard test concerns their joint implication of a *differential image sensitivity*: as image rises from μ_L to μ_H, the increase in contributions should be more pronounced for *MPL* than for *DE*. Panel B of Table 1 thus presents an OLS regression interacting *High Image* with *MPL,* using *DE-Low Image* as baseline; the interaction is positive and significant at the 1-percent level.

Robustness Experiment. One may worry that features of the elicitation methods unrelated to image concerns might be at play in our results. Note first that these would have to generate not just different *DE* versus *MPL* contributions, but also a flipping of that gap as image rises from low to high, which seems unlikely. Nonetheless, to rule out potential confounding factors we ran the *DE* versus *MPL* treatments on another 366 subjects, with the donation replaced by a non-moral good (university-shop voucher). For this "placebo,"$\mu = 0$, and indeed we find no significant differences between *MPL* and *DE*: see Panel C of Table 1, and Online Appendix C for implementation details.

5 Conclusion

Our model and experiment show that image concerns affect the measurement of moral preferences in ways that *interact with the elicitation method.* Regardless of whether one is interested in image-inclusive preferences (for positive predictions) or in purely intrinsic ones (for normative judgements), behavior will differ between direct and price-list mechanisms. These results argue for caution in interpreting standard estimates of moral preferences from experiments and contingent-valuation surveys,[8] but also provide potential guidance for maximizing public-goods contributions and image manipulations.[9]

In particular, even purely utilitarian individuals may act, when facing *BDM*- or *MPL*-like situations, as if deontologically motivated: refusing all proposed prices in exchange for what is perceived as having a dignity. With necessarily finite budgets, a definitive test of how many "real Kantians" there are is ultimately impossible, but our experiment provides both an upper bound and some grounds for skepticism about public positions on the subject. The former is given by the 26.4% of subjects who choose to save a life over the maximum offer of $200€$ in the Low Image *MPL* condition. The latter stems from the fact that this proportion nearly doubles to 43.82% with a mild visibility manipulation. These results can also help to account

[8]A related point is made by Chen and Schonger (2022) for other forms of preferences involving moral "duties".

[9]Individual WTP's, which include the value of social and self-image, are the right measures to predict, explain or alter behavior. To inform policy, however, they can substantially overstate the true social value of the public good. Thus, in our model, reputation is a positional good, the image gains and losses of contributors and non-contributors exactly offsetting each other. In general, the image game can have negative, zero, or positive sum, depending on the curvature of the reputation functional; Butera et al. (2022) find evidence for negative sum, which reinforces the previous point.

for the common resistance to estimating and using a "statistical value of life." Despite the fact that we implicitly engage in trading off costs and statistical lives all the time, explicit reference to putting a price tag on life typically produces conspicuously displayed righteous indignation (e.g., Sandel, 2012).

On the empirical side, an interesting avenue for further research would be to estimate the distributions of intrinsic preferences and image concerns in a population, from those of *MPL* bids for the desired outcome (as in the work on auctions) and for making one's choices visible (as in Butera et al., 2022).

6 Appendix

Proof of Proposition 1. From (2)-(4), it follows that:

$(P_0) : a_H = a_L = 0$, sustained by out-of equilibrium belief (OEB) $\hat{v} = v_H$ following $a = 1$ (by the D1 criterion), is an equilibrium if and only if $c \geq c_H^{DE}$. When

$$\bar{c}_L^{DE} = v_L e + \mu(v_H - v_L) \leq c \leq v_H e + \mu(v_H - v_L) \equiv \bar{c}_H^{DE},$$

it coexists with a separating equilibrium S in which $a_H = 1 = 1 - a_L$, plus a mixed-strategy one in-between. A shown earlier, however, P_0 is Pareto dominant, and therefore selected.

$(P_1) : a_H = a_L = 1$, sustained by OEB $\hat{v} = v_L$ following $a = 0$ (by D1), is an equilibrium if and only if $c \leq \underline{c}_L^{DE}$.

$(S) : a_H = 1 - a_L = 1$ is an equilibrium if and only if $\bar{c}_L^{DE} \leq c \leq \bar{c}_H^{DE}$.

$(SS_1) : 0 < a_L < 1 = a_H$, with belief $\hat{v} \in (v_L, \bar{v})$ following $a = 1$, is an equilibrium if and only if $\underline{c}_L^{DE} < c < \bar{c}_L^{DE}$. The low type's mixed strategy $a_L(c) \in (0, 1)$ is then given by combining the indifference condition $v_L e - c + \mu(\hat{v}(a_L) - v_L) = 0$ and the Bayesian posterior $\hat{v}(c) = [\rho v_H + (1 - \rho) a_L v_L] / [\rho v + (1 - \rho) a_L]$:

$$v_L e - c + \frac{\mu\rho(v_H - v_L)}{\rho + (1 - \rho) a_L(c)} \equiv 0, \tag{8}$$

so $a_L(c)$ decreases with c, while the reputation $\hat{v}(c)$ following $a = 1$ increases.

$(SS_0) : 0 = a_L < a_H < 1$, with beliefs $\hat{v} \in (\bar{v}, v_H)$ following $a = 0$, is an equilibrium if and only if $c_H^{DE} < c < \bar{c}_H^{DE}$. It always coexists with P_0, and is always dominated by it.

These results jointly imply that:

(a) If $\underline{c}_L^{DE} < \bar{c}_L^{DE} < c_H^{DE}$, the unique equilibrium is P_1 for $c < \underline{c}_L^{DE}$; SS_1 for $c \in [\underline{c}_L^{DE}, \bar{c}_L^{DE}]$; and S for $c \in [\bar{c}_L^{DE} < c_H^{DE}]$. For $c \geq c_H^{DE}$, the dominant equilibrium is P_0.

(b) If $\underline{c}_L^{DE} < c_H^{DE} < \bar{c}_L^{DE}$, the unique equilibrium is P_1 for $c < \underline{c}_L^{DE}$, and SS_1 for $c \in [\underline{c}_L^{DE}, c_H^{DE}]$. For $c > c_H^{DE}$, the dominant equilibrium is P_0.

(b) If $c_H^{DE} < \underline{c}_L^{DE} < \bar{c}_L^{DE}$, the unique equilibrium is P_1 for $c < c_H^{DE}$, and for $c \geq c_H^{DE}$ the dominant equilibrium is P_0. ■

Proof of Proposition 2. The proof of existence is standard. For example, for a separating equilibrium to obtain, it must be: that (i) type v_L obtains his symmetric-information allocation (otherwise, he would be better off selecting $c_L^{MPL} = v_L e$), and (ii) he does not want to mimic type v_H: $\mu(v_H - v_L) \leq L(c_H^{MPL})$ and $c_H^{MP} < c_{max}$. It is easily verified that the proposed strategies satisfy these conditions, and similarly for the semi-separating and pooling equilibria .

The equilibrium is not unique absent refinement, however. For example, there is a pooling equilibrium at $c^{MPL} = v_H e < c_{max}$ when $\mu(\bar{v} - v_L) \geq L(v_H e)$, sustained by OBE $\hat{v} = v_L$ following any declared price $c \neq v_L e$. Note, however, that sorting implies monotonicity, so there is at most one price, denoted c^*, that can be chosen with positive probability by both types; any other price claimed by type v_H (respectively, v_L) exceeds c^* (respectively, lies below it) c^*) . Denote $\hat{v}(c)$ the mean belief following a price c, and consider a deviation to $c' = c^* + \varepsilon$, for $\varepsilon > 0$ arbitrarily small, together with the set of belief responses that raise both types' utilities relative to equilibrium

$$\hat{V}_L \equiv \{\hat{v}(c^* + \varepsilon) \mid \mu\left[\hat{v}(c^* + \varepsilon) - \hat{v}(c^*)\right] > L_L(c^* + \varepsilon) - L_L(c^* + \varepsilon)\},$$
$$\hat{V}_H \equiv \{\hat{v}(c^* + \varepsilon) \mid \mu\left[\hat{v}(c^* + \varepsilon) - \hat{v}(c^*)\right] > L_H(c^* + \varepsilon) - L_H(c^* + \varepsilon)\}.$$

Clearly $V_L \subset V_H$, so by D1 the deviation must induce a probability-one belief on v_H; thus, the only possible pooling price is $c = c_{max}$. Consequently, the equilibrium must take one of the three forms described in the proposition, and because it is obtained on disjoint sets of parameters, it is unique under D1. ■

Richer type spaces. Our two-type model brings to light three channels through which image and choice mechanisms interact. With more types they still operate, though less can be said about their net balance when comparing *DE* and *BDM*. The cheap-talk and cheap-act effects arising under *MPL*, one attenuating and the other strengthening with image concerns, are both very general, extending even to a continuum: equilibrium bids naturally rise with μ, which increases the implementation probability and reduces the effective price of image; see (5). For the discouragement effect, with $n > 2$ types it remains the case that, for μ positive but low enough, *MPL*'s richer information hinders pooling. With a continuum, however, separation can no longer be costless, for any reputational stakes. Overall, with a distribution $F(v)$ over $[0, v_{max}]$ (see Online Appendix A for details):

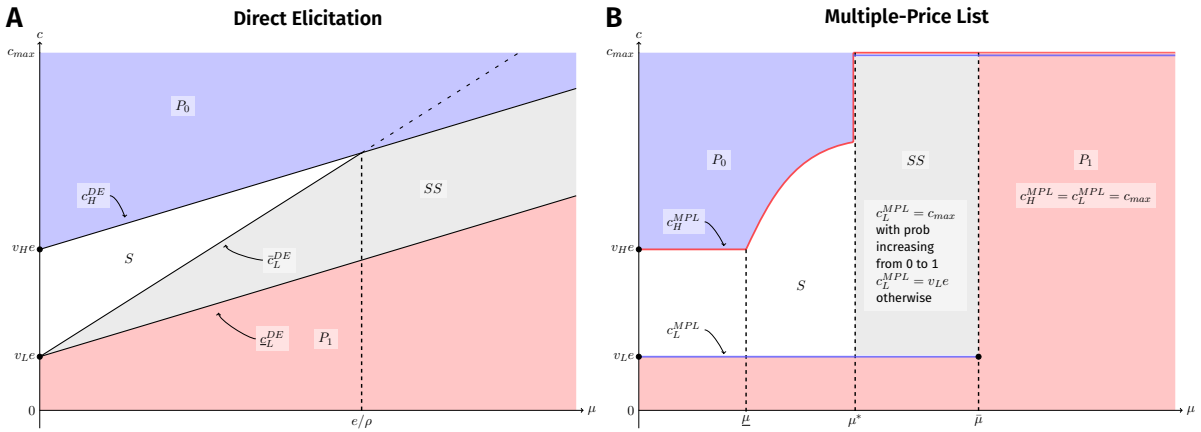
1. The characterization of *DE* (Proposition 1) carries over, with type v now contributing at c if $b^{DE}(v) \equiv v + \mu(E[v'|v' > v] - E[v'|v' < v]) > c$, defining a threshold $v^*(c, \mu)$ under appropriate regularity conditions (see Bénabou and Tirole (2006)).

2. So does that of *MPL* (Proposition 2), except for costless revelation. As with discrete types, equilibrium involves: (i) separation up to some $v^\dagger$, decreasing in μ, with bids solving $b^{MPL}(0) = 0$ and $b^{MPL}(v) = \arg\max_b\{-\int_v^b (\tilde{c} - v)g(\tilde{c})d\tilde{c} + \mu\hat{v}(b)\}$, hence

$b'(v)[b(v)-v] = \mu/g(b(v)) > 0$; (ii) observationally deontological pooling at $b^{MPL}(v) = c_{max}$ by all $v > v^{\dagger}$.

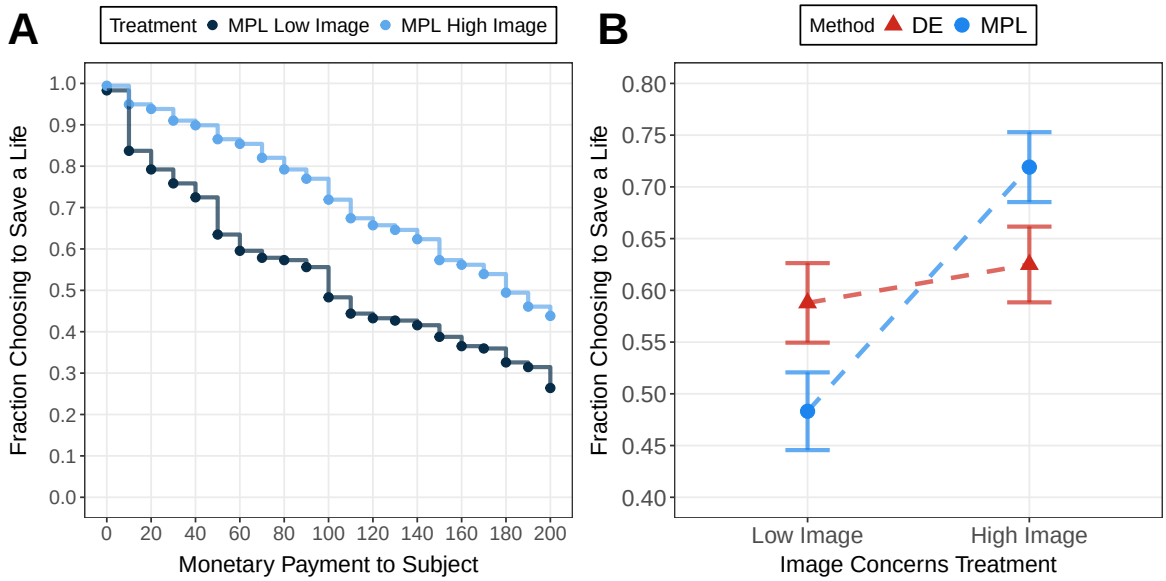
3. In Proposition 3, the first and third results are unchanged: contributions under both schemes are sincere for $\mu = 0$, then increase continuously with μ, for each type and at any cost level (H1); and *MPL* delivers more contributions than *DE* for large μ (H3), as the cheap-act effect induces Kantian-like pooling at c_{max} by more (lower) types. What becomes ambiguous is the comparison at low μ (H2), which depends in complex ways on the agent's type (low enough v's always contribute more under *DE,* high enough ones under *MPL*), the cost level c, and the entire distributions $G(c)$ and $F(v)$.

Figure 1: Equilibrium under Direct Elicitation (panel A) and Multiple-Price List (panel B)



Notes: P_0: pooling at $a_H = a_L = 0$; S: separation, $a_H = 1, a_L = 0$; SS: semi-separation, $a_H = 1, a_L \in (0, 1)$; P_1: pooling at $a_H = a_L = 1$.

Figure 2: Main Experimental Results



Notes: Panel A displays the fractions of subjects that choose to save a life for each offered price in the MPL Low Image and MPL High Image treatments. Panel B shows the interaction effect of elicitation method and image concerns, by displaying the fractions of subjects that choose to save a life with MPL and DE, under either the Low Image or the High Image treatment. Error bars indicate the standard error of the mean.

Table 1: Regression analyses of the effect of the elicitation method on prosocial behavior

Panel A:				
Dependent variable:	Choice to Save a Life (vs. 100€)			
	Low Image		*High Image*	
	(1)	(2)	(3)	(4)
MPL	−0.105	−0.103	0.094	0.091
	(0.054)	(0.053)	(0.050)	(0.050)
Constant (*DE*)	0.588	0.626	0.625	0.622
	(0.038)	(0.049)	(0.037)	(0.046)
Controls		X		X
Observations	343	343	354	354

Panel B:		
Dependent variable:	Choice to Save a Life (vs. 100€)	
	(1)	(2)
MPL	-0.105	-0.097
	(0.054)	(0.053)
High Image	0.037	0.052
	(0.053)	(0.052)
MPL X High Image	0.199	0.190
	(0.073)	(0.072)
Constant (*DE Low Image*)	0.588	0.595
	(0.038)	(0.044)
Controls		X
Observations	697	697

Panel C:		
Dependent variable:	Choice of Voucher (vs. 10€)	
	(1)	(2)
MPL No-Image	0.045	0.051
	(0.047)	(0.047)
Constant (*DE No-Image*)	0.253	0.227
	(0.033)	(0.047)
Controls		X
Observations	366	366

Notes: The table shows OLS regression coefficients. The dependent variable in Panel A is an indicator variable equal to one if the subject chose a donation that saves a human life and zero if the subject chose 100€ for themselves. "MPL" is an indicator variable equal to one if the subject was part of the *MPL* treatment and zero if the subject was part of the *DE* treatment. Columns (1) and (2) display the results for the *Low Image* treatment, and columns (3) and (4) for the *High Image* treatment. The dependent and independent variables in Panel B are the same as in Panel A, with the addition of the variable "High Image", which is an indicator variable equal to one if the subject was part of the *High Image* treatment and zero if the subject was part of the *Low Image* treatment. The dependent variable in Panel C is an indicator variable equal to one if the subject chose a voucher to a university online shop and zero if the subject chose 10€ for themselves. "MPL No-Image" is an indicator variable equal to one if the subject was part of the *MPL No-Image* treatment and zero if the subject was part of the *DE No-Image* treatment. Robust standard errors in parentheses. Controls include age, gender, income, religiousness, educational level, and high school grade.

References

Andreoni, James (1989). "Giving with Impure Altruism: Applications to Charity and Ricardian Equivalence". *Journal of Political Economy* 97 (6): 1447–58. [3]

Andreoni, James (1990). "Impure Altruism and Donations to Public Goods: A Theory of Warm-Glow". *Economic Journal* 100 (401): 464–77. [3]

Ariely, Dan, Anat Bracha, and Stephan Meier (2009). "Doing Good or Doing Well? Image Motivation and Monetary Incentives in Behaving Prosocially". *American Economic Review* 99 (1): 544–55. [2]

Ashraf, Nava, Oriana Bandiera, and B. Kelsey Jack (2014). "No Margin, No Mission? A Field Experiment on Incentives for Public Service Delivery". *Journal of Public Economics* 120: 1–17. [2]

Baillon, Aurelien, Yoram Halevy, and Chen Li (2022). "Randomize at Your Own Risk: On the Observability of Ambiguity Aversion". *Econometrica* 90 (3): 1085–107. [3]

Barmettler, Franziska, Ernst Fehr, and Christian Zehnder (2012). "Big Experimenter Is Watching You! Anonymity and Prosocial Behavior in the Laboratory". *Games and Economic Behavior* 75 (1): 17–34. [9]

Bartling, Björn, Vanessa Valero, Roberto Weber, and Yao Lan (2022). "Public Discourse and Socially Responsible Market Behavior". *Working Paper*, [2]

Bénabou, Roland, Armin Falk, and Luca Henkel (2022). "Ends versus Means: Kantians, Utilitarians and Moral Decisions". *Working Paper*, [3]

Bénabou, Roland, and Jean Tirole (2006). "Incentives and Prosocial Behavior". *American Economic Review* 96 (5): 1652–78. [2, 14]

Bénabou, Roland, and Jean Tirole (2011a). "Identity, Morals, and Taboos: Beliefs as Assets". *Quarterly Journal of Economics* 126 (2): 805–55. [2]

Bénabou, Roland, and Jean Tirole (2011b). "Laws and Norms". *NBER Working Paper 17579*, [2]

Berry, James, Greg Fischer, and Raymond Guiteras (2020). "Eliciting and Utilizing Willingness to Pay: Evidence from Field Trials in Northern Ghana". *Journal of Political Economy* 128 (4): 1436–73. [3]

Bock, Olaf, Ingmar Baetge, and Andreas Nicklisch (2014). "Hroot: Hamburg Registration and Organization Online Tool". *European Economic Review* 71: 117–20. [10]

Bos, Olivier, and Martin Pollrich (2020). "Optimal Auctions with Signaling Bidders". *Working Paper*, [3]

Bos, Olivier, and Tom Truyts (2022). "Entry in First-Price Auctions with Signaling". *Working Paper*, [3]

Brandts, Jordi, and Gary Charness (2011). "The Strategy versus the Direct-Response Method: A First Survey of Experimental Comparisons". *Experimental Economics* 14 (3): 375–98. [3]

Brock, J. Michelle, Andreas Lange, and Erkut Y. Ozbay (2013). "Dictating the Risk: Experimental Evidence on Giving in Risky Environments". *American Economic Review* 103 (1): 415–37. [3]

Bursztyn, Leonardo, and Robert Jensen (2017). "Social Image and Economic Behavior in the Field: Identifying, Understanding, and Shaping Social Pressure". *Annual Review of Economics* 9: 131–53. [1]

Butera, Luigi, Robert Metcalfe, William Morrison, and Dmitry Taubinsky (2022). "Measuring the Welfare Effects of Shame and Pride". *American Economic Review* 112 (1): 122–68. [12, 13]

Charness, Gary, Uri Gneezy, and Brianna Halladay (2016). "Experimental Methods: Pay One or Pay All". *Journal of Economic Behavior and Organization* 131: 141–50. [3]

Charness, Gary, Uri Gneezy, and Alex Imas (2013). "Experimental Methods: Eliciting Risk Preferences". *Journal of Economic Behavior and Organization* 87: 43–51. [3]

Chen, Daniel L., and Martin Schonger (2016). "A Theory of Experiments: Invariance of Equilibrium to the Strategy Method of Elicitation and Implications for Social Preferences". *TSE Working Paper, no. 16-724*, [3]

Chen, Daniel L., and Martin Schonger (2022). "Social Preferences or Sacred Values? Theory and Evidence of Deontological Motivations". *Science Advances* 8 (19): eabb3925. [3, 12]

Chen, Daniel L., Martin Schonger, and Chris Wickens (2016). "oTree-An Open-Source Platform for Laboratory, Online, and Field Experiments". *Journal of Behavioral and Experimental Finance* 9: 88–97. [10]

Cohen, Jonathan, Keith Marzilli Ericson, David Laibson, and John Myles White (2020). "Measuring Time Preferences". *Journal of Economic Literature* 58 (2): 299–347. [3]

Cole, Shawn, A Nilesh Fernando, Daniel Stein, and Jeremy Tobacman (2020). "Field Comparisons of Incentive-Compatible Preference Elicitation Techniques". *Journal of Economic Behavior and Organization* 172: 33–56. [3]

Cox, James C., Vjollca Sadiraj, and Ulrich Schmidt (2015). "Paradoxes and Mechanisms for Choice under Risk". *Experimental Economics* 18 (2): 215–50. [3]

Dana, Jason, Roberto A. Weber, and Jason Xi Kuang (2007). "Exploiting Moral Wiggle Room: Experiments Demonstrating an Illusory Preference for Fairness". *Economic Theory* 33 (1): 67–80. [2]

DellaVigna, Stefano, John A. List, and Ulrike Malmendier (2012). "Testing for Altruism and Social Pressure in Charitable Giving". *Quarterly Journal of Economics* 127 (1): 1–56. [2]

Exley, Christine L. (2016). "Excusing Selfishness in Charitable Giving: The Role of Risk". *Review of Economic Studies* 83 (2): 587–628. [2]

Falk, Armin (2021). "Facing Yourself - A Note on Self-Image". *Journal of Economic Behavior and Organization* 186: 724–34. [2]

Falk, Armin, and Thomas Graeber (2020). "Delayed Negative Effects of Prosocial Spending on Happiness". *Proceedings of the National Academy of Sciences* 117 (12): 6463–68. [8]

Falk, Armin, Thomas Neuber, and Nora Szech (2020). "Diffusion of Being Pivotal and Immoral Outcomes". *Review of Economic Studies* 87 (5): 2205–29. [2, 3]

Feddersen, Timothy, Sean Gailmard, and Alvaro Sandroni (2009). "Moral Bias in Large Elections: Theory and Experimental Evidence". *American Political Science Review* 103 (2): 175–92. [2]

Giovannoni, Francesco, and Miltiadis Makris (2014). "Reputational Bidding". *International Economic Review* 55 (3): 693–710. [3]

Gneezy, Uri, Elizabeth A. Keenan, and Ayelet Gneezy (2014). "Avoiding Overhead Aversion in Charity". *Science* 346 (6209): 632–35. [3]

Goeree, Jacob K. (2003). "Bidding for the Future: Signaling in Auctions with an Aftermarket". *Journal of Economic Theory* 108 (2): 345–64. [3]

Goeree, Jacob K., Charles A. Holt, and Susan K. Laury (2002). "Private Costs and Public Benefits: Unraveling the Effects of Altruism and Noisy Behavior". *Journal of Public Economics* 83 (2): 255–76. [3]

Grossman, Zachary (2015). "Self-Signaling and Social-Signaling in Giving". *Journal of Economic Behavior and Organization* 117: 26–39. [2]

Grossman, Zachary, and Joël J. van der Weele (2017). "Self-Image and Willful Ignorance in Social Decisions". *Journal of the European Economic Association* 15 (1): 173–217. [2]

Kolappan, C., R. Subramani, V. Kumaraswami, T. Santha, and P. R. Narayanan (2008). "Excess Mortality and Risk Factors for Mortality among a Cohort of TB Patients from Rural South India". *International Journal of Tuberculosis and Lung Disease* 12 (1): 81–86. [8]

Ledyard, John O. (1995). "Public Goods: A Survey of Experimental Research". In *The Handbook of Experimental Economics*. Alvin E. Rot and John H. Kagel, ed. vol. 1, Princeton University Press, 111–94. [3]

Miller, Klaus M., Reto Hofstetter, Harley Krohmer, and Z. John Zhang (2011). "How Should Consumers' Willingness to Pay Be Measured? An Empirical Comparison of State-of-the-Art Approaches". *Journal of Marketing Research* 48 (1): 172–84. [3]

Sandel, Michael J. (2012). *What Money Can't Buy: The Moral Limits of Markets*. Farrar, Straus and Giroux. [13]

Selten, Reinhard (1967). "Die Strategiemethode Zur Erforschung Des Eingeschränkt Rationalen Verhaltens Im Rahmen Eines Oligopolexperiments". In *Beiträge Zur Experimentellen Wirtschaftsforschung*. H. Sauermann, ed. Tübingen: Mohr, 136–68. [3]

Straetemans, Masja, Philippe Glaziou, Ana L. Bierrenbach, Charalambos Sismanidis, and Marieke J. van der Werf (2011). "Assessing Tuberculosis Case Fatality Ratio: A Meta-Analysis". *PLoS ONE* 6 (6): [8]

Tiemersma, Edine W., Marieke J. van der Werf, Martien W. Borgdorff, Brian G. Williams, and Nico J.D. Nagelkerke (2011). "Natural History of Tuberculosis: Duration and Fatality of Untreated Pulmonary Tuberculosis in HIV Negative Patients: A Systematic Review". *PLoS ONE* 6 (4): [8]

Van Leeuwen, Boris, and Ingela Alger (2021). "Estimating Social Preferences and Kantian Morality in Strategic Interactions". *Working Paper*, [3]