Eliciting Moral Preferences

Roland Bénabou[1], Armin Falk[2], Jean Tirole[3]

This version: August 12, 2018[4]

[1]Princeton University, NBER, CEPR, CIFAR, briq, IZA, BREAD, and THRED.

[2]Institute of Behavior and Inequality (briq) and Department of Economics, University of Bonn.

[3]Toulouse School of Economics (TSE) and Institute for Advanced Study in Toulouse (IAST), University of Toulouse Capitole.

[4]Ana Luisa Dutra, Juliette Fournier, Pierre-Luc Vautrey and Ben S. Young provided superb research assistance. 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 Programme Grant Agreement no. 249429 and no. 340950, as well as European Union's Horizon 2020 research and innovation programme, Grant Agreement no. 669217).

Abstract

We investigate how much of a person's deep moral preferences can be retrieved from observing their choices, for instance via experiments, and in particular how one should interpret behaviors that appear deontologically rather than consequentially motivated. Comparing the performance of the direct elicitation (DE) and multiple-price list or Becker-DeGroot-Marschak (BDM) mechanisms, we characterize in each case how (social or self) image motives inflate the extent to which agents behave prosocially –e.g., refuse "bribes" for causing harm. More surprisingly, the signaling bias is shown to depend on the elicitation method, both per se and interacted with the level of visibility: it is greater under DE for low enough reputation concerns, and greater under BDM when they become high enough. We also provide conditions ensuring a single crossing. We further show that, even when all agents are consequentialists, certain Kantian behaviors and postures easily emerge under BDM (but not DE) when reputation becomes important enough, with both high and low-morality agents turning down all prices within the offered range.

Keywords: Moral behavior, Kantian reasoning, utilitarianism, consequentialism, social norms, deontology, preference elicitation, multiple price list

JEL Codes: D62, D64, D78.

"In the kingdom of ends everything has either a price or a dignity. What has a price can be replaced by something else as its equivalent; what on the other hand is above all price and therefore admits of no equivalent has a dignity" (Kant, 1785).

1 Introduction

This paper investigates two related questions. How should one interpret Kantian-like behaviors that appear to be deontologically rather than consequentially motivated? More generally, how much of a person's deep moral preferences can be retrieved from observing their choices, in particular through different experimental methods?

A long-standing debate within Western moral philosophy opposes consequentialist versus deontological reasoning as the proper foundation of ethics. Utilitarians justify normative principles and actions in terms of their consequences, aiming at maximizing the Good –pleasure, happiness, desire satisfaction, or welfare (Bentham, 1789, Mill 1861, Johnson 2014, Sinnott-Armstrong 2003). In contrast, rule-based ethics value actions *per se* rather than through their consequences, thus giving priority to the Right over the Good (Alexander and Moore 2015). In particular, Kant's *practical imperative* commands one to "Act in such a way that you always treat humanity, whether in your own person or in the person of any other, never simply as a means, but always at the same time as an end."[1] This means, essentially, a *prohibition on tradeoffs* between personal or even social gains and any harm to others, or to (future) self. Such reasoning is commonly invoked by people rejecting "indecent" proposals, or to argue that certain "sacred values" –human life, integrity of the body, autonomy, freedom, etc.– are incommensurable with money (should have an infinite price), so that transactions or markets where they could be seen as being "for sale" ought to be prohibited, irrespective of participants' willingness and potential gains from trade.

Empirically, behavior seems to reflect a mix of utilitarian and deontological behaviors –both across individuals and sometimes within each one. The literature on cooperation and voluntary contribution to public goods provides evidence that prosocial choices are generally sensitive to the implied consequences (Kagel and Roth 1995, Chapter 2; Goeree et al. 2002). Likewise, charitable giving decreases when the risk of having no impact rises (Brock et al. 2013), or when overhead increases (Gneezy et al. 2014). At the same time, there is also evidence of "warm glow" altruism in which utility seems to be derived from the act as such, with donations fairly insensitive to the level of contributions by others or that of government funding (e.g., Andreoni 1989, 1990). In the well-known trolley dilemma, or less abstractly in the equivalent "organ-transplant" problem (should a benevolent State sacrifice every year the lives of some N randomly selected individuals so as to save $10N$ others from death?), consequentialist reasoning requires "actively" killing one to save many, while deontological reasoning calls for obeying the imperative not to kill, regardless of consequences and without reference to any ends. Faced with such dilemmas, different people choose differently, and even minor variations in framing can drastically alter their (hypothetical) decisions.

[1] This is sometimes also referred to as the "second formulation" of the *categorical imperative*. The first (and better known) one, which is to "Act only on that maxim through which you can at the same time will that it should become a universal law," can be understood as helping the individual visualize, and thus internalize, the externalities involved his choices. In that sense it is more consequentialist (or rather, "rule-consequentialist") in nature, and we formalize related mechanisms in our companion paper (Bénabou, et al. 2018).

Experiments in which subjects choose between money and a charitable act, under varying probabilities that their decision will actually be implemented, also point to a mix of motives. Feddersen et al. (2009) and Chen and Schonger (2013) note that if the same probability applies to costs and benefits, the behaviors of pure consequentialists (in the traditional sense that excludes image concerns) and that of people with purely expressive or/and deontological preference should both be entirely invariant to the odds. In practice, they find that the implementation probability does matter, which reveals a tradeoff between the two types of motives. In Falk and Szech (2017), participants in a group vote simultaneously on killing a (surplus) laboratory mouse, or on destroying a charitable donation, in return for money. In both paradigms (mice and donation), the decision of whether or not to take the "bribe" is now individual and non-contingent, while the likelihood that the harm materializes varies. The frequency of moral choices is found to decrease as the probability of being pivotal falls, in line with consequentialism, but even as it reaches zero there remain about 18% of subjects (in both cases) who turn down the money "on principle," even though they understand (as verified by elicited beliefs) that the chance this will actually do any good is zero. In a sequential version of the charity paradigm, which makes not being pivotal even clearer, there are still 5% who choose to contribute, even conditional on knowing that the overall donation has already been destroyed, rendering their "moral" decision completely non-consequential.

In the field, people widely participate in elections in spite of infinitesimal odds of being pivotal, presumably an expression of a moral or civic "duty;" at the same time, participation is sensitive to the expected closeness of the election, and also responds to the social visibility of the voting process (Funk 2010). Moral attitudes also differ widely with respect to the regulation of so-called repugnant goods (Roth 2007, Elias et al. 2016). While advocates of markets for organs, sex or surrogate mothers point to the large potential welfare gains, among the opponents are some who fully acknowledge those benefits, yet give priority to the imperative never to treat human beings as the means to an end, no matter how beneficial the latter might be.

To address these questions, we build on a simple workhorse model of individual moral behavior (Bénabou and Tirole 2006, 2011). An agent can take or abstain from a "moral" action, namely one that confers a positive externality on others, at some cost to himself. Individuals differ in their intrinsic valuations for doing so and, in line with the abundant evidence on the pursuit of social and self image, derive reputational benefits from being perceived, or seeing themselves, as having high moral values. In our companion paper (Bénabou et al. 2018), we expand this basic framework to formalize the construction and dissemination of *moral arguments,* distinguishing two broad types: narratives, which provide situation-specific justifications for some course of action (formally, signals and messages about the tradeoff between private costs and social benefits); and imperatives, which are broad "commandments" to follow some rule unconditionally, issued without providing explicit reasons.

The present paper leaves aside these communication aspects to explore a different set of questions, though also relating to imperatives: we ask here what can be learned about a person's moral values by observing their choices not just in a single instance, but over a broad range of situations. The objective terms of the moral tradeoff (cost, externality, reputational impact) are thus now always common knowledge between actor and observer(s), but allowed to vary across choice situations. This occurs naturally in a long-run relationship, and is also the very paradigm of controlled experiments.

We thus extend the basic moral-choice framework to compare the two revealed-preference methods most commonly used, namely direct elicitation (DE) and the multiple-price list or Becker-DeGroot-Marschak (BDM) mechanism. In both cases, (self) reputational motives predictably inflate the extent to which agents behave prosocially –e.g., refuse "bribes" for causing harm, with the formal analysis showing precisely by how much. More surprisingly, the signaling bias is shown to *depend on the elicitation method,* both *per se* and *interacted* with the level of visibility: it is greater under DE for low enough reputation concerns, and greater under BDM when they become high enough. We also provide conditions ensuring a single crossing.

We further show that, even when all agents are consequentialists, certain Kantian behaviors and postures easily emerge under BDM (but not DE): when reputation become important enough, both high and low-morality agents *turn down all prices* within the offered range. This second set of results provides a methodological caveat for both experiments and contingent-valuation surveys, and also fits well with the common propensity to respond to moral dilemmas through "righteous" indignation and rigid postures rather than cost-benefit analysis.

Related economics literature. There is now a substantial literature, both theoretical and experimental, exploring the interactions between moral behavior, material or social incentives, and markets. The paper bridges the two lines of work that respectively focus on the role of signaling or moral-identity concerns (e.g., Bénabou and Tirole 2006a, 2011a,b, Ellingsen and Johannesson 2008, Mazar et al. 2008, Ariely et al. 2009, Exley 2016, DellaVigna et al. 2016, Gino et al. 2016, Grossman and van der Weele 2017) and on that of pivotality, whether real or deontologically "imagined" à la Kant (Brekke et al. 2003, Roemer 2010, Falk and Szech 2013, 2017, Ambuehl 2016, Elias et al. 2016). With respect to experimental methodology, it also contributes to the study of prosocial preferences and alternative elicitation mechanisms (e.g., Goeree et al. 2002, Feddersen at al. 2009, Brandts and Charness 2011, Chen and Schonger 2013, Charness et al. 2016).

2 Model

1. Preferences. We start here from the basic framework from Bénabou and Tirole (2006, 2011), which we will extend to richer choice situations such as multiple-price elicitation mechanisms. Agents are risk-neutral and have a two-period horizon, $t = 1, 2$. At date 1, an individual has an opportunity to engage in moral behavior ($a = 1$) or act selfishly ($a = 0$). Choosing $a = 1$ is prosocial in that it involves a personal cost $c > 0$ but generates a valuable externality or public good, normalized to $e \in [0, 1]$; for instance, e may be the probability of an externality of fixed size 1.[2] Agents differ by their intrinsic motivation to act morally: given e, it is either $v_H e$ (high, moral type) or $v_L e$ (low, immoral type), with probabilities ρ and $1 - \rho$ and $v_H > v_L \geq 0$; the average type will be denoted as $\bar{v} = \rho v_H + (1 - \rho) v_L$.

Besides the externality, the second feature of action $a = 1$ that ties it to the moral domain is that it can be reputationally valuable, conferring on the individual a social- or self-image benefit at date 2. In the social context, the agent knows his private type but the audience

[2] In Bénabou et al. (2018) we also allow agents to have imperfect willpower at the moment of choice: the ex-ante cost c of "doing the right thing" is momentarily perceived as c/β, where $\beta \leq 1$ is the individual's degree of self-control or (inverse) hyperbolicity. Because ex-ante versus ex-post choices play no role here, we set $\beta = 1$ throughout (equivalently, c can be reinterpreted as c'/β).

(peer group, firms, potential partners) does not. In the self-signaling context, the individual has an immediate, "intuitive" sense of his deep preferences at the moment of action –say, how much empathy or spite he experiences– but later on the intensity of that feeling is imperfectly accessible ("forgotten"). 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 choosing action $a \in \{0, 1\}$ and μ measures the strength of self or social image concerns, common to all agents. This utility level may be augmented additively by any externalities or public goods generated by others, but since this term is independent of the agents' action we omit it here.

Note that these preferences are consequentialist: an agent's desire to behave prosocially reflects and trades off the externality he expects his actions to have, the personal costs involved, and the reputational consequences.[3]

2. Behavior. As is common in signaling models, multiple equilibria may coexist. For instance, when

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

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. Intuitively, when the high type is expected to abstain the stigma from doing so is lessened, which in turn reduces net reputational value of acting morally. In this and all other cases of multiplicity (see the Appendix), we choose the equilibrium that is best for both types, namely the no-contribution pooling equilibrium. Indeed, the separating equilibrium yields lower payoffs: $\mu \cdot v_L < \mu\bar{v}$ for the low type and $v_H e - c + \mu v_H \leq \mu\bar{v}$ for the high one.[4]

This very simple framework readily implies that an agent is more likely to act morally the higher the perceived externality e and/or his image concern μ, and the lower his initial level of reputational "capital" ρ (keeping actual preferences fixed). As we discuss in Bénabou et al. 2018), these basic predictions align quite well with a broad range of empirical evidence.

3 Measuring Moral Preferences

Consider now what can be learned from people's willingness to accept some harm to other(s) as the means to sufficiently desirable ends, or on the contrary their blanket refusal of such (im)moral tradeoffs. Our framework makes clear, from the start, that two types of choice

[3] The model also allows for genuinely deontological agents, $v_H = +\infty$, but a key point is that they are not required to generate "observationally deontological" behavior in BDM-like experiments, i.e., subjects rejecting all prices (for harming someone else) on some randomly implemented price list with finite or even infinite support.

[4] Pareto dominance is understood here as better for both types of a single individual. Since $a = 1$ has positive externalities, this may be different from (even opposite to) Pareto efficiency understood in the sense of making everyone in society better off. If we instead selected the separating equilibrium there would be more of an alignment for high values of e and less for low values, but all comparative statics would remain the same.

situations must be distinguished.

The first one is where the agent faces a tradeoff is between private cost and public benefit, c and e, so that Kantian attitudes consist in refusing to entertain any implicit or explicit price between personal gain and harm to others –as if v was infinite. This blanket refusal signifies that one's morality or dignity is "not for sale," and often accompanied by indignation at the proposal. Our model will show that such behavior readily emerges even when all actors are actually consequentialists. While there are undeniably exceptional individuals ready to sacrifice even their own life to do "the right thing," morally weaker or just average people have incentives to posture as deontologically motivated, when they really are not.

The second type of "moral-tradeoff aversion" involves choices between two public harms, without any material stake for the decision-maker; as in the trolley or transplant dilemma, however, one harm will occur *"by default"* and the other only through an *"active"* decision. In terms of payoffs this is now just a choice between e and (say) $5e$, making clear that consequentialism dictates a positive answer –and this all more so for a more moral agent (v_H versus v_L). The common refusal of explicit tradeoffs of this second kind is thus harder to rationalize than the first one.[5] One could incorporate it into the model by allowing (some) individuals to incur a "visceral", unobserved fixed cost c' of *actively* inflicting bodily harm on anyone else, and indeed that is what many variations of the trolley paradigm (pushing a person versus a button, etc.), as well as neuroimaging evidence, strongly suggest. Reputational concerns over this private "repugnance" will then kick in as powerfully as they do for v, because in most daily interactions the moral choices faced are of the first kind (me versus others). This makes an agent with a high c' a very desirable and reliable partner,[6] even though in situations of the second type they would fail to make the necessary painful arbitrage; and typically, societies choose for those leadership roles more "steely" individuals.

We shall not pursue this second route here, for two reasons. First, individual trolley-like dilemmas are less frequent than private cost/public benefits tradeoffs. Second, the basic logic is the same as for choice problems of the first type, except that reputational and self-image concerns now considerably *magnify,* rather than completely substitute for, the presence of some truly deontological types (i.e., with a significant c'). In both cases, the "popularity of Kantians" –the esteem accorded to those perceived to hold sacred values and treat moral decisions as a matter of dignity rather than price– leads many who do not really have such preferences to adopt similar choices and postures.

3.1 Direct Elicitation

We first derive an agent's choice of how to act when faced with any given value $c \in \mathbb{R}_+$ of the cost of moral action, or equivalently an incentive or "bribe" to behave immorally. In experimental settings this correspond to a mechanism of direct elicitation, and the strength of reputational

[5] It is also typically accompanied by other puzzles. First, it is highly specific to a few domains (life, health, but not money), to perceived losses rather than gains, and to action versus inaction, all of which make decisions highly sensitive to framing. Second, societies and their citizens make such choices all the time, albeit much less explicitly: conscripting soldiers to defend the country, allocating resources to developing drugs that will cure common rather than rare diseases, etc.

[6] Everett et al. (2016) show that subjects who make deontological judgments in trolley-type moral dilemmas are indeed preferred as social partners, being perceived as more moral and trustworthy.

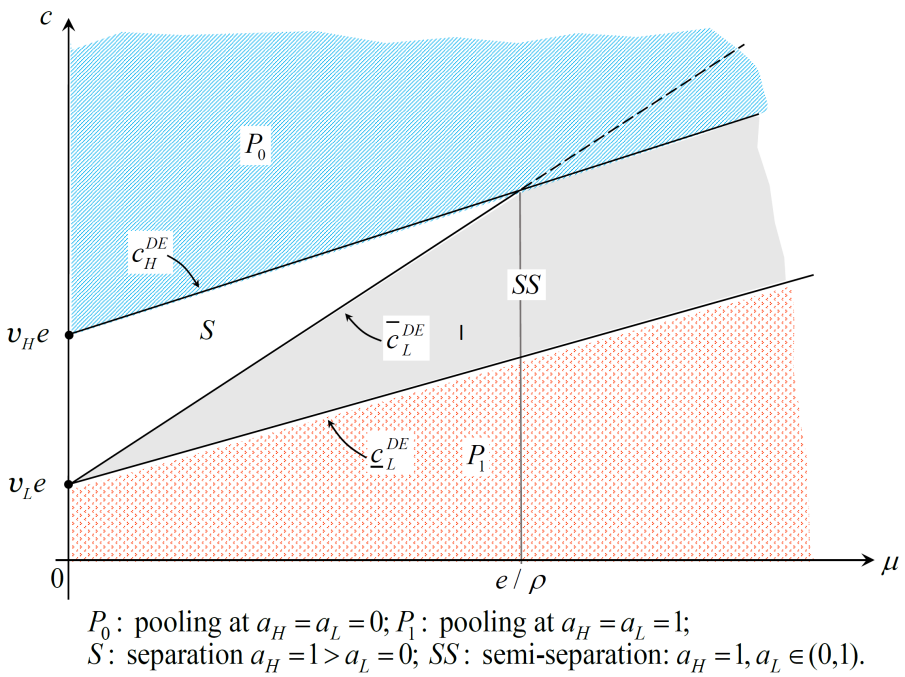

Figure 1: Direct Elicitation

concerns μ can also be modulated by varying the public visibility or private memorability of choices. Equilibrium behavior will be fully characterized by three cost (or incentive) thresholds, as illustrated in Figure 4:

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

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

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

Denoting a_H, a_L the strategies (probabilities of choosing $a = 1$) of the two types, we show the following intuitive results, illustrated in Figure 1 (for the case $\rho < 1/2$).

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

1. *For low prices, $c < \min\{\underline{c}_L^{DE} c_H^{DE}\}$, everyone behaves morally, $a_H = a_L = 1$.*

2. *For intermediate prices, $c \in (\underline{c}_L^{DE}, c_H^{DE})$, the high type still behaves morally ($a_H = 1$), but the low type's probability $a_L(c)$ of doing so decreases with c, and is then equal to 0 for $c \geq \min\{\bar{c}_L^{DE}, c_H^{DE}\}$.*

3. *For high prices, $c \geq c_H^{DE}$, both types behave immorally, $a_H = a_L = 0$.*

Relative to "pure" moral preferences ve, 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 (and hence full revelation of actual preferences) occurs shrinks with μ, and becomes empty for $\mu > e/\rho$.

Having solved for each type's behavior under any opportunity cost c of acting morally, we can also compute the aggregate behavior of agents facing a distribution $G(c)$ on $[0, +\infty)$:

$$\bar{a}^{DE} = \rho G(c_H^{DE}) + (1-\rho) \int_0^{\min\{\bar{c}_L^{DE}, c_H^{DE}\}} a_L(\tilde{c}) dG(\tilde{c}). \tag{5}$$

3.2 Multiple-Price List

We now turn to a standard BDM-type mechanism: each subject is asked at what minimum level of reward c he is willing to take the immoral action $a = 0$, knowing that the actual $\tilde{c}$ will be drawn according to some cumulative distribution $G(\tilde{c})$ on an interval $[0, c_{\max})$. In experiments, this distribution is typically uniform, but here we allow other cases as well, including $c_{\max} = +\infty$. Let $L(c)$ denote the low type's net loss associated with refusing any reward below level c :

$$L(c) \equiv \int_{v_L e}^{c} [\tilde{c} - v_L e] \, dG(\tilde{c}),$$

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

We now solve for both types' willingnesses to accept (WTA) under the multiple-price list, denoted c_H^{BDM} and c_L^{BDM}, as well as the resulting average behavior

$$\bar{a}^{BDM} = \rho G(c_H^{BDM}) + (1-\rho) G(c_L^{BDM}), \tag{6}$$

and ultimately compare them to their counterparts under direct elicitation.

Note first that, *absent reputation concerns,* both mechanisms are *equivalent* and reveal each type's *true* moral preference: $c_H^{DE} = c_H^{BDM} = v_H e$, $\underline{c}_L^{DE} = \bar{c}_L^{DE} = c_L^{BDM} = v_L e$.

When $\mu > 0$, there are again three cases to consider, summarized in Figure 5.

Proposition 2 (BDM elicitation). *The outcome of the BDM mechanism is as follows:*

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

   Initially, for $\mu \leq L(v_H e)/(v_H - v_L)$, 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. Semi-separation: *when μ is intermediate, $\mu \in [L(c_{\max})/(v_H - v_L), L(c_{\max})/\rho(v_H - v_L)]$, the low type seeks to pool while the high type seeks to separate, so their WTA's escalate until the high type becomes observationally deontological, $c_H^{BDM} = c_{\max}$, while the low type randomizes between that same "virtuousness" ($c_L^{BDM} = c_{\max}$) and revealing himself (accepting $c_L^{BDM} = v_L e$), with a probability $a_L(\mu)$ increasing in μ.*

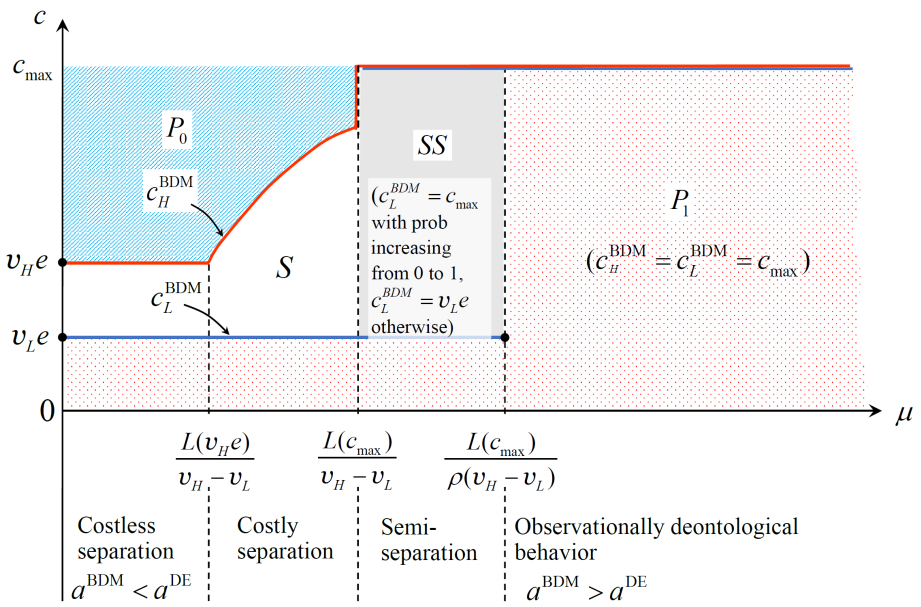

Figure 2: Multiple-Price List

3. Pooling: *when* $\mu > L(c_{\max})/\rho(v_H - v_L)$, *(self) image concerns are strong enough that both types' behavior is observationally deontological*: $c_H^{BDM} = c_L^{BDM} = c_{\max}$.

3.3 Comparison of BE vs. BDM

While the general comparison hinges on the specification of $G(c)$, the extreme cases of low and high μ are relatively simple.

1. Weak image concerns: the discouragement effect. When μ is low enough that separation is costless, we have $c_H^{BDM} = v_H e < c_H^{DE}$ and $c_L^{BDM} = v_L e < \underline{c}_L^{DE}$, therefore $\bar{a}^{DE} > \bar{a}^{BDM}$. Intuitively, BDM raises the cost to the low type of mimicking the high one, since to do so he must be willing to forego up to $v_H e$, and for low reputational gain (small μ) 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 initially flat in Figure 2 until μ becomes high enough, whereas it is linear under DE.

2. Strong image concerns and the cheap-signaling effect. Conversely, at high values of μ reputational concerns become paramount, and the cost of signaling is lower (for both types) under BDM than under DE, as high values of c must be paid only with probability less than 1. It is even bounded by $L(c_{\max}) < \infty$, which limits the extent to which the high type can separate himself, and so for μ high enough full pooling occurs at this maximum: $c_H^{BDM} = c_L^{BDM} = c_{\max} \leq +\infty$ and hence $\bar{a}^{BDM} = 1 > \bar{a}^{DE}$, provided the support of G extends beyond $\underline{c}_L^{DE}$.

Away from the extremes, these two offsetting discouragement and cheap-signaling effects imply that there is generally no unambiguous relationship between the value of μ and the sign of $\bar{a}^{DE} - \bar{a}^{BDM}$. For the uniform distribution universally used in experimental work, however, we are able to provide relatively simple conditions (given in Appendix) ensuring *a single-crossing*

property, namely that $\bar{a}^{BDM} > \bar{a}^{DE}$ if and only if μ exceeds a critical cutoff.

4 Discussion

Our analysis yields two important lessons –one concerning potential bias in the measurement of "true" moral preferences, the other related to policies and welfare judgements based on contingent-valuation surveys.

1. Experiments. A first insight is that even purely utilitarian individuals may act, when facing BDM-like situations, as if deontologically motivated, refusing to consider any of the tradeoffs proposed to them. Their behavior will then resemble that in the initial Kant quote, namely an unwillingness to accept any price in exchange for what is perceived as having dignity. Of course, a definitive empirical test is ultimately impossible: since only finite (realized and expected) amounts can be offered, one cannot rule out that some even higher prices would have been accepted. With this caveat in mind, it is notable that in recent experiments on willingness to kill (surplus) mice, Falk and Szech (2013) find that a sizable fraction of participants refused all offered prices, which ranged up to €50 or even €100. The distribution of switching points also exhibited a bimodal distribution with masses at the minimum and maximum prices, as in Proposition 2(2).[7]

More generally, Propositions 1-2 show that *image concerns* affect the measurement of moral preferences in ways that *interact with the elicitation method.* This also means that regardless of whether one is interested in image-inclusive moral preferences (for positive predictions) or in purely intrinsic ones (for normative judgements), the estimates will differ between direct and price-list mechanisms. These results contrast with the invariance across methods commonly assumed in experimental work, thus raising the possibility of potential biases in the estimation of moral preferences and arguing for caution in the interpretation of those estimates. A related point is made by Chen and Schonger (2015) for other forms of preferences involving moral "duties" in addition to material payoffs.

The three most commonly used methods of elicitation are direct response, direct response with random implementation (not paying every for everyone's decision) and price list (BDM). As explained above: (i) random implementation makes signaling "cheaper," thus inducing more "moral" (but also more hypothetical) decisions when reputation is at stake, and (ii) BDM involves an additional "hurdle" effect that may further inflate, or on the contrary partially deflate, the first one. While there is a fair amount of research comparing how DE versus BDM affect the prevalence of prosocial behavior in *strategic* settings such as ultimatum, trust or public goods games (Brandts and Charness 2011), there is –to the best of our knowledge– none for non-

[7] There were four uniform price-list treatments, offering 20 choices each (in steps of €2.50 or €5), with two price ranges and two framings of the decision context. The fractions of "observationally deontological" participants were 18.8 and 27.1% (with $n = 192$) for a price support of $[e2.50, e5]$, and still 7.4 and 11.1% (with $n = 248$) for $[e5, e100]$. The price-list data were used to calibrate parameters in the paper's main experiment, and are available upon request.

interactive contexts like the one analyzed here.[8] Concerning simple random versus deterministic implementation, a recent overview article by Charness et al. (2016) reports generally mixed effects.[9] Propositions 1-2 suggest a new type of experiment, interacting elicitation method with degree of visibility (or/and prosocial nature of the choices) and testing in particular whether the single-crossing condition discussed above holds empirically.

2. Policy The second issue is the interpretation and use for policy-making of contingent-valuation surveys. The fact that respondents' stated willingness to pay for non-market goods such as environmental preservation or risk abatement often seems unreasonably high, and invariant to the size of the public good, is usually attributed to the "cheap talk" nature of most such surveys (the actual c faced is zero; see Figure 2). Note, however, that even for incentive-compatible elicitations and realistically envisioned tradeoffs (e.g., calibrated from charitable contributions), the willingness to pay includes the value of social and self-image (the latter, even under anonymity). If the purpose of measurement is a positive one of predicting or explaining behavior this is appropriate (people will actually pay up to this amount), but our results caution that different elicitations methods may yield quite different results. If the purpose is normative, moreover, then even real-money WTP's can substantially overstate the true *social* value of the public good in question. That is because they incorporate the *private* image gains from contributing, even though in the aggregate these are exactly offset by the image losses of non-contributors.

A related phenomenon is the common resistance to estimating and using a "statistical value of life." Despite the fact that we implicitly engage in trading off costs and lives all the time (limiting access to medical treatments, setting pollution standards or health-and-safety regulations), explicit reference to putting a price tag on life typically produces righteous and conspicuously displayed indignation (e.g., Sandel 2012).

3. Future research. We have focussed here on people's (un)willingness to engage in moral tradeoffs along one very important dimension, namely those involving private costs or/and benefits on one hand (c and $\mu\hat{v}$), social consequences of their own actions on the other (e). Another dimension, which we leave for further study concerns tradeoffs between alternative public goods or bads (e vs. e') where one involves an "active" choice and the other not (as in trolley-like problems), or/and both reflect the purely private utility gains or losses derived by other people from immoral or "repugnant" transactions.

[8] Two exceptions can be obtained from experiments that used price lists to calibrate a subsequent binary-choice paradigm. In Falk (2017), subjects faced a uniform price list from €1to €20 to determine a point of indifference between receiving money and inflicting an (actual) electric shock on another subject. The cumulative fraction willing to do so for $c \leq$€8 was considerably lower than the corresponding fraction in direct-response binary choice of whether or not to shock in return for €8. In the data from Falk and Szech (2013), on the other hand, the fraction willing to kill mice for €10 in a binary choice is not statistically different from that willing to do so for €10 or less when facing a 20-step uniform price list (ranging from €2.50 to €50).

[9] For example, Sefton (1992) finds that dictators were less generous in a condition where all participants are paid, relative to only a random subset. Clot et al. (2018), however, find no significant difference.

Appendix

Proof of Proposition 1 From conditions (2)-(4), it is straightforward to characterize the regions in which each possible equilibrium exists:

(P_0) Pooling at $a_H = a_L = 0$, sustained by out-of equilibrium beliefs v_H following $a = 1$ (by the D1 criterion), is an equilibrium if and only if $c \geq c_H^{DE}$. By the same reasoning as given below Assumption 1, it is then either unique or better for both types (Pareto dominant) than any other equilibrium with which it may coexist.

(P_1) Pooling at $a_H = a_L = 1$, sustained by out-of equilibrium beliefs v_L following $a = 0$ (by the D1 criterion), is an equilibrium if and only if $c \leq \underline{c}_L^{DE}$.

(S) Separation, namely $a_H = 1 - a_L = 1$, is is an equilibrium if and only if $\bar{c}_L^{DE} \leq c \leq \bar{c}_H^{DE}$, where $\bar{c}_H^{DE} > c_H^{DE}$ is defined by $v_H e - c_H^{DE} + \mu\,(v_H - \bar{v}) \equiv 0$.

(SS_1) Semi-separation with $0 < a_L < 1 = a_H$, and beliefs $\hat{v} \in (v_L, \bar{v})$ 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 and $v_L e - c + \mu(\hat{v}(a_L) - v_L) = 0$ and the Bayesian posterior $\hat{v}(c) = [\rho v_H + (1-\rho)a_L v_L] \,/\, [\rho v + (1-\rho)a_L]$, which leads to:

$$v_L e - c + \frac{\mu\rho(v_H - v_L)}{\rho + (1-\rho)a_L(c)} \equiv 0, \tag{7}$$

so that $a_L(c)$ is decreasing in c, and the reputation $\hat{v}(c)$ following $a = 1$ conversely increasing.

(SS_0) Semi-separation with $0 = a_L < a_H < 1$, and 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 thus 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 below the first cutoff, SS_1 between the first and second, and S between the second and third. Above the third, the dominant equilibrium is P_0.

(b) If $\underline{c}_L^{DE} < c_H^{DE} < \bar{c}_L^{DE}$ (where the second inequality means that $\mu\rho > e$), the unique equilibrium is P_1 below the first cutoff, and SS_1 between the first and second; above that, the dominant equilibrium is P_0.

(b) If $c_H^{DE} < \underline{c}_L^{DE} < \bar{c}_L^{DE}$ (where the first inequality means that $\mu(2\rho - 1) > e$), the unique equilibrium is P_1 below the first cutoff, and above it the dominant equilibrium is P_0. ■

Sorting condition for the comparison of DE and BDM elicitation. We proceed in five steps.

1. Direct elicitation. Equations (2)-(4) can be rewritten as:

$$\begin{aligned} c_H^{DE} &= v_H e + \mu(1-\rho)(v_H - v_L), \\ \bar{c}_L^{DE} &= v_L e + \mu(v_H - v_L), \\ \underline{c}_L^{DE} &= v_L e + \mu\rho(v_H - v_L), \end{aligned}$$

while the low types' mixed strategy for $c \in [\underline{c}_L^{DE}, \min\{\bar{c}_L^{DE}, c_H^{DE}\})$ is

$$a_L(c) = \frac{\rho}{1-\rho}\left[\frac{\mu(v_H - v_L)}{c - v_L e} - 1\right].$$

Substituting into (5) for a uniform distribution and focusing on the case where $\underline{c}_L^{DE} \leq c_H^{DE} \leq c_{\max}$ yields

$$\bar{a}^{DE} = \rho\frac{c_H^{DE}}{c_{\max}} + (1-\rho)\frac{\underline{c}_L^{DE}}{c_{\max}} + \int_{\underline{c}_L^{DE}}^{\bar{c}_L^{DE}} (1-\rho)a_L(\tilde{c})\frac{d\tilde{c}}{c_{\max}} \quad \mu < e/\rho,$$
$$\bar{a}^{DE} = \rho\frac{c_H^{DE}}{c_{\max}} + (1-\rho)\frac{\underline{c}_L^{DE}}{c_{\max}} + \int_{\underline{c}_L^{DE}}^{c_H^{DE}} (1-\rho)a_L(\tilde{c})\frac{d(\tilde{c})}{c_{\max}} \quad \text{for } \mu \geqslant e/\rho.$$

Hence:

$$\bar{a}^{DE}(\mu) = \begin{cases} \frac{1}{c_{\max}}\left[\bar{v}e + \mu\rho(v_H - v_L)(1-\rho-\log\rho)\right], & \text{if } \mu < e/\rho \\ \frac{1}{c_{\max}}\left[v_L e + \mu\rho(v_H - v_L)\left(1 + \log\left(\frac{e}{\mu} + 1 - \rho\right) - \log\rho\right)\right], & \text{if } \mu \geqslant e/\rho \end{cases}. \| \tag{8}$$

2. Multiple Price List. With the uniform distribution $L(c) = (c - v_L e)^2/(2c_{\max})$, so the three cutoffs for μ defined in Proposition 2 are given by:

$$\mu_0 = \frac{(v_H - v_L)e^2}{2c_{\max}}, \quad \mu_1 = \frac{1}{2c_{\max}}\frac{(c_{\max} - v_L e)^2}{v_H - v_L}, \quad \mu_2 = \frac{\mu_1}{\rho}, \tag{9}$$

and the two types' willingnesses to accept equal:

$$c_H^{BDM}(\mu) = \begin{cases} v_H e, & \text{if } \mu < \mu_0 \\ v_L e + \sqrt{2c_{\max}\mu(v_H - v_L)}, & \text{if } \mu \in [\mu_0, \mu_1) \\ c_{\max}, & \text{if } \mu \geqslant \mu_1 \end{cases}, \tag{10}$$

$$c_L^{BDM}(\mu) = \begin{cases} v_L e, & \text{if } \mu < \mu_1 \\ \begin{cases} v_L e, & \text{w.p. } a_L(\mu) \\ c_{\max}, & \text{w.p. } 1 - a_L(\mu) \end{cases}, & \text{if } \mu \in [\mu_1, \mu_2) \\ c_{\max}, & \text{if } \mu \geqslant \mu_2 \end{cases} \tag{11}$$

Substituting into (6), we have

(a) If $\mu < \mu_0$, then

$$\bar{a}^{BDM} = \rho\frac{v_H e}{c_{\max}} + (1-\rho)\frac{v_L e}{c_{\max}} = \frac{\bar{v}e}{c_{\max}}.$$

(b) If $\mu \in [\mu_0, \mu_1)$, then

$$\bar{a}^{BDM} = \rho \frac{(v_L e + \sqrt{2c_{\max}\mu(v_H - v_L)})}{c_{\max}} + (1-\rho)\frac{v_L e}{c_{\max}} = \frac{v_L e}{c_{\max}} + \rho\sqrt{\frac{2\mu(v_H - v_L)}{c_{\max}}}$$

(c) If $\mu \in [\mu_1, \mu_2)$, then

$$\bar{a}^{BDM} = \rho + (1-\rho)\left[a_L(\mu) + (1 - a_L(\mu))\frac{v_L e}{c_{\max}}\right] = \frac{\mu}{\mu_2} + \left(1 - \frac{\mu}{\mu_2}\right)\frac{v_L e}{c_{\max}},$$

since $a_L(\mu) = (\mu - \rho\mu_2) / [(1-\rho)\mu_2]$

For $\mu > \mu_2$, finally, we saw that $\bar{a}^{BDM} = 1$. Summarizing, we have

$$\bar{a}^{BDM}(\mu) = \begin{cases} \frac{\bar{v}e}{c_{\max}}, & \text{if } \mu < \mu_0 \\ \frac{v_L e}{c_{\max}} + \rho\sqrt{\frac{2\mu(v_H - v_L)}{c_{\max}}}, & \text{if } \mu \in [\mu_0, \mu_1) \\ \frac{\mu}{\mu_2} + \left(1 - \frac{\mu}{\mu_2}\right)\frac{v_L e}{c_{\max}}, & \text{if } \mu \in [\mu_1, \mu_2) \\ 1, & \text{if } \mu \geqslant \mu_2 \end{cases} \|$$

3. Auxiliary assumptions. We will focus for the moment on values $\mu \leq 1/\rho$ (the maximum value for which Assumption 1 is feasible for some e). Recall that, in computing $\bar{a}^{DE}$ we assumed that $c_{\max} \geqslant c_H^{DE}(\mu)$. Since c_H^{DE} is increasing, we need only impose this at $\mu = 1/\rho$, which means:

$$c_{\max} \geqslant v_H e + \frac{1-\rho}{\rho}(v_H - v_L). \tag{12}$$

Second, $\bar{a}^{BDM}(\mu) = 1 > \bar{a}^{DE}(\mu)$ for $\mu \geqslant \mu_2$, we need only look for intersections at $\mu < \mu_2$. The above range restriction for μ then requires that $\mu_2 \leq 1/\rho$ (equivalently, $\mu_1 \leq 1$), that is:

$$(c_{\max} - v_L e)^2 \leqslant 2c_{\max}(v_H - v_L). \tag{13}$$

Finally, we also imposed that $c_H^{DE} \geq \underline{c}_L^{DE}$, which means that

$$e + \mu(1-\rho) > \mu\rho. \tag{14}$$

Thus, it must be that either $\rho \leqslant 1/2$, or else $\rho > 1/2$ and $e \geqslant 2 - 1/\rho$.

With these three assumptions, it is always the case that $\mu_0 < e/\rho$, but we still have three possible cases for the remaining cutoffs: $\mu_0 < e/\rho < \mu_1 < \mu_2$, $\mu_0 < \mu_1 < e/\rho < \mu_2$, and $\mu_0 < \mu_1 < \mu_2 < e/\rho$. $\|$

4. Single-Crossing Condition. We will show that, together with (12)–(14) above, the following condition ensures that $\bar{a}^{BDM}(\mu)$ and $\bar{a}^{DE}(\mu)$ cross only once:

$$e + \mu_1(1 - \rho - \log\rho) < 2\mu_1 \frac{c_{\max}}{c_{\max} - v_L e} = \frac{c_{\max} - v_L e}{\rho} \tag{15}$$

It will be useful to define $V_L \equiv v_L e/c_{\max}$ and $V_\Delta = (v_H - v_L)/c_{\max}$, and then from these:

$$A^{DE}(\mu) \equiv \frac{\bar{a}^{DE}(\mu) - V_L}{\rho V_\Delta} = \begin{cases} e + \mu(1 - \rho - \log \rho), & \text{if } \mu < e/\rho \\ \mu\left(1 + \log\left(\frac{e}{\mu} + 1 - \rho\right) - \log \rho\right), & \text{if } \mu \geqslant e/\rho \end{cases}, \tag{16}$$

$$A^{BDM}(\mu) \equiv \frac{\bar{a}^{BDM}(\mu) - V_L}{\rho V_\Delta}. = \begin{cases} e, & \text{if } \mu < \mu_0 \\ \sqrt{\frac{2\mu}{V_\Delta}}, & \text{if } \mu \in [\mu_0, \mu_1) \\ \frac{2\mu}{1 - V_L}, & \text{if } \mu \in [\mu_1, \mu_2] \end{cases}. \tag{17}$$

The BDM cutoffs are thus given by

$$\mu_0 = \frac{V_\Delta e^2}{2}, \quad \mu_1 = \frac{1}{2V_\Delta}(1 - V_L)^2, \quad \mu_2 = \frac{1}{2\rho V_\Delta}(1 - V_L)^2, \tag{18}$$

and Assumptions C1-C3 can be rewritten as:

$$V_L + V_\Delta\left(e + \frac{1 - \rho}{\rho}\right) \leqslant 1, \tag{19}$$

$$(1 - V_L)^2 \leqslant 2V_\Delta, \tag{20}$$

$$\rho \leqslant 1/2, \text{ or } \rho > 1/2 \text{ and } e \geqslant 2 - 1/\rho. \tag{21}$$

As to the single-crossing condition, it takes the form.

$$e + \mu_1(1 - \rho - \log \rho) < 2\mu_1 \frac{1}{1 - V_L}. \tag{22}$$

Let us first show that when it holds, then

$$A^{DE}(\mu_1) \leqslant e + \mu_1(1 - \rho - \log \rho) < A^{BDM}(\mu_1).$$

If $\mu_1 \leqslant e/\rho$, this follows from the definition of A^{DE}. If $\mu_1 > e/\rho$, then note that

$$A'^{DE}(\mu) = \begin{cases} 1 - \rho - \log \rho, & \text{if } \mu < e/\rho \\ 1 + \log\left(\frac{e}{\mu} + 1 - \rho\right) - \log \rho - \frac{e}{e + \mu(1-\rho)}, & \text{if } \mu \geqslant e/\rho, \end{cases},$$

and that these left- and right-derivatives coincide at e/ρ. Moreover, for $\mu > e/\rho$ we have $A''^{DE} = -e^2/\left[\mu(e + \mu(1 - \rho))^2\right] < 0$, and therefore $A^{DE}(\mu) < e + \mu(1 - \rho - \log \rho)$.

Next, note that (22) implies that $e + \mu(1 - \rho - \log \rho)$ and $A^{BDM}(\mu)$ cross exactly once for $\mu \in [0, \mu_1]$, since the first function is linear and the second concave, and we know that for $\mu < \mu_0$, $A^{DE} > A^{BDM}$. By the previous bounding argument, it follows that $A^{DE}(\mu)$ and $A^{BDM}(\mu)$ also cross exactly once in this region. Finally, (22) also implies that

$$\frac{1 - V_L}{2}(1 - \rho - \log \rho) < 1 - \frac{V_\Delta e}{1 - V_L} < 1,$$

so $A'^{BDM}(\mu) > 1 - \rho - \log\rho \geqslant A'^{DE}(\mu)$, for $\mu > \mu_1$. Therefore, $A^{BDM}(\mu) > A^{DE}(\mu)$ for all $\mu > \mu_1$. Thus, condition (22) indeed guarantees that A^{DE} and A^{BDM} (and so $\bar{a}^{DE}$ and $\bar{a}^{BDM}$) cross exactly once, and we can find μ^* such that $\bar{a}^{BDM} > \bar{a}^{DE}$ iff $\mu > \mu^*$.‖

5. Compatibility of the four conditions (for $\rho \leqslant 1/2$). We now verify that the intersection of (19)–(19) and the sorting condition (22) define a nonempty region region of parameters.

First, we can always find $0 < \rho \leqslant 1/2$ (ensuring (19)), $e \in [0,1]$ and $V_L \in [0,1]$ such that

$$e - \rho + \frac{1}{\rho} < \frac{2}{1 - V_L}.$$

Then, take $x \in (e - \rho + 1/\rho, 2/(1 - V_L))$ and let $V_\Delta = (1 - V_L)/x$. Condition (20) then holds, since

$$\frac{(1 - V_L)^2}{2V_\Delta} = x\frac{1 - V_L}{2} < 1,$$

and similarly for (19), since

$$e + \frac{1 - \rho}{\rho} < e - \rho + \frac{1}{\rho} < x = \frac{1 - V_L}{V_\Delta}.$$

The single-crossing condition, finally, requires that

$$e + \mu_1(1 - \rho - \log\rho) = e + \frac{(1 - V_L)^2}{2V_\Delta}(1 - \rho - \log\rho) < e + 1 - \rho - \log\rho.$$

But $\log\rho > 1 - 1/\rho$, so

$$e + \mu_1(1 - \rho - \log\rho) < e - \rho + \frac{1}{\rho} < x = \frac{1 - V_L}{V_\Delta} = \frac{2\mu_1}{1 - V_L}.$$

Therefore, all conditions can be simultaneously satisfied.

6. Extending the results to $\mu > 1/\rho$. Note that in this case in may be that $c_L^{DE}(\mu) > \underline{c}_H^{DE}(\mu)$. We will show that (22) alone guarantees single crossing. First, we establish that, for all μ :

$$A^{DE}(\mu) \leqslant e + \mu(1 - \rho - \log\rho). \tag{23}$$

Defining μ_H and μ_L, respectively, by $c_H^{DE}(\mu_H) \equiv c_{\max}$ and $\underline{c}_L^{DE}(\mu_L) \equiv c_{\max}$, there are two cases to consider.

(a) Case $\mu_H < \mu_L$. The function $A^{DE}(\mu)$ is given by:

$$A^{DE}(\mu) = \begin{cases} e + \mu(1 - \rho - \log\rho), & \text{if } \mu \leqslant e/\rho \\ \mu\left(1 + \log\left(\frac{e}{\mu} + 1 - \rho\right) - \log\rho\right), & \text{if } e/\rho < \mu \leqslant \mu_H \\ \mu\left(1 + \log\left(\frac{1-V_L}{\rho V_\Delta}\right) - \log\mu\right), & \text{if } \mu_H < \mu \leqslant \mu_L \\ \frac{1-V_L}{\rho V_\Delta}, & \text{if } \mu > \mu_L \end{cases},$$

with first derivative

$$A'^{DE}(\mu) = \begin{cases} 1-\rho-\log\rho, & \text{if } \mu \leqslant e/\rho \\ 1+\log\left(\frac{e}{\mu}+1-\rho\right)-\log\rho-\frac{e}{e+\mu(1-\rho)}, & \text{if } e/\rho < \mu \leqslant \mu_H \\ \log\left(\frac{1-V_L}{\rho V_\Delta}\right)-\log\mu, & \text{if } \mu_H < \mu \leqslant \mu_L \\ 0, & \text{if } \mu > \mu_L \end{cases}$$

Note that $A^{DE}(\mu)$ is continuous, but $A'^{DE}(\mu)$ is discontinuous at μ_H. In particular, as long as $v_H e < c_{\max}$,

$$\lim_{\mu \nearrow \mu_H} A'^{DE}(\mu) > \lim_{\mu \searrow \mu_H} A'^{DE}(\mu).$$

Also, it is still the case that $A''^{DE}(\mu) \leqslant 0$. Combining everything, we can see that (23) indeed holds for all μ.

(b) Case $\mu_L < \mu_H$. Define $\bar{\mu}$ by $\underline{c}_L^{DE}(\bar{\mu}) \equiv \underline{c}_H^{DE}(\bar{\mu})$, which implies that $\bar{\mu} < \mu_L < \mu_H$. Note that this requires that $\rho > 1/2$.

The function $A^{DE}(\mu)$ is given by:

$$A^{DE}(\mu) = \begin{cases} e+\mu(1-\rho-\log\rho), & \text{if } \mu \leqslant e/\rho \\ \mu\left(1+\log\left(\frac{e}{\mu}+1-\rho\right)-\log\rho\right), & \text{if } e/\rho < \mu \leqslant \bar{\mu} \\ \frac{e+\mu(1-\rho)}{\rho}, & \text{if } \bar{\mu} < \mu \leqslant \mu_H \\ \frac{1-V_L}{\rho V_\Delta}, & \text{if } \mu > \mu_H \end{cases},$$

with first derivative:

$$A'^{DE}(\mu) = \begin{cases} 1-\rho-\log\rho, & \text{if } \mu \leqslant e/\rho \\ 1+\log\left(\frac{e}{\mu}+1-\rho\right)-\log\rho-\frac{e}{e+\mu(1-\rho)}, & \text{if } e/\rho < \mu \leqslant \bar{\mu} \\ \frac{1-\rho}{\rho}, & \text{if } \bar{\mu} < \mu \leqslant \mu_H \\ 0, & \text{if } \mu > \mu_H \end{cases}$$

Thus $A''^{DE}(\mu) \leqslant 0$ and $A'^{DE}(\mu)$ is continuous at $\bar{\mu}$, so again (23) holds for all μ.

Finally, recall that we showed (22) to imply that $e+\mu_1(1-\rho-\log\rho) < A^{BDM}(\mu_1)$, which in turn ensures that $A'^{BDM}(\mu) > 1-\rho-\log\rho$ for $\mu > \mu_1$. Therefore

$$A^{DE}(\mu) \leqslant e+\mu(1-\rho-\log\rho) < A^{BDM}(\mu), \text{ for all } \mu > \mu_1,$$

ensuring again that A^{DE} and A^{BDM} cross at a single point, which lies in $[0, \mu_1]$.■

References

Alexander, L., and M. Moore (2015) "Deontological Ethics," *The Stanford Encyclopedia of Philosophy*, Edward N. Zalta (ed.). Web link.

Ambuehl, S. (2017) "An Offer You Can't Refuse? Incentives Change What We Believe," University of Toronto mimeo, October.

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

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

Ariely, D., Bracha, A., and S. Meier (2009) "Doing Good or Doing Well? Image Motivation and Monetary Incentives in Behaving Prosocially," *American Economic Review*, 99(1): 544–555.

Bénabou, R., Falk, A. and J. Tirole (2018) "Narratives, Imperatives, and Moral Reasoning," NBER Working Paper 24798, July.

_____ (2006) "Incentives and Prosocial Behavior," *American Economic Review*, 96(5): 1652–1678.

_____ (2011) "Identity, Morals and Taboos: Beliefs as Assets," *Quarterly Journal of Economics*, 126(2): 805–855.

Bentham, J. (1789) *An Introduction to the Principles of Morals.* London: Athlone.

Brekke, K. A., S. Kverndokk, and K. Nyborg (2003) "An Economic Model of Moral Motivation," *Journal of Public Economics,* 87 (9-10), 1967-1983.

Brandts, J. and G. Charness (2011). "The Strategy Versus the Direct-Response Method: A First Survey of Experimental Comparisons," *Experimental Economics,* 14(3): 375-398.

Brock, J. M., Lange A., and E. Y. Ozbay (2013), "Dictating the Risk: Experimental Evidence on Giving in Risky Environments, " *American Economic Review*, 103(1): 415–437.

Charness, G., U. Gneezy, and B. Halladay (2016) "Experimental Methods: Pay One or Pay All," *Journal of Economic Behavior and Organization,* 131: 141-150.

Chen, D. L. and M. Schonger (2013) "Social Preferences or Sacred Values? Theory and Evidence of Deontological Motivations," Discussion Paper, ETH Zurich.

Chen, D. L. and M. Schonger (2015) "A Theory of Experiments: Invariance of Equilibrium to the Strategy Method and Implications for Social Preferences," mimeo.

Clot, S., G. Grolleau, and L. Ibanez (2018). "Shall We Pay All? An Experimental Test of Random Incentivized Systems," *Journal of Behavioral and Experimental Economics*, forthcoming.

DellaVigna, S., List, J. and U. Malmendier (2012) "Testing for Altruism and Social Pressure in Charitable Giving," *Quarterly Journal of Economics,* 127: 1-56.

Elias, J., Lacetera, N., and M. Macis (2016) "Efficiency-Morality Trade-Offs In Repugnant Transactions: A Choice Experiment," NBER W.P. 22632. September.

Ellingsen, T., and M. Johannesson (2008) "Pride and Prejudice: The Human Side of Incentive Theory," *American Economic Review*, 98(3): 990–1008.

Everett, J. A. C., Pizarro, D., and M. J. Crockett (2016) "Inference of Trustworthiness from Intuitive Moral Judgments," *Journal of Experimental Psychology: General,* 145(6):772-787.

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

Falk, A. (2017) "In Face of Yourself - A Note on Self-Image," mimeo.

Falk, A. and N. Szech (2013) "Morals and Markets," *Science*, 340: 707–711.

_____ (2017) "Diffusion of Being Pivotal and Immoral Outcomes," Discussion Paper, University of Bonn.

Feddersen, T., Gailmard, S. and A. Sandroni (2009) "Moral Bias in Large Elections: Theory and Experimental Evidence," *American Political Science Review*, 103(2): 175-192.

Gino, F., Norton, M. and R. Weber (2016) "Motivated Bayesians: Feeling Moral While Acting Egoistically," *Journal of Economic Perspectives*, 30(3), Summer, 189-212.

Gneezy, U., Keenan, E. A., and A. Gneezy (2014) "Avoiding Overhead Aversion in Charity," *Science*, 346(6209): 632–635.

Goeree, J. K., Holt, C. A., and S. K. Laury (2002) "Private Costs and Public Benefits: Unraveling the Effects of Altruism and Noisy Behavior," *Journal of Public Economics*, 83(2): 255–276.

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

Johnson, R. (2014) "Kant's Moral Philosophy," *The Stanford Encyclopedia of Philosophy*, Edward N. Zalta (ed.). Web link.

Kagel, J. H. and A. E. Roth (1995) *The Handbook of Experimental Economics.* Princeton, Princeton University Press.

Kant, I. (1785) *Grundlegung zur Metaphysik der Sitten.*

Mazar N., Amir O. and D. Ariely (2008) "The Dishonesty of Honest People: A Theory of Self-Concept Maintenance," *Journal of Marketing Research*, XLV: 633–644.

Mill, J. S. (2002) *Utilitarianism: edited with an introduction by Roger Crisp*, New York: Oxford University Press, Originally published in 1861.

Roemer, J. (2010) "Kantian Equilibrium," *Scandinavian Journal of Economics,* 112(1): 1-24,

Roth, A. E. (2007) "Repugnance as a Constraint on Markets," *Journal of Economic Perspectives*, 21(3): 37–58.

Sandel, M. (2012) *What Money Can't Buy: The Moral Limits of Markets.* New York, NY: Farrar, Strauss and Giroux.

Sinnott-Armstrong, W. (2003) "Consequentialism," *Stanford Encyclopedia of Philosophy,* Edward N. Zalta (ed.). Web link.