Eliciting Moral Preferences: Theory and Experiment

Roland Bénabou, Armin Falk, Luca Henkel and Jean Tirole

Paper available via web page

ESA World Meeting – September 2020

Introduction

- What can be learned about a person's or a population's moral preferences from observing their choices, including in experiments?
  - How should this be used to inform policy, or to maximize prosocial actions?
  - How to interpret behaviors that seem deontologically rather than consequentially motivated: refusing tradeoffs that involve harm to others, assigning infinite price to "sacred values" such as life, freedom, dignity?
- Show how, whenever image concerns are present, the answers depend crucially on how choices are elicited / contributions sollicited:
  - Single (or, separate) decisions, vs. multiple simultaneous decisions; e.g., yes/no to an offer, versus stating a willingness to pay
  - Ex-ante commitments under uncertainty, vs. known, ex-post choices; e.g., random realized situation, random implementation

Concrete Setting and Application

- Use model & experiment to study and compare properties of two most commonly used revealed-preference methods:
  - Direct elicitation (*DE*)
  - Multiple-price list (*MPL*)
- Compared to *DE*, *MPL* features multiple decisions, of which only one is implemented for real, at random
- In standard situations (e.g., for non-moral decisions), we know both schemes give the same, and correct, answer. For instance:
  - Ask people in a population to make a DE choice, each one at different price
  - Ask each person the same MPL choice question

$\Rightarrow$ Get same distribution of outcomes, estimate same distribution of preferences

Key Results

1. As soon as image concerns are present, *DE* and *MPL* give different answers
   - Unrelated to risk aversion: tradeoffs now differ in expected value
2. Gap between results varies with the importance of image concerns (interaction), not just in magnitude, but even in sign! At any given price:
   - *DE* will generate more prosocial behavior than *MPL* when image concerns are weak (but positive)
   - *MPL* will generate more prosocial behavior than *DE* when image concerns are strong
3. Image-minded consequentialists will display Kantian-like price insensitivity much more readily under MPL than under DE
4. Results due interplay of three general effects, also at work in public-goods contributions mechanisms sharing key features with DE / MPL.
   - *Discouragement effect, cheap-talk effect, and cheap-act effect*
5. Model's most distinctive prediction: "crossing pattern" between *DE* and *MPL* contributions, as image concern go from weak to strong
6. Test it in an experiment: taking money or "Saving a Life"

Model

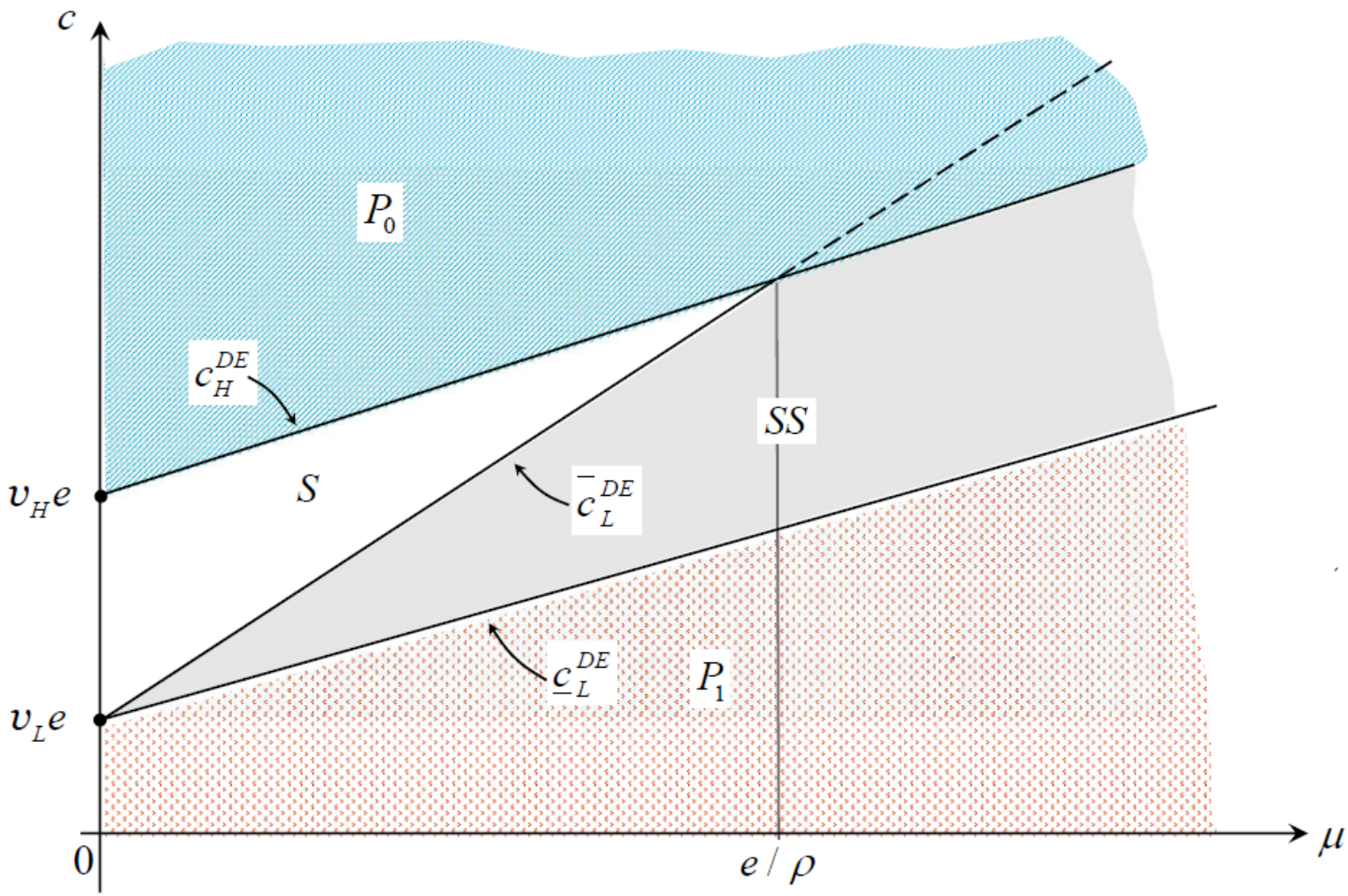
- Choice: engage in moral behavior ($a = 1$) or act selfishly ($a = 0$)
  - $a = 1$ involves personal cost c but generates positive externality e
- Agents differ in their motivation to act morally:
  - High type $v_H e$, with prob ρ, Low type $v_L e$, with prob $1 - \rho$; $v_H > v_L \geq 0$
- Final utility for type $\tau = L, H$:

$$U_\tau(a) = (v_\tau e - c)a + \mu E\left[v | a, \text{ choice conditions}\right]$$

- $\mu \geq 0$: strength of self or/and social image concerns. Image / esteem based on agent's expected type, conditional on action a and choice conditions
- Situation, even experiment, is now a signaling game $\Rightarrow$ behavior reflects not just individual preferences, but equilbrium
  - Pareto dominance as selection criteria in case of multiple equilibria

Behavior under Direct Elicitation

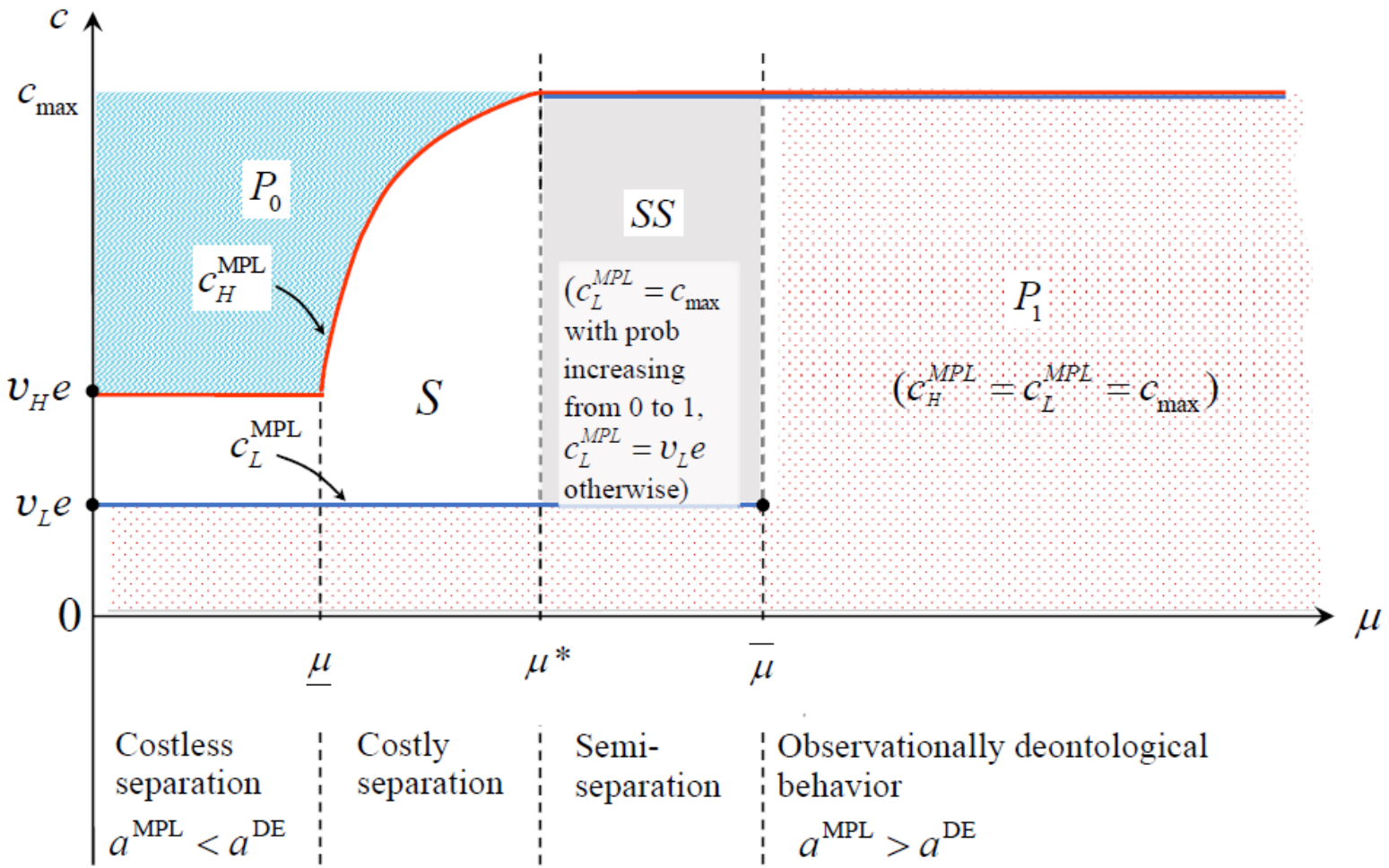
- Agents face choice $a \in \{0, 1\}$, for given value $c \in [0, c_{\max}]$



- 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 $a_H = a_L = 1$.

Behavior under Multiple-Price List

- Agents state maximum level of c to take $a = 1$ (WTP)
- Actual $\tilde{c}$ drawn from $G(\tilde{c})$ on $[0, c_{\max})$, implement $a = 1$ at cost $\tilde{c}$ iff $\tilde{c} \leq c$



- P_0: pooling at $a_H = a_L = 0$; S: separation, $a_H = 1$, $a_L = 0$; SS: semi-separation: $a_H = 1, a_L \in (0, 1)$; iP_1: pooling at $a_H = a_L = 1$.

Intuition: Three Key Effects from DE to MPL I

1. **Discouragement effect:** because it reveals multiple decisions at the same time, *MPL* raises the cost to the Low type of mimicking the High type:
   - Say, $v_L e = 50$, $v_H e = 75$. Low type might be willing to pool at *DE* price of $c = 60$, but under *MPL* would have to be willing to pool up to 75.
   - If μ is positive but not very large, not worth it $\Rightarrow$ will simply state WTP of 50, and thus does not contribute when $\tilde{c} = 60$ is drawn
   - This effect dominates at low $\mu > 0 \Rightarrow$ *DE* induces more prosocial decisions than *MPL*

2. **Cheap-talk effect:** under DE, if say yes to c, then $(e, -c)$ occurs for sure. Under *MPL*, if state WTP of c there is a probability $1 - G(c)$ that won't be "called on it", and neither e nor $-c$ occurs.
   - This effect tends to induces more prosocial decisions (especially by the low type) under *MPL*, relative to *DE*
   - However, as μ rises it weakens and ultimately vanishes, as the cutoff c_τ^{MPL} rises toward $c_{\max}$, driving the probability of implementation toward 1.

Intuition: Three Key Effects from DE to MPL II

3. **Cheap-act effect:** under DE, if say yes to c, then I pay c for e. Under *MPL*, if state WTP of c and am "called on it", will pay some random $\tilde{c} \leq c$.

   Because $c - E_G[\tilde{c}|c \leq \tilde{c}] > 0$, this effect also tends to induces more prosocial decisions under *MPL*, relative to *DE*.

   - Moreover, for the experimentally standard uniform distribution, and more generally for any distribution $G(\tilde{c})$ satisfying *MLRP*, the previous difference increases with c.
   - Therefore, as μ rises , pushing up all cutoffs, this effect strengthens. It is thus the one that dominates at high μ.

- Intermediate $\mu's$: all three effects operate, not much can be said in general
  - Paper derives a sufficient condition for single crossing of aggregate contributions under *DE* vs. *MPL* in the case of uniform G.

Main Result: Comparing DE and MPL

Proposition (interactions and reversal)

For each type (hence also on average):

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

2. *For all* $\mu \in (0, \underline{\mu})$,
$$\mathbf{a_\tau^{DE}(c, \mu) \geq a_\tau^{MPL}(c, \mu)},$$
with strict inequality for $c \in (v_L e, \underline{c}_L^{DE}(\mu))$ *and* $c \in (v_H e, c_H^{DE}(\mu))$, *both nonempty.*

3. *For all* $\mu \geq \bar{\mu}$,
$$\mathbf{a_\tau^{DE}(c, \mu) \leq a_\tau^{MPL}(c, \mu)},$$
with strict inequality for $c \in (\underline{c}_L^{DE}(\mu), c_{\max})$, *which is nonempty whenever* $\mu \in (\bar{\mu}, \mu^{**})$.

Empirical Tests

- **Hypothesis 1:** For both *DE* and *MPL*, total contributions increase in μ
- **Hypothesis 2:** For low $\mu_L > 0$, total contributions are higher under *DE* than under *MPL*
- **Hypothesis 3:** For high μ_H, total contributions are higher under *MPL* than under *DE*
- **Corrollaries:**
  - **Differential image sensitivity:** as μ changes from μ_L to μ_H, contributions rise by more under *MPL* than under *DE*
  - **Observationally deontological behavior:** at μ_H, more people will choose the moral action "whatever it costs" , i.e. up to the highest price $c_{\max}$ under *MPL*, than under *DE*. Different estimated fractions of "Kantians".

Experiment: Saving a Life

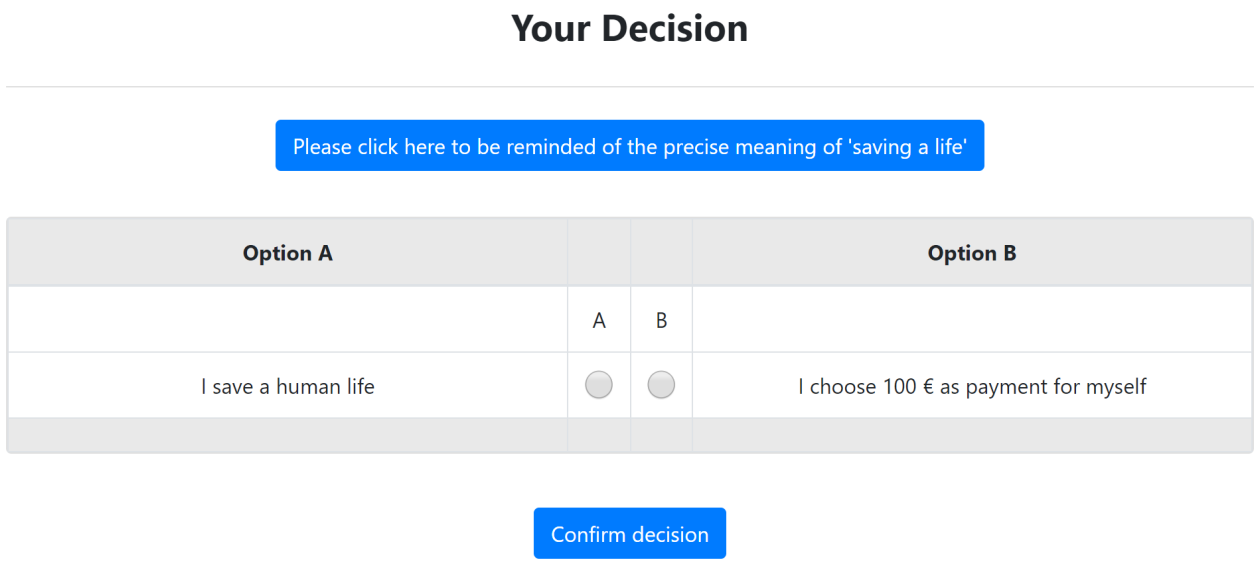
- **Choices:**
  - Moral action $(a = 1)$: induce a 350€ donation that, in expectation / on average, will save one patient from death by tuberculosis. Major $e >> 0$
  - Selfish action $(a = 0)$: take money for oneself.

    Amount c, can range from 10 to 200€

- **High stakes:**
  - Subjects provided with detailed, verifiable (on site) evidence of death risk for tuberculosis patients in India, effectiveness of treatment, track record of NGO doing it (Operation ASHA), expected value calculation

- **Treatments:** 2 x 2 between-subjects design, varying both:
  - Elicitation method: Direct elicitation (*DE*) vs. Multiple-price list (*MPL*)
  - Level of image concerns, μ : choices kept private (*Low Image*), or made publicly visible & morally salient (*High Image*)

Decision Screens



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

Manipulating Moral-Image Concerns

- To ensure some minimal social and self-image $\mu > 0$, subjects are anonymously paired, will learn partner's choices (benchmarking). Then:

▶ **Low Image** (μ_L) :

- Experiment is double blind. Use procedure of Barmettler, Fehr, and Zehnder (2012): one subject carries out final payment, without participating in experiment. Self-image still presumably operating.

▶ **High Image** (μ_H) :

- Subject's choices are publicly observed and compared to those of their partners by a committee, upon receiving payment
  - Own and partner's choices projected on a wall with subject present, must read them aloud.
  - Committee of three, sitting in the room, evaluates morality of each choice
    - ★ Morality scores not disclosed, but really given, and subjects know that

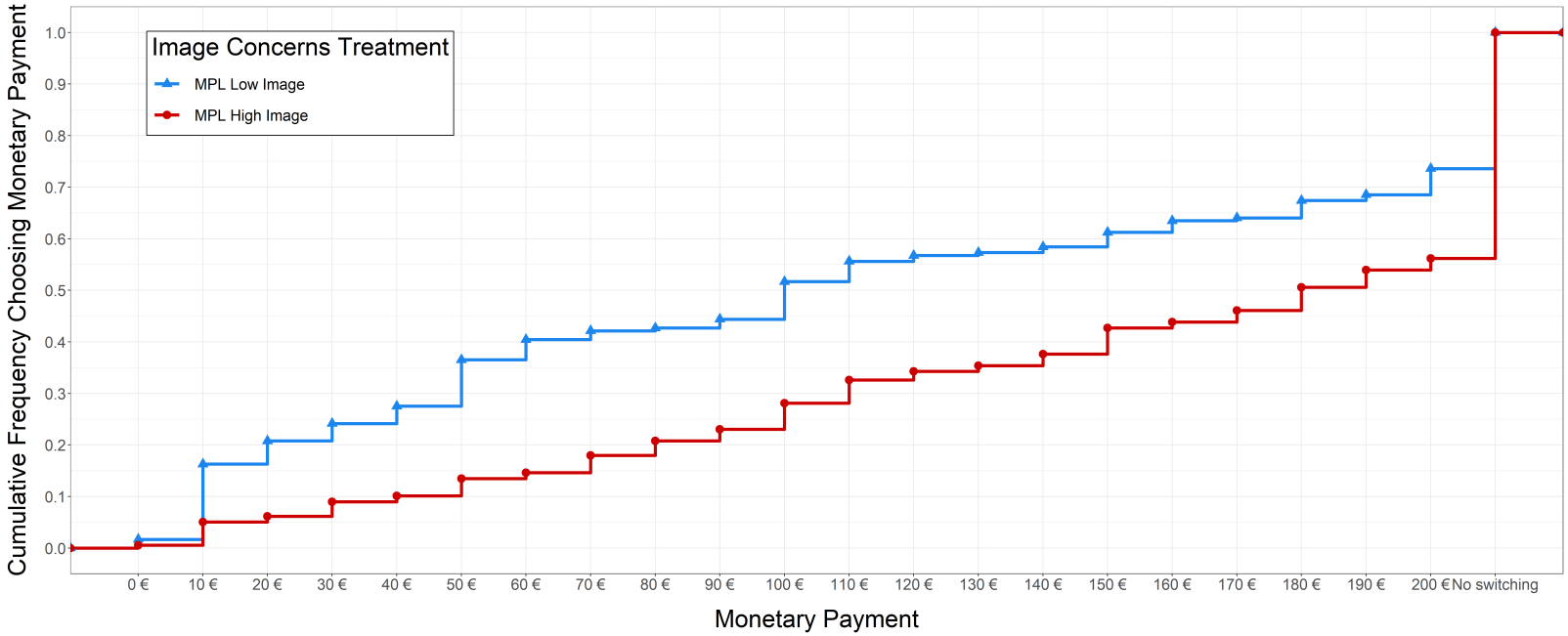
Procedure

- Bonn Lab: 697 subjects, mostly students, 58% female, mean age = 24.01
- 12€ show-up fee. Receive extensive background information on donation, decisions must pass comprehension test. For each session ($\approx$ 20 subjects) the decisions of one pair are implemented for real.
- Implement, under High and Low Image:
  - *DE* at (preset) price of $c = 100$€
  - *MPL* with $\tilde{c}$ uniform over $0, 10, ...c_{\max} = 200$€, in increments of 10€
- When comparing the two schemes, do so at same price level:
  - $\bar{a}^{DE}(100, \mu)$: fraction who save a life rather than take $c = 100$€, under *DE* (would then have done so under *DE* at any $c' \leq 100$)
  - $\bar{a}^{MPL}(100, \mu)$: fraction who state WTP ≥ 100€ under *MPL*, and thus commit to saving a life at any $\tilde{c} \leq 100$ that may be drawn

Hypothesis 1: direct effect of image

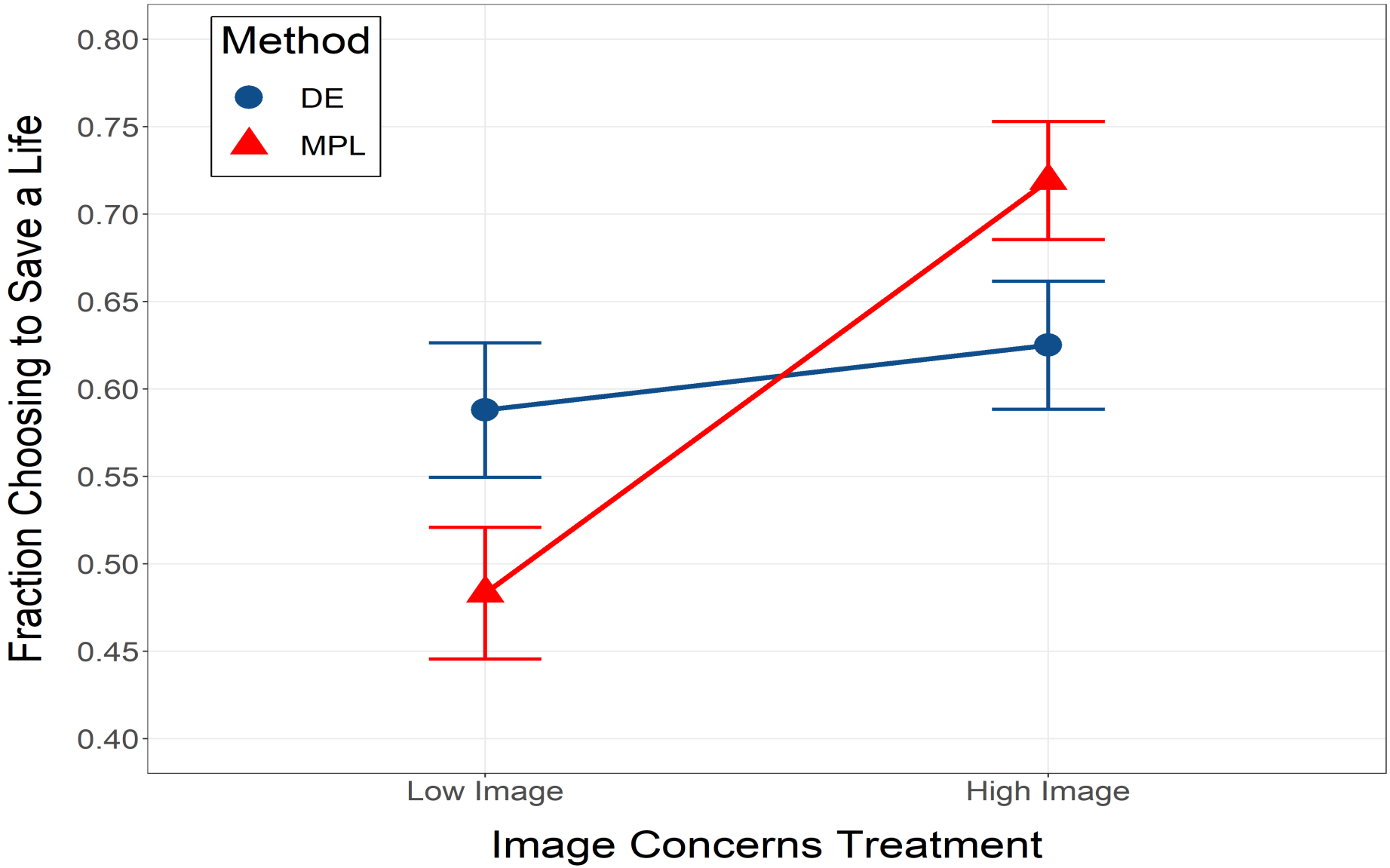
- **DE:** 58.8% of subjects choose to save a life (vs. 100€) in *Low Image*, and 62.5% in *High Image*. But difference not significant

- **MPL:**



  - CDF from Low Image lies above that from MPL-High Image, for all monetary payments ($p < 0.001$, Kolmogorov–Smirnov test). Difference $> 15\%$ for almost all payments; largest at 60€, of 26%.
  - Obs. deontological: 26.4% under μ_L,nearly doubles to 48.4% under μ_H!

Hypotheses 2 and 3: interaction and reversal



Hypotheses 2 + 3: differential image sensitivity

Dependent variable:	Choice to Save a Life (vs. 100€)			
	Low Image Concerns		*High Image Concerns*	
	(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
R^2	0.011	0.077	0.010	0.062

Robust standard errors in parentheses. Controls include age, gender, income, religiousness, educational level, and high school grade. Significance levels: $^{*}p<0.1$, $^{**}p<0.05$ and $^{***}p<0.01$.

Heterogeneity among subjects

- Use independent measure of altruism (validated in Falk et al. 2018):
  - "How willing are you to give to good causes without expecting anything in return?"
  - "Today you unexpectedly received 1,000€. How much of the money would you donate to a good cause?"

$\Rightarrow$ Median split

- Measure is correlated with "saving a life" decision, but independent of treatment

Heterogeneity among subjects

Dependent variable:	Choice to Save a Life (vs. 100€)							
	Below-median Altruism				Above-median Altruism			
	Low Image		High Image		Low Image		High Image	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
MPL	−0.187**	−0.187**	0.040	0.030	−0.032	−0.007	0.118*	0.138**
	(0.075)	(0.078)	(0.075)	(0.079)	(0.073)	(0.072)	(0.068)	(0.066)
Constant (DE)	0.512***	0.611***	0.592***	0.586***	0.663***	0.591***	0.667***	0.647***
	(0.056)	(0.084)	(0.050)	(0.072)	(0.052)	(0.070)	(0.054)	(0.062)
Controls		X		X		X		X
Observations	342	342	342	342	355	355	355	355
R^2	0.036	0.133	0.002	0.035	0.001	0.101	0.017	0.109

Robust standard errors in parentheses. Controls include age, gender, income, religiousness, educational level, and high school grade. Significance levels: *p<0.1, **p<0.05 and ***p<0.01.

Conclusion

- Image concerns interact differently with different elicitations methods, solicitation schemes. The introduction of multiple decisions and random implementation give rise to three key effects:
  - *Discouragement effect:* multiple decisions (WTP) decrease contributions. Dominates at low μ
  - *Cheap-act effect:* random cost increases contributions. Dominates at high μ
  - Cheap-talk effect: operates in middle range.
- Experimental evidence: *DE* and *MPL* "crossing" in high-stakes experiment
- Implications:
  1. Caveat for measurement of moral preferences (and other reputation-bearing behaviors), whether one is interested in descriptive / predictive questions (how people behave, including from reputation-seeking), or normative ones (how much they truly value public goods, behaviors)
  2. Caveat about / upper bound on / estimating the proportion of "Kantians"
  3. Possible applications / extensions to other types of preference elicitation schemes, charitable-contributions solicitations, etc.