Eliciting Moral Preferences under Image Concerns: Theory and Experiment

Roland Bénabou Armin Falk Luca Henkel Jean Tirole

November 1, 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 Chicago and University of CEMA, CESifo, JILAEE. 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).

Research transparency: The experimental instructions can be found at https://osf.io/354sb.

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.[5]

[5]See Appendix D for further discussion of our paradigm's advantages relative to standard small-stakes games, e.g., the dictator game.

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).[6] 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.

[6] 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€.[7]

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 E.

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: $\bar{a}^{DE}(c,\mu_H) > \bar{a}^{DE}(c,\mu_L)$, $\bar{a}^{MPL}(c,\mu_H) > \bar{a}^{MPL}(c,\mu_L)$.

[7] 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).[8] 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).

[8]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,[9] but also provide potential guidance for maximizing public-goods contributions and image manipulations.[10]

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

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

[10] 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):

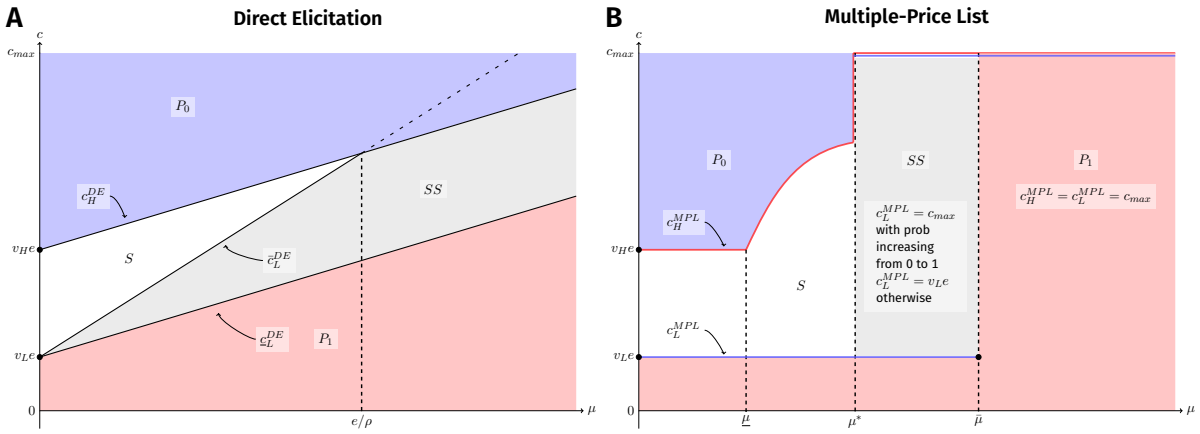
1. The characterization of *DE* (Proposition 1) carries over, with type v now contributing at c if $b^{DE}(v) \equiv v + \mu(E[v'|v' > v] - E[v'|v' < v]) > c$, defining a threshold $v^*(c, \mu)$ under appropriate regularity conditions (see Bénabou and Tirole (2006)).

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 $v > v^{\dagger}$.

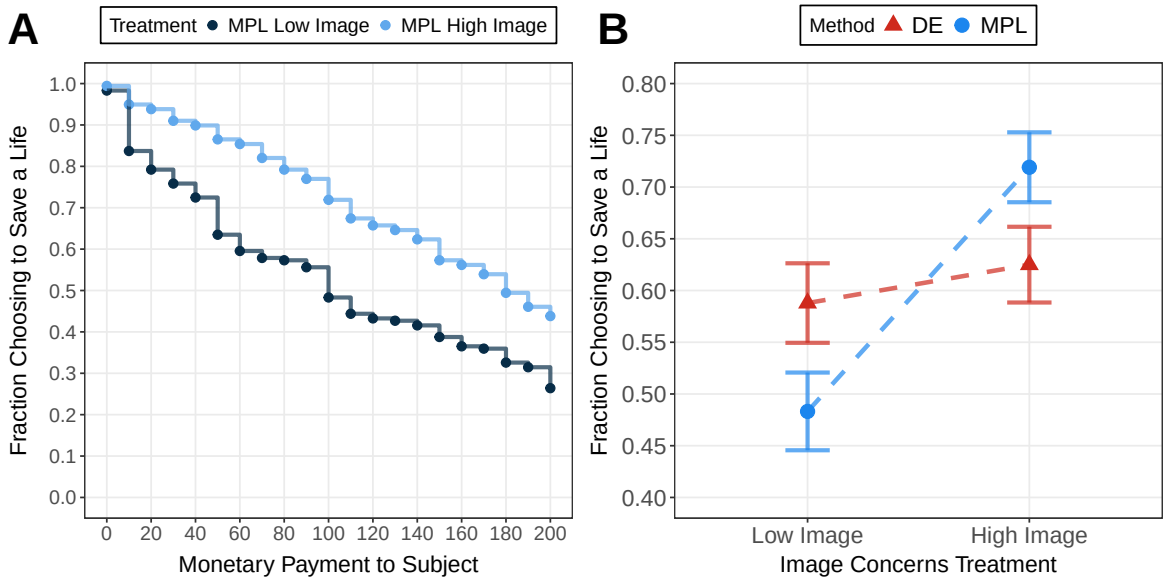
3. In Proposition 3, the first and third results are unchanged: contributions under both schemes are sincere for $\mu = 0$, then increase continuously with μ, for each type and at any cost level (H1); and *MPL* delivers more contributions than *DE* for large μ (H3), as the cheap-act effect induces Kantian-like 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, 34]

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]

ONLINE APPENDIX

A Extension: continuum of types

Agents differ in their prosociality v, according to a uniform distribution on $[0, v_{max}]$, where $v_{max} < \infty$. Denote the average v by $\bar{v} = v_{max}/2$. Besides their intrinsic motivation v to do good and their extrinsic motivation (the cost of the action), they also care about their reputational payoff $R \equiv \mu\widehat{v}$, where $\widehat{v}$ is their reputation and μ the (known) intensity of their image concerns.

A.1 Direct elicitation

The supply of prosociality under *DE* is $\bar{a}^{DE}(c) \equiv 1 - \frac{v^{DE}(c)}{v_{max}}$, where the cutoff $v^{DE}(c)$, when interior, solves

$$v^{DE}(c) - c + \mu\Delta = 0$$

and $\Delta \equiv \mathbb{E}(v|v > v^*) - \mathbb{E}(v|v < v^*) = \mu\bar{v}$.[11] We can define v's implicit bid as

$$\begin{aligned} b^{DE}(v) &\equiv \max\left\{c \mid a^{DE}(v,c) = 1\right\} \\ &= \max\left\{c \mid v \geq v^{DE}(c)\right\} = \min\{v + \mu\bar{v}, c_{max}\} \end{aligned}$$

and the image-induced inflation in the willingness to pay (henceforth, WTP) as

$$x^{DE}(v) \equiv b^{DE}(v) - v = \min\{\mu\bar{v}, c_{max} - v\}.$$

A.2 Multiple-Price-List

We consider a uniform cost distribution (as is the case in most experiments, including the one conducted in this paper) with $g(c) = 1$, meaning that we normalize $c_{max} = 1$. An agent's utility when having type v and bidding b is:

$$U(v,b) \equiv \int_0^b (v - \tilde{c})d\tilde{c} + \mu R(b) = vb - \frac{b^2}{2} + \mu R(b)$$

We look for a *separating equilibrium* on an interval $[0, v^\dagger]$ combined with pooling at $c_{\max} = 1$ on $(v^\dagger, v_{max}]$ (a fully separating equilibrium corresponds to $v^\dagger = v_{max}$).

[11] Corner solutions are:

- $v^{DE}(c) = 0$ when $c \leq \mu\bar{v}$, and
- $v^{DE}(c) = v_{max}$ when $c \geq v_{max} + \mu\bar{v}$.

Let $b(v)$ stands for the equilibrium bid of type v. In the separating part of the equilibrium

$$b(0) = 0,$$

and $b(v) = \max_b\{vb - \frac{b^2}{2} + \mu\hat{v}(b)\}$, which satisfies:

$$\frac{db}{dv}[b(v) - v] = \mu.$$

Finally, types $v > v^\dagger$ select $b(v) = 1$: they behave in an observationally deontological (or price insensitive) way, choosing $a = 1$ at any $c \leq c_{max}$ At the jump point $v^\dagger$ (when interior), this type gains extra reputation $\mu[M^\dagger(v^\dagger) - v^\dagger] = \mu\left[\frac{1-v^\dagger}{2}\right]$ by pooling with higher types in this way, but increases the cost of signaling from $\left[b(v^\dagger) - v^\dagger\right]^2/2$ to $\left[1 - v^\dagger\right]^2/2$.

And so,

$$(1 - v^\dagger)^2 - (b(v^\dagger) - v^\dagger)^2 = \mu(v_{max} - v^\dagger). \tag{9}$$

This jump point could be 0 (everyone mimics “Kantian” behavior"), or could be v_{max} when the equilibrium is fully separating. We will consider all of these cases when comparing *DE* and *MPL*. Let

$$x^{MPL}(v) \equiv b(v) - v$$

denote type v's inflation of WTP. Solving for the differential equation yields, for $v < v^\dagger$,

$$x^{MPL}(v) = \mu\big[1 + W(-e^{-1-v/\mu})\big],$$

where $W : [-e^{-1}, \infty) \to [-1, \infty)$ is the Lambert W-function, i.e., the inverse function of xe^x. Since $W(-e^{-1}) = -1$, we have $x^{MPL}(0) = 0$. Moreover, $x^{MPL}(v)$ goes to μ as v goes large, because $W(0) = 0$. Note that $x^{MPL}(v)$ is strictly increasing in v, and so a fortiori is $b^{MPL}(v)$. Moreover, the inflation is increasing in μ:

$$\frac{\partial x^{MPL}}{\partial \mu} = 1 + w + \frac{v}{\mu} \cdot \frac{w}{1+w} = \frac{1 + w - w\ln(-w)}{1+w}$$

where $w = W(-e^{-1-v/\mu}) \in [-1, 0)$, and using

$$we^w = -e^{-1-v/\mu} \implies \frac{v}{\mu} = -\big(\ln(-w) + w + 1\big).$$

The derivative of x^{MPL} with respect to μ is positive because the numerator $1+w-w\ln(-w)$ equals 0 at $w = -1$ and is increasing in w (the denominator is obviously positive).

A.3 Comparing *DE* and *MPL*.

Depending on the magnitudes of v_{max} and μ, different cases may happen.

Case 1: $\frac{2}{\mu} \leq v_{max} \Leftrightarrow c_{max} \leq \mu\bar{v}$. Behavior of all types is observationally deontological, under both *DE* and *MPL:*

$$\forall v: \quad b^{MPL}(v) = b^{DE}(v) = 1.$$

Case 2: $\frac{1}{\mu} \leq v_{max} < \frac{2}{\mu} \Leftrightarrow \bar{c} \leq \mu\bar{v} < c_{max}$. Behavior is always observationally deontological under *MPL,* but not under *DE:*

$$\forall v: \quad b^{MPL}(v) = 1 \quad \text{and} \quad b^{DE}(v) = \min\{v + \mu\bar{v}, 1\}.$$

Case 3: $\leq v_{max} < \frac{1}{\mu}$ (this can only happen when $\mu < 0.5$). The two curves do not intersect because $\Delta \geq 1$. Moreover, b^{MPL} jumps to c_{max} at $v^\dagger \in (0, v_{max})$, but there could be two cases:

(a) If $v^\dagger < 1 - \mu\bar{v}$ (that is, $b^{DE}(v^\dagger) = v^\dagger + \mu\bar{v} < 1$), then b^{MPL} jumps above b^{DE} at $v^\dagger$.

(b) Otherwise, b^{MPL} jumps to c_{max} when b^{DE} is already there. Thus b^{MPL} is below b^{DE} everywhere.

Proof. Suppose first that the two curves intersect. Then we have:

$$\mu\bar{v} = \mu + \mu W(-e^{-1-v/\mu}) \Rightarrow W(-e^{-1-v/\mu}) = \bar{v} - 1 = \frac{v_{max}}{2} - 1 \geq 0,$$

while $W(-e^{-1-v/\mu})$ is always negative, a contradiction. Therefore, the two curves cannot intersect when $\Delta \geq 1$.

Moreover, MPL is fully separating if and only if

$$b(v_{max}) = v_{max} + \mu + \mu W(e^{-1-v_{max}/\mu}) \leq 1.$$

But here, $v_{max} \geq 2$ and we know that $\mu + \mu W(e^{-1-v/\mu}) > 0$ for all v. Thus, MPL cannot be fully separating, and $v^\dagger < v_{max}$. Furthermore, we know that

$$(1 - v^\dagger)^2 - (b(v^\dagger) - v^\dagger)^2 = 2\mu(\bar{v} - \frac{v^\dagger}{2}),$$

which shows that $v^\dagger \neq 0$: $v^\dagger = 0$ would imply that $1 = 2\mu\bar{v} = \mu v_{max}$, a contradiction.

Therefore, because DE reaches $c_{max} = 1$ at $1 - \mu\bar{v}$, either $v^\dagger < 1 - \mu\bar{v}$, and MPL jumps above b^{DE} at $v^\dagger$, or MPL jumps to c_{max} when DE is already there. Q.E.D.

Case 4: $v_{max} < \min\{2, \frac{1}{\mu}\}$

Denote the potential crossing point by $\hat{v} = -\mu\big[\ln(1 - \bar{v}) + \bar{v}\big] > 0$. How does $\hat{v}$ compare with $v^\dagger$, where MPL jumps?

(a) If $\hat{v} \leq v^\dagger$, then two curves intersect at $\hat{v}$, and we have

$$b^{MPL}(\hat{v}) = b^{DE}(\hat{v}) = -\mu \ln(1-\bar{v}).$$

This case includes the situation where MPL is fully separating. Indeed, whenever MPL is fully separating the two curves intersect and none of them reaches c_{max}.

(b) If $\hat{v} > v^\dagger$, b^{MPL} jumps above b^{DE} at $v^\dagger$, and they do not intersect.

Proof. The only claim that we have to prove in Case 4 is that: *whenever MPL is fully separating the two curves intersect and none of them reaches c_{max}.* Note that MPL can be fully separating only under Case 4's conditions, i.e. $v_{max} < \min\{2, \frac{1}{\mu}\}$. Moreover, when MPL is fully separating, $v_{max} \leq b(v_{max}) \leq 1$, and so, $\bar{v} \leq 0.5$. Considering these, we next prove that $\hat{v} = -\mu\big(\ln(1-\bar{v}) - \bar{v}\big) < 2\bar{v} = v_{max}$ for any $\bar{v} \leq 0.5$. Since $\mu < \frac{1}{v_{max}} = \frac{1}{2\bar{v}}$ and $-\big(\ln(1-\bar{v}) + \bar{v}\big) > 0$, we have:

$$\hat{v} = -\mu\big(\ln(1-\bar{v}) + \bar{v}\big) < \frac{-\ln(1-\bar{v})}{2\bar{v}} - \frac{1}{2}.$$

To prove that

$$\frac{-\ln(1-\bar{v})}{2\bar{v}} - \frac{1}{2} < 2\bar{v},$$

we show:

$$h(\bar{v}) \equiv -\ln(1-\bar{v}) - \bar{v} - 4\bar{v}^2 \leq 0 \quad \forall \bar{v} \in [0, 0.5].$$

This is true because $h(0) = h'(0) = 0$ and $h''(v) = \frac{1}{(1-\bar{v})^2} - 8 < 0 \; \forall \bar{v} \in [0, 0.5]$.

Finally, as we showed, the two curves intersect once and only once in the interval $[0, v_{max}]$, because MPL is fully separating. Moreover, after the intersection, $b^{DE}(v_{max})$ is below $b^{MPL}(v_{max})$. Therefore, $b^{DE}(v_{max}) < b^{MPL}(v_{max}) \leq 1$. Q.E.D.

Comparing DE and MPL for large and small μ. It is easy to see that, fixing v_{max}, for μ large (but not excessively large, i.e. we are in Case 2), all contribute under MPL, but not under DE. In contrast, as μ becomes small:

- If $v_{max} < 2$, Case 4(a) applies and b^{MPL} cuts b^{DE} at some v converging to 0 as μ goes to 0, and remains higher afterward.
- If $v_{max} \geq 2$, Case 3(a) applies and b^{MPL} is below b^{DE} before $v^\dagger$ and jumps above b^{DE} at $v^\dagger < 1 - \mu\bar{v}$.

Proof. Note that $v_{max} < \frac{1}{\mu}$ for small μ. When $v_{max} < 2$, Case 4 applies. Moreover, since $\hat{v}$ goes to 0 and $v^\dagger$ goes to $\min\{1, v_{max}\}$ as μ goes to 0, we have $\hat{v} < v^\dagger$ for small μ. On the other hand, when $v_{max} \geq 2$, Case 3 applies. Moreover, $v^\dagger < 1 - \mu\bar{v}$ for small enough μ, by

the following lemma (putting $k = \bar{v}$ and $\alpha = 1$). We actually prove a stronger version than is needed here, as this will prove useful later on. Q.E.D.

Note that $v^\dagger$ depends on μ. Thus, we sometimes use $v^\dagger_\mu$ to emphasize this point.

Lemma 1. *When $2 \leq v_{max} < \frac{1}{\mu}$, for any k and α such that $\alpha > \frac{1}{2}$ or ($\alpha = \frac{1}{2}$ and $k < 2\bar{v} - 1$), there exists $\delta > 0$ such that for all $\mu < \delta$:*

$$v^\dagger_\mu < 1 - (k\mu)^\alpha.$$

Proof. Let us define

$$h(v, \mu) \equiv (1 - v)^2 - (b(v) - v)^2 - 2\mu(\bar{v} - \frac{v}{2}).$$

Note that $v^\dagger_\mu$ must solves $h(v^\dagger_\mu, \mu) = 0$. Moreover, $h(0, \mu) = 1 - 2\mu\bar{v} > 0$, and $h(1, \mu) = -2\mu(\bar{v} - \frac{1}{2}) < 0$. Additionally,

$$\begin{aligned}\frac{\partial h(v,\mu)}{\partial v} &= -2(1 - v) - 2(b'(v) - 1)(b(v) - v) + \mu \\ &= -2(1 - v) - 2\mu + 2(b(v) - v) + \mu = 2(b(v) - 1) - \mu < 0.\end{aligned}$$

So $v^\dagger_\mu$ exists and we can prove that $v^\dagger_\mu < 1 - (k\mu)^\alpha$ by showing $h(1 - (k\mu)^\alpha, \mu) < 0$. We have:

$$h(1 - (k\mu)^\alpha, \mu) = (k\mu)^{2\alpha} - \Big(\mu + \mu W(-e^{-1-\frac{1}{\mu}+k^\alpha\mu^{\alpha-1}})\Big)^2 - 2\mu\bar{v} + \mu(1 - (k\mu)^\alpha).$$

Therefore, denoting $w = W(-e^{-1-\frac{1}{\mu}+k^\alpha\mu^{\alpha-1}})$,

$$\begin{aligned}\frac{d}{d\mu}h(1 - k\mu^\alpha, \mu) &= 2\alpha k^{2\alpha}\mu^{2\alpha-1} - 2(\mu + \mu w)\Big(1 + w + \mu \cdot \frac{1}{\mu^2}\big(1 + (\alpha - 1)(k\mu)^\alpha\big)\frac{w}{1 + w}\Big) \\ &\qquad - 2\bar{v} + 1 - (1 + \alpha)(k\mu)^\alpha \\ &= 2\alpha k^{2\alpha}\mu^{2\alpha-1} - 2(1 + w)\Big(\mu + \mu w + \big(1 + (\alpha - 1)(k\mu)^\alpha\big)\frac{w}{1 + w}\Big) \\ &\qquad - 2\bar{v} + 1 - (1 + \alpha)(k\mu)^\alpha.\end{aligned}$$

Thus,

$$\frac{d}{d\mu}h(1 - (k\mu)^\alpha, \mu)|_{\mu=0} = 1 - 2\bar{v} + 2\alpha k^{2\alpha}\mu^{2\alpha-1}|_{\mu=0} = \begin{cases} 1 - 2\bar{v} & \text{if } \alpha > \frac{1}{2} \\ 1 - 2\bar{v} + k & \text{if } \alpha = \frac{1}{2} \end{cases}.$$

This is negative since $\bar{v} > 1$, and when $\alpha = \frac{1}{2}$ we have $k < 2\bar{v} - 1$. Moreover, $h(1, 0) = 0$. Therefore, for μ close enough to zero, $h(1 - (k\mu)^\alpha, \mu) < 0$. Q.E.D.

Comparing average inflations. Let us also compute the average bid inflations from 0 to v, $A^{MPL}(v)$ and $A^{DE}(v)$. We do not multiply by the uniform density $f(v) = \frac{1}{v_{max}}$, which means we calculate averages times v_{max}. For $v \leq 1 - \mu\bar{v}$ we have:

$$A^{DE}(v) = \int_0^v x^{DE}(s)ds = \int_0^v \mu\bar{v}ds = \mu\bar{v}v.$$

Moreover, for $v \leq v^\dagger$ we have:

$$A^{MPL}(v) = \int_0^v x^{MPL}(s)ds = \mu v + \mu \int_0^v W(-e^{-1-s/\mu})ds.$$

Denoting $w = W(-e^{-1-s/\mu})$, we have $we^w = -e^{-1-s/\mu}$ by the definition of Lambert W function, and so,

$$s = -\mu\big(\ln(-w) + w + 1\big) \implies ds = -\mu\big(\frac{1}{w} + 1\big)dw.$$

Define

$$\begin{aligned} AW(v) \equiv \int_0^v W(-e^{-1-s/\mu})ds &= -\mu \int_{-1}^{W(-e^{-1-v/\mu})} (1+w)dw \\ &= -\mu\Big(W(-e^{-1-v/\mu}) + \frac{1}{2}W(-e^{-1-v/\mu})^2 + \frac{1}{2}\Big). \end{aligned}$$

Let us focus on the case where μ is small. Therefore, $W(-e^{-1-v/\mu}) \approx 0$, and the first-order approximation is:

$$AW(v^\dagger) \approx -\frac{\mu}{2} \implies A^{MPL}(v^\dagger) = \mu v^\dagger + \mu AW(v^\dagger) \approx \mu v^\dagger.$$

- Suppose $v_{max} < 2$ and denote $v^+ = \min\{v^\dagger, 1 - \mu\bar{v}\}$. Thus,

  $$A^{DE}(v^+) = \mu v^+ \bar{v} < \mu v^+ \approx A^{MPL}(v^+)$$

  Moreover, for $v > v^+$ we know that b^{MPL} is above b^{DE} because b^{MPL} cuts b^{DE} at $\hat{v}$ close to 0, when μ is small. Therefore,

  $$A^{DE}(v_{max}) < A^{MPL}(v_{max}).$$

- On the other hand, if $v_{max} > 2$, then $v^\dagger < 1 - \mu\bar{v}$, as we showed before. Let us ignore

the part after $1-\mu\bar{v}$, since both bids are c_{max}. Focusing on values up to that point,

$$\begin{aligned} A^{MPL}(1-\mu\bar{v}) &= A^{MPL}(v^\dagger) + \int_{v^\dagger}^{1-\mu\bar{v}} (1-s)ds \\ &= A^{MPL}(v^\dagger) + \frac{1}{2}\big((1-v^\dagger)^2 - (\mu\bar{v})^2\big) \\ &\approx \mu v^\dagger + \frac{1}{2}(1-v^\dagger)^2. \end{aligned}$$

Furthermore,

$$A^{DE}(1-\mu\bar{v}) = \mu\bar{v}(1-\mu\bar{v}) \approx \mu\bar{v}.$$

Therefore,

$$\begin{aligned} (A^{MPL} - A^{DE})(1-\mu\bar{v}) &\approx \mu v^\dagger + \frac{1}{2}(1-v^\dagger)^2 - \mu\bar{v} \\ &> \mu v^\dagger + \frac{1}{2}(2\bar{v} - \frac{3}{2})\mu - \mu\bar{v} = \mu(v^\dagger - \frac{3}{4}). \end{aligned}$$

The inequality holds for small enough μ, and comes from Lemma 1, using $k = 2\bar{v} - \frac{3}{2}$ and $\alpha = \frac{1}{2}$. Consequently, since $v^\dagger$ goes to 1 as μ goes to zero, for small enough μ we have:

$$A^{DE} < A^{MPL}.$$

Thus, the average inflation factor is always higher under *MPL* for μ small when the distributions of types and costs are both uniform. Note that we only considered first-order approximations with respect to μ and found that $A^{MPL} - A^{DE}$ is positive and of order μ.

B Decision Screens

Figure B.1: Decision Screen DE

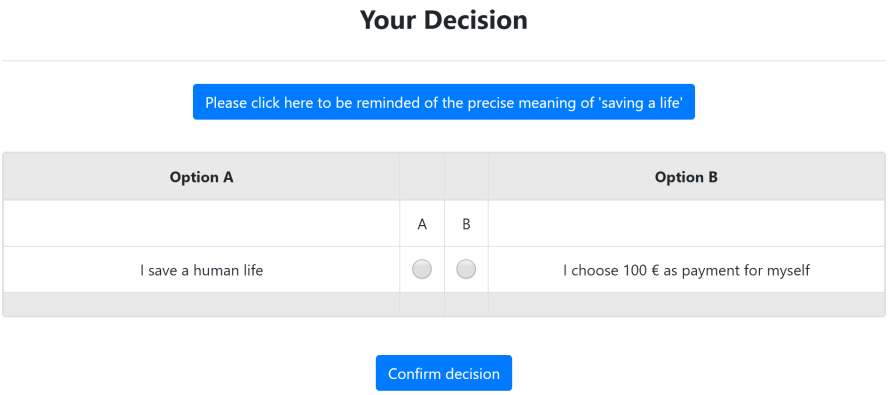


Figure B.2: Decision Screen MPL

Your Decisions

Please click here to be reminded of the precise meaning of 'saving a life'

Option A				Option B
	A		B	
I save a human life	○	1	○	I choose 0 € as payment for myself
I save a human life	○	2	○	I choose 10 € as payment for myself
I save a human life	○	3	○	I choose 20 € as payment for myself
I save a human life	○	4	○	I choose 30 € as payment for myself
I save a human life	○	5	○	I choose 40 € as payment for myself
I save a human life	○	6	○	I choose 50 € as payment for myself
I save a human life	○	7	○	I choose 60 € as payment for myself
I save a human life	○	8	○	I choose 70 € as payment for myself
I save a human life	○	9	○	I choose 80 € as payment for myself
I save a human life	○	10	○	I choose 90 € as payment for myself
I save a human life	○	11	○	I choose 100 € as payment for myself
I save a human life	○	12	○	I choose 110 € as payment for myself
I save a human life	○	13	○	I choose 120 € as payment for myself
I save a human life	○	14	○	I choose 130 € as payment for myself
I save a human life	○	15	○	I choose 140 € as payment for myself
I save a human life	○	16	○	I choose 150 € as payment for myself
I save a human life	○	17	○	I choose 160 € as payment for myself
I save a human life	○	18	○	I choose 170 € as payment for myself
I save a human life	○	19	○	I choose 180 € as payment for myself
I save a human life	○	20	○	I choose 190 € as payment for myself
I save a human life	○	21	○	I choose 200 € as payment for myself

Confirm decisions

C Robustness Experiment

In the main experiment, we showed how image concerns lead to differences in moral behavior between elicitation methods. One concern is that there are factors present in our experiment that lead to differences between DE and MPL independent of image concerns. In particular, the previous literature has identified two main factors that could potentially confound the comparison between the two elicitation methods in our case.

First, in our experiment, only a subset of subjects had their decision implemented for real. In the MPL treatments, another randomization takes place, which is absent in DE: if selected for payout, one decision of the price list is randomly selected. If subjects violate the independence axiom and view these two randomization processes not separately but rather as a meta-lottery, this could potentially affect the comparison. This issue is also present in the many experiments that study decisions over lotteries and pay only one lottery out for real. In this context, it is usually assumed that subjects evaluate the different random processes in isolation, an assumption that has been repeatedly validated empirically[12]. It is natural to assume that subjects also perceive the two processes in isolation in our experiment since they were introduced and explained at two different points in the instructions.

The second factor is the so-called compromise effect (Andersen et al., 2006; Birnbaum, 1992; Simonson, 1989). When presenting a price list, the focus lies perceptually on the center. This in turn could change the attractiveness of the options appearing in the middle of the price list, biasing answers away from the subject's true valuations. To control for this effect, we carefully selected the DE value to correspond to the value precisely in the middle of the price list in the MPL treatments. As such, it seems unlikely that differences in perceptions could explain discrepancies between the elicitation methods.

Therefore, we would not expect differences between DE and MPL in our experiment once image concerns are absent. Nevertheless, in order to document this empirically, we conducted a robustness experiment, which is explained next.

C.1 Setup and Treatments

For the robustness experiment, we used a good that is unrelated to prosocial and moral considerations, so that image concerns are plausibly absent. For this non-moral good, we chose a 35 € voucher for the University of Bonn's online shop. With the voucher, subjects can buy sweatshirts, T-shirts, and accessories related to the university. The voucher cannot be returned and is only valid for purchases in the shop. There were two between-subject treatments: *DE No-Image* and *MPL No-Image*. In the former, subjects could choose between 10 € and the voucher, while in the latter they faced a price list from 0€ to 20€ in 1€ increments. Note that this closely mimics the decisions in the main experiment. The only

[12] See e.g., Starmer and Sugden (1991), Cubitt, Starmer and Sugden (1998) and Hey and Lee (2005).

difference is that all values are divided by 10. As in the main experiment, subjects were paired with another subject, and only a subset of subjects had their choices implemented for real.

Accordingly, instructions for the decisions were identical, with the sole difference being that descriptions related to the saving a life paradigm were replaced with descriptions of the voucher. Consequently, any factors influencing the comparison between *DE* and *MPL* in the main experiment should also manifest in the robustness experiment.

C.2 Procedure

Subjects were recruited from the same subject pool as the main experiment, with the restriction that they had not previously participated in the main experiment. The experiment was conducted as a virtual lab experiment since in-person lab sessions were not possible due to the ongoing Covid-19 pandemic. That is, the experiment started and ended at a prespecified date and time, and the experimenter was available during the experiment in case of problems.

In total, 366 subjects (227 female, mean age 26.88, SD 7.87) took part, 188 in the *MPL No-Image*, and 178 in the *DE No-Image* treatment, respectively. The experiment lasted on average 13 minutes, for which the subjects received a show-up fee of 3€. Subjects were grouped in virtual sessions consisting of roughly 24 subjects, and one pair was randomly selected for payout out of each virtual session. Exactly as in the main experiment, for these two subjects, either their DE decision was implemented or a randomly chosen decision from the MPL list.

C.3 Results

Assessing subjects' general valuation of the voucher, we observe considerable variation in switching behavior in the *MPL No-Image* treatment. In total, 76% had an interior switching value, meaning they preferred the voucher in the initial decision but switched to preferring the monetary value at some point. The variation compares quite favorably to the *MPL-Low Image* treatment, where this was the case for 72% of subjects. Comparing the choice at 10 € in *MPL No-Image* with *DE No-Image*, we find that 29.8% choose the voucher in *MPL* and 25.3% in *DE*. This difference is small in magnitude and not statistically significant ($p = 0.35$; two-sided Fisher's exact test). It is also in the opposite direction of what we find in the main experiment for the *Low Image* case, which is the natural comparison. Table C.1 replicates this null result in an OLS-regression, with column (2) using the same variables as control variables as in the main experiment, compare Table 1, columns (2) and (4). Thus, we do not observe any meaningful differences between the two elicitation methods in our setting once image concerns are removed.

Table C.1: Regression analyses of the effect of the elicitation method on voucher choice

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
R^2	0.003	0.039

The table shows OLS regression coefficients. Robust standard errors in parentheses. Controls include age, gender, income, religiousness, educational level, and high school grade.

References

Andersen, Steffen, **Glenn W. Harrison**, **Morten Igel Lau**, and **E. Elisabet Rutström** (**2006**). "Elicitation using multiple price list formats ". *Experimental Economics*, 9(4), 383-405.

Birnbaum, Michael H. (**1992**). "Violations of Monotonicity and Contextual Effects in Choice-Based Certainty Equivalents ". *Psychological Science*, 3(5), 310-315.

Cubitt, Robin P., **Chris Starmer**, and **Robert Sugden** (**1998**). "On the Validity of the Random Lottery Incentive Mechanism ". *Experimental Economics*, 1, 115-132.

Hey, John D., and **Jinkwon Lee** (**2005**). "Do Subjects Separate (or Are They Sophisticated)? ". *Experimental Economics*, 8, 233-265.

Simonson, Itamar (**1989**). "Choice Based on Reasons: The Case of Attraction and Compromise Effects ". *Journal of Consumer Research*, 16(2), 158.

Starmer, Chris, and **Robert Sugden** (**1991**). "Does the Random-Lottery Incentive System Elicit True Preferences? An Experimental Investigation ". *American Economic Review*, 81, 971-978.

D Further details on the experimental paradigm

We chose the *Saving a Life* paradigm in order to generate situations in which one choice is unambiguously perceived as the more moral one. As described in Section 3.1, the paradigm faces subjects with a decision between triggering a live-saving donation $(a = 1)$ or taking money for themselves $(a = 0)$. This decision engages a universal notion of morality: helping a person who is suffering from a potentially deadly, yet easily preventable disease. Hence, it is clear to (almost) everyone what the "right thing to do" is and what a moral person should choose. This is reinforced by the donation being cost-effective: the amount is well above all monetary payments possible for the subjects themselves, as described later, and the money is directly used to treat patients, without any administrative or transaction cost.

The potentially extreme consequences of the donation decision also increase the likelihood that subjects will take the choice task and its signaling/reputational implications seriously. To further ensure that subjects make choices with sufficient deliberation, we provided them with comprehensive explanations about the disease, the treatment, the veracity of the experiment, and so on. In particular, the experimental instructions explain in detail the consequences of the donation and, conversely, what the absence of the donation entails (see Section E). The charity to which the donation is designated has identified individuals who are suffering from tuberculosis in the regions where it operates. Because tuberculosis is a highly lethal disease if untreated, many people die there every year. The donation was calculated to enable the charity to cure as many patients as needed (five) so that, in expectation, one among them is saved from death by tuberculosis. The calculation includes information on the treatment success rate of Operation ASHA, the mortality rate for alternative treatments by the state tuberculosis program in India, the different detection rates for new cases of tuberculosis, other causes of mortality, etc. Without the donation, conversely, five patients are not treated who would otherwise have been, implying that in expectation one will succumb to the disease.

As our experimental results will show (Section 4.2), subjects do take this information and the decision seriously. If they were not convinced of the credibility or effectiveness of the donation, they would simply take the money for themselves, no matter how little was offered. However, as indicated by choices in the *MPL* treatments, subjects show a clear sensitivity to the monetary amount offered, and even at 200€ under anonymity, there are still 26% who do not take the money. For similar findings, see Falk and Graeber (2020).

E Instructions

This appendix shows an excerpt of the experimental instructions, focusing on the description of the *Saving a Life* paradigm. The full instructions are available at the following link: https://osf.io/354sb.

E.1 Announcement by the Experimenter

The following text was read aloud by the experimenter after all subjects were placed in their cubicles, establishing common knowledge among all subjects of a session. The content depended on the image treatment.

E.1.1 Treatment *Low Image*

Welcome to today's study. In today's study, you will make decisions on a computer. These decisions will take place under complete anonymity. To ensure this, we will now apply the following procedure: You should all have two notes with your cubicle number in front of you. We will soon collect one of the two notes and randomly draw one out of all collected. The person in the drawn cubicle is responsible for the payment in today's study. At the end of the study, we prepare sealed envelopes with your payments. Those envelopes are then passed to the soon to be randomly drawn person, who will hand them out to each of you sequentially in the adjacent room. The envelopes are designed so that you cannot see the contents from the outside, i.e., not on weight or similar clues. Hence at no time can there be a connection drawn between your payment and your decisions. Please hold now one of the notes with your cubicle number onto out of your cubicle. (Responsible person is drawn and placed in the adjacent room) The study will begin shortly. If you have at any time have questions, just hold your hand out of the cubicle.

E.1.2 Treatment *High Image*

Welcome to today's study. In today's study, you will make decisions on your computer. Your decisions will subsequently be evaluated by a committee consisting of three students from the University of Bonn. For this, after you have made your decisions, you will go to the adjacent room, where your decisions will be projected on a wall with a projector. You will then briefly communicate your decisions to the committee, and the committee will evaluate them. Afterward, you will receive the result of the evaluation. Detailed information about your decisions, the committee, and the evaluation will be given to you at the appropriate time on your computer. The study will begin shortly. If you have at any time have questions, just hold your hand out of the cubicle.

E.1.3 Further Procedure

After the text was read aloud, in the *Low Image* conditions the experimenter then collected one note from each subject indicating their respective cabin number. All notes were thrown into a bag, and one was drawn in front of all participants to make clear that the person responsible for the payment procedure was a randomly determined participant. In the *High Image* conditions, subjects were shown the adjacent room and the setup with the committee, which consisted of student research assistants. The members of the committee did not interact with the subjects in any way.

E.2 Introduction

All further instructions were displayed on the subjects screens. The following introduction was the same for all treatments.

E.2.1 Welcome to the study

Welcome, and thank you for your interest in today's study!
For your participation, you will receive a fixed payment of 12€ given to you at the end. In this study, you will make decisions on the computer. Depending on how you choose, you can earn additional money.
During the entire study, communication between participants is prohibited. Please turn off your phone so that other participants are not disturbed. Please only use the designated functions on the computer and make the entries with the mouse and keyboard. If you, at some point, have questions, please make a hand signal. Your question will be answered at your seat.
On the next screens, you will receive specific information about participation in this study. To proceed, click "Next".

E.2.2 Your Partner

As part of this experiment, a partner has been assigned to you. This partner is a participant in today's experiment, just like you. He or she was randomly assigned to you and will receive the same instructions as you.
In today's experiment, you and your partner will both receive the exact same information and subsequently face the exact same decisions. These decisions have certain consequences, which will be described in detail later.
At the end of today's experiment, one pair is randomly drawn from all participants in today's experiment. Only the decisions of this pair will be implemented, as described in the instructions. Please note: The random draw of a pair is completely independent of the participants' decisions. Each pair has the same probability of being drawn. Since your decision can be

actually implemented for real, you should think carefully about how you will decide in the experiment.

E.2.3 Information about Tuberculosis

What follows is important information that is relevant to the decisions you will later be asked to make. It concerns the illness tuberculosis and its possible treatment. Please read through all the information carefully.

What is Tuberculosis?

Tuberculosis – also called Phthisis or White Death – is an infectious disease, which is caused by bacteria. Roughly one-third of all humans are infected with the pathogen of Tuberculosis. Active Tuberculosis breaks out among 5 to 10% of all those infected. Tuberculosis is primarily airborne. This is also why quick treatment is necessary.

Tuberculosis patients often suffer from very unspecific symptoms like fatigue, the feeling of weakness, lack of appetite, and weight loss. At an advanced stage of lung tuberculosis, the patient coughs up blood, leading to the so-called rush of blood. Without treatment, a person with Tuberculosis dies with a probability of 43%.

How prevalent is Tuberculosis?

In the year 2014, 6 million people have been recorded as falling ill with active Tuberculosis. Almost 1.5 million people die of Tuberculosis each year. This means more deaths due to Tuberculosis than due to HIV, malaria, or any other infectious disease.

Is tuberculosis curable?

According to the World Health Organization (WHO), the United Nations agency for international public health, “tuberculosis is preventable and curable”. Treatment takes place by taking antibiotics several times a week over a period of 6 months. It is important to take the medication consistently. Since 2000, an estimated 53 million lives have been saved through effective diagnosis and treatment of tuberculosis.

The success rate of treatment for a new infection is usually over 85%.

The preceding figures and information have been provided by the WHO and are freely available. Click here for more details.

Operation ASHA

Operation ASHA is a charity organization specialized since 2005 on treating Tuberculosis in disadvantaged communities. The work of *Operation ASHA* is based on the insight that the biggest obstacle for the treatment of Tuberculosis is the interruption of the necessary 6-month-long regular intake of medication.

For a successful treatment, the patient has to come to a medical facility twice a week – more than 60 times in total – to take the medication. Interruption or termination of the treatment is fatal because this strongly enhances the development of a drug-resistant form of Tuberculosis. This form of Tuberculosis is much more difficult to treat and almost always

Figure E.1: Typical appearance of a tuberculosis patient.

Figure E.2: A worker from Operation ASHA delivers medication to a tuberculosis patient.

leads to death.

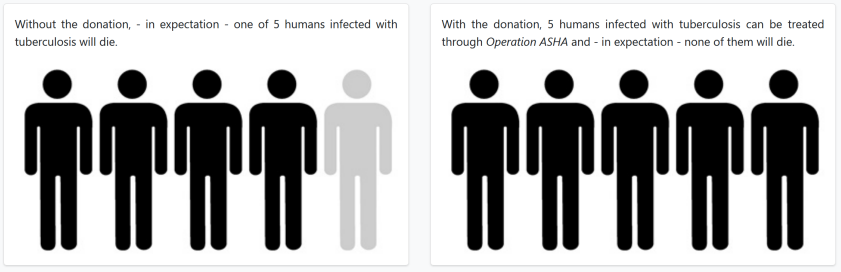
The Concept of *Operation ASHA*

To overcome this problem, *Operation ASHA* developed a concept that guarantees regular treatment through immediate spatial proximity to the patient. A possible non-adherence is additionally prevented by visiting the patient at home.

By now, *Operation ASHA* runs more than 360 treatment centers, almost all of which are located in the poorer regions of India. More than 60,000 sick persons have been identified and treated that way.

Operation ASHA is an internationally recognized organization, and its success has been covered by the New York Times, BBC, and Deutsche Welle, for example. The MIT and the University College London have already conducted research projects about the fight against Tuberculosis in cooperation with *Operation ASHA*. The treatment method employed by *Operation ASHA* is described by the World Health Organization (WHO) as “highly efficient and cost-effective”.

Figure E.3: Relationship between the donation and the saving of a life



The Impact of a Donation to *Operation ASHA*

It is now possible to save people from death by Tuberculosis by donating to *Operation ASHA*. To save a person's life means here to successfully cure a person with Tuberculosis, who otherwise would die because of the Tuberculosis. A donation of 350€ ensures that at least one human life can be expected to be saved. The information used to calculate the donation amount is obtained from public statements from the World Health Organization (WHO), peer-reviewed research studies, Indian Government statistics, and published figures from *Operation ASHA*.

In the calculation, information was conservatively interpreted, or a pessimistic number was used so that the donation amount of 350€ is in the case of doubt higher than the actual costs to save a human life. In addition, in the calculation of the treatment success rate of *Operation ASHA*, the mortality rate for alternative treatment by the state tuberculosis program in India and the different detection rates for new cases of Tuberculosis are included.

In the context of this study, an agreement made with *Operation ASHA* will ensure that 100% of the donation will be used exclusively for the diagnosis and treatment of tuberculosis patients. This means that every Euro of the donation amount goes directly to saving human lives, and no other costs will be covered. Based on a very high number of cases, the contribution of a donation of 350€ can be simplified visualized as follows:

With a donation of 350€ 5 additional patients infected with Tuberculosis can be treated through *Operation ASHA*.

If these 5 persons are not treated through *Operation ASHA*, it is expected that one patient will die.

If, through the donation of 350€ all 5 patients are treated, it is expected that no patient will die.

Based on this experience, this means that through a donation of 350€ the life of a human will be saved. The relationship between a donation of 350€ and the saving of a human is illustrated in the following graphic: [Figure E.3 here]

Summary

Tuberculosis is a worldwide common bacterial infectious disease. The success rate of medical treatment of a new disease is very high. Nevertheless, close to 1.5 million people die every year from Tuberculosis. The biggest obstacle to the curing of Tuberculosis is the po-

tential stopping of continuous treatment with antibiotics. The concept of *Operation ASHA* is therefore based on the immediate proximity to the patient as well as the control and recording of the regular intake of medication. Through a donation of 350€ to *Operation ASHA*, a life will be saved.

How is the donation connected to the saving of a life?

The donation of 350€ already accounts for the fact that someone inflicted with the illness could have survived without treatment by *Operation ASHA*; i.e., instead of through *Operation ASHA*, they could have received treatment through other actors (such as the public health system). The amount is, therefore, sufficient for the diagnosis and complete treatment of multiple sufferers.

What does it mean to "save a life"?

To save a life means here the successful curing of a person suffering from Tuberculosis, who otherwise would die because of Tuberculosis. In particular, this means that the amount of the donation is sufficient to identify and cure so many tuberculosis patients that there is at least one person among them who otherwise could be anticipated to have died of Tuberculosis.

Note

Click on "Next" once you have finished carefully reading through the information.

You can only click on the button "Next" once you have spent at least 5 minutes on the tabs of this page.