INCENTIVES AND PROSOCIAL BEHAVIOR

Roland Bénabou - Jean Tirole

Princeton University IDEI-Toulouse

American Economic Review (2006), forthcoming

MOTIVATING FACTS AND PUZZLES

- ✓ People vote, volunteer, give to charitable organizations, help strangers, join rescue squad, risk life,… Field + numerous experiments. [See, e.g. Fehr-Schmidt (2003) for survey]

- ✓ Some phenomena cannot be explained by sole presence of people with other-regarding preferences:

  (a) *Crowding-out effects*

  [Festinger-Carlsmith 59, Deci-Ryan 85, Gneezy-Rustichini 2000a,b, Frey 1997, Fehr and Götte 1999, Falk and Kösfeld 2004, Mellström-Johannesson (2005)]

  (b) *Social glory / shame. Imitation effects.*

  [Codes of honor, shame (Batson (1998)); conspicuous donations (only a few % anonymous)]. People contribute / cooperate more when observed (e.g, Funk (2005), Bandiera (2006), List (2005)] and when they know others are doing it (e.g., Bardsley-Sausgruber (2005))]

  (c) *Self-image concerns / information aversion*

  [Adam Smith 1776; Dana et al 2003a,b, Murnighan et al. 2001]
  [Kahneman-Knetsch 1992 “purchase of moral satisfaction”]

The Red Cross
on contributing, volunteering:

“You will be surprised at how good it makes you feel and what a terrific response you will get from loved ones”.

“Helping others feels good and makes you feel good about yourself”.

TWO MAIN THEMES / SETS OF QUESTIONS

- Theme I: why do people engage in prosocial / altruistic / reciprocal behaviors?
  - "pure" altruism: care for others, for public goods
  - "joy of giving", "warm glow"; self-image
  - social esteem or stigma; reputational concerns
  - explicit incentives: rewards, punishments
- Essential to take into account how multiple motives, varying across people, *interact* together in shaping behavior.
- Example: are legal sanctions substitutes or complements for social sanctions and norms?
- Theme II: How effective are *incentives*, whether *material* (rewards, punishments) or *image-related* (public honor and shame) at inducing desired behavior? [Econ vs. Psy. Reconcile?]

Two (informational) crowding-out mechanisms

	"Intrinsic and Extrinsic Motivation" (*RES* 03)	*"Incentives and Prosocial Behavior"*
Mechanism	Conveys bad news about nature of task, its payoffs, or individual's ability (principal's trust of the agent) *(informed principal)*	Sullies the meaning of good deeds by creating doubt as to true motivation (overjustification effet) *(multidimensional signaling)*
Examples	Rewarding child for task, reading. Closely monitoring agent...	Help, blood donation,…
Impact of rewards	• Limit immediate reinforcement. • Crowd out future re-engagement	Immediate
Related work	Souvorov (2004), Herold (2004), Souvorov - van de Ven (2005) Ellingsen-Johannesson (2006)…	[Bernheim (1994), Corneo(1997), Denrell (1998)] Seabright (2005)

Model of prosocial behavior that combines:

- ✓ Heterogeneity in individuals' degrees of altruism and greed.
- ✓ Concerns for social reputation and / or self-respect
- ➢ Analysis of how the *three motives for prosocial behavior*,

  intrinsic + extrinsic + (self) reputational

  interact and how this balance changes with power of incentives, disclosure of rewards, prominence of actions…
- ➢ *Information-based explanation for crowding out*

  Rewards or punishments spoil reputational value of good deeds by creating (self-) doubt as to the underlying motivation

  In line with the "overjustification effect" in psychology
- ➢ *Welfare analysis*: what is the socially optimal level of incentives? Will private sponsors (NGO's, charities, etc). provide too low / too high incentives, or the wrong kind.

Outline

- Model.
- Heuristics of image-spoiling effect of rewards.
- Crowding out: signal-extraction & the overjustification effect
- Social norms: what sustains them? When are individual decisions strategic complements or substitutes?
- Setting of incentives by public or private sponsors. Social optimum, sponsor competition.
- Conclusions

I. THE MODEL

- Agents choose their participation level a in some prosocial activity; may be discrete (0/1) or continuous.
- Contribute $a \Rightarrow$ *c*ost $C(a)$, monetary reward $y \cdot a$
- Incentive rate y may reflect proportional subsidy or tax on a.
- Variant: monetary donation $a \Rightarrow$ receive "perks" y per €

❑ *Direct benefits from pro (or anti) social activity:*

$$(v_a + v_y y)\, a - C(a)$$

intrinsic extrinsic / material

v_a: valuation of public good + "joy of giving" ⇒ *"altruism"*

v_y: valuation of money / private consumption ⇒ *"greed"*

Individual's preference type or "identity" $\boldsymbol{v} = (v_a, v_y)$ is drawn from continuous distribution; private information.

- *Social esteem / status or self-esteem: reputational concerns*

$$x\,[\gamma_a E(v_a|a,y) - \gamma_y E(v_y|a,y)]$$

wants to appear / see himself as *prosocial*

wants to appear / see himself as *disinterested* (or: wealthy)

probability that behavior is observed by others *(social signaling)* or used to assess "what kind of a person I am" *(self signaling).*

Also length of record, number of people who hear of it, memorability, etc.

Let $\mu_a \equiv x\,\gamma_a$ and $\mu_y \equiv x\,\gamma_y \Rightarrow$ agent chooses a to maximize:

$$U = (v_a + v_y y)a - C(a) + \mu_a E(v_a|a,y) - \mu_y E(v_y|a,y),$$

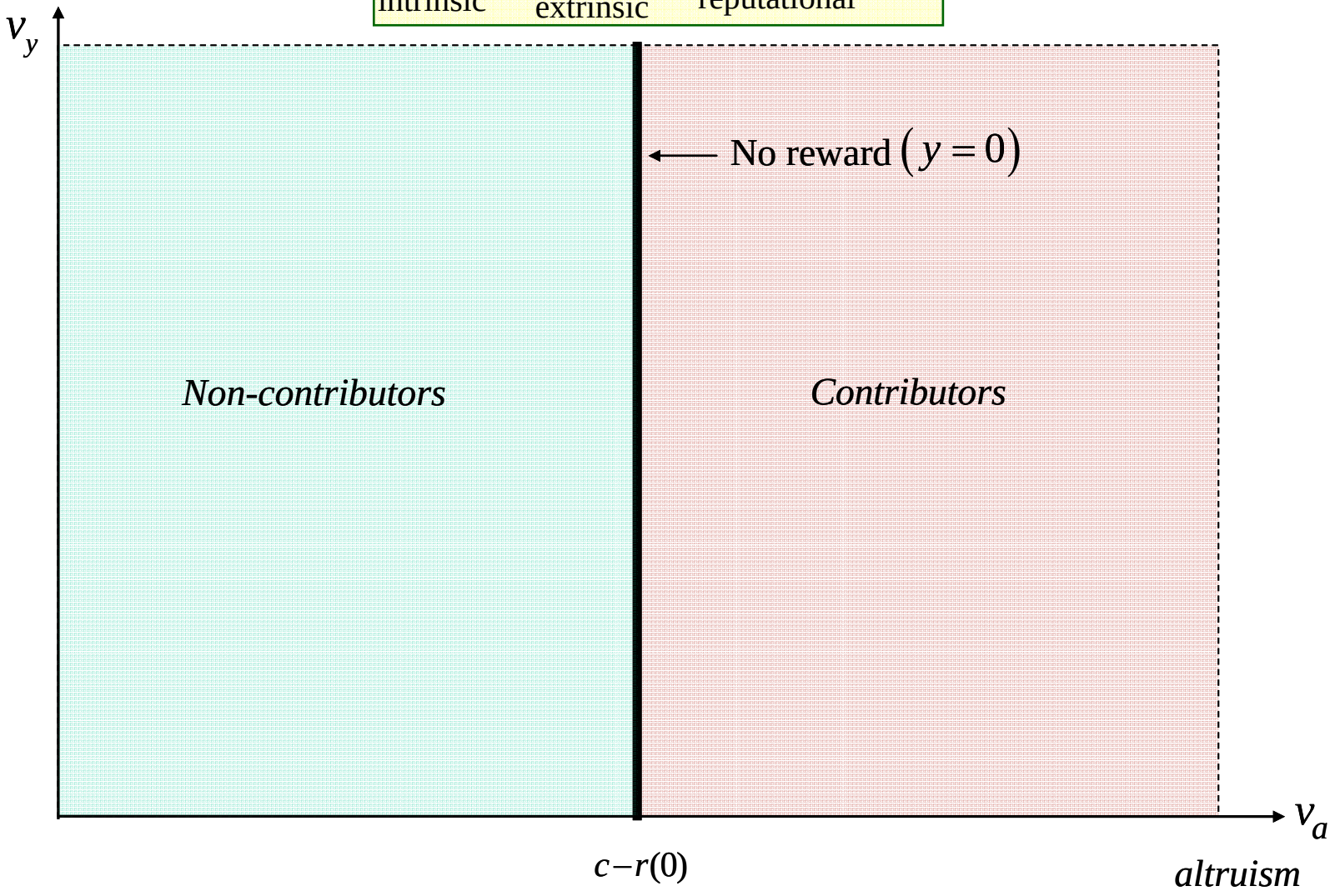
taking into account how his behavior will be *interpreted.*

Policy parameters: *material rewards*, y and *publicity*, x.

THE IMAGE-SPOILING EFFECT OF REWARDS: BASIC INTUITIONS

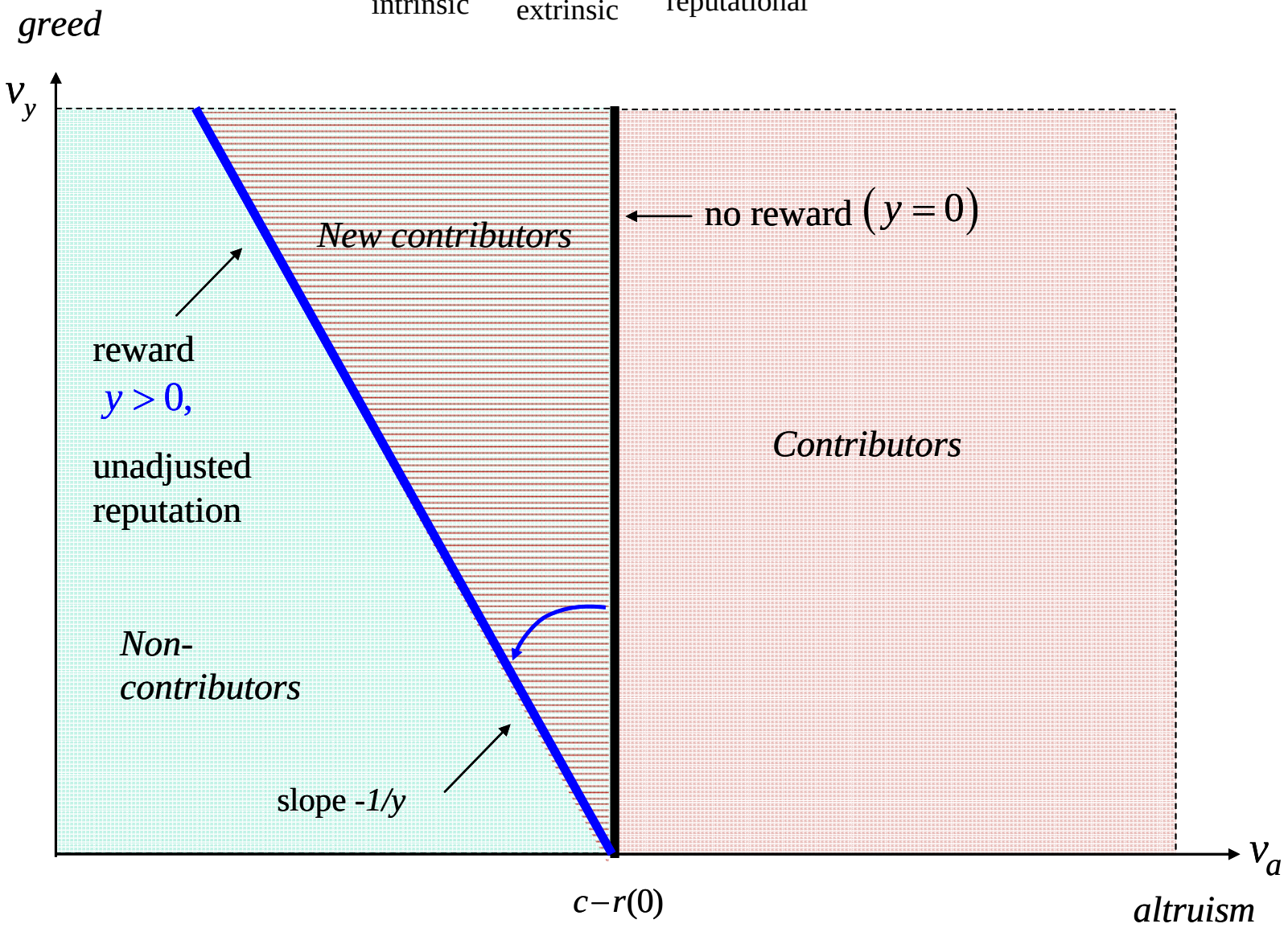
❑ Binary decision: $a = 0$ (cost = 0) or $a = 1$ (cost c) => participate if:

$$\underbrace{v_a}_{\text{intrinsic}} + \underbrace{v_y y}_{\text{extrinsic}} + \underbrace{r(y)}_{\text{reputational}} \geq c$$



❑ Introduce reward $y > 0$. First step: if reputation remains $r(0)$

$$\underbrace{v_a}_{\text{intrinsic}} + \underbrace{v_y y}_{\text{extrinsic}} + \underbrace{r(0)}_{\text{reputational}} \geq c$$

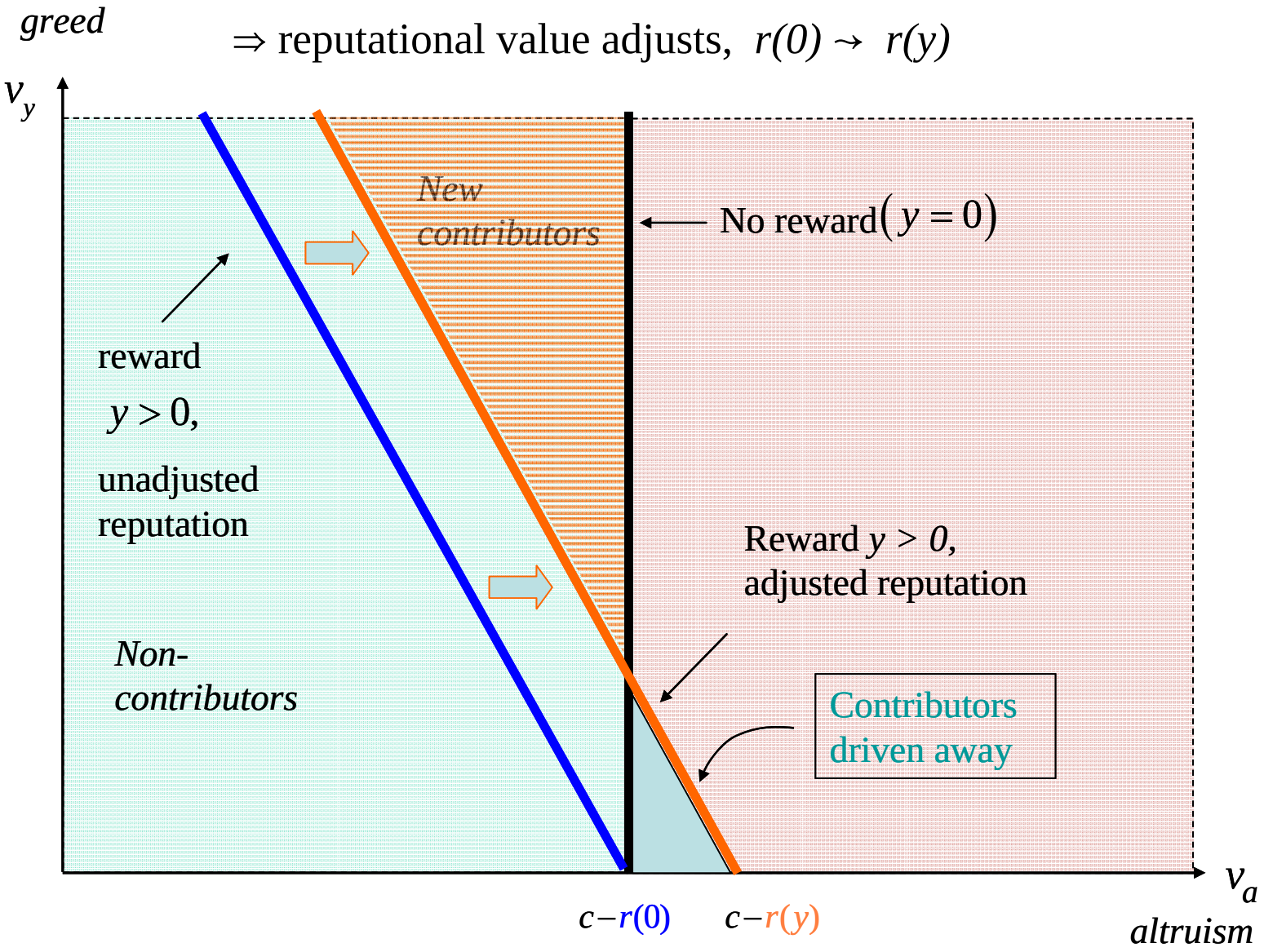


❑ But: compared to original contributors, new ones are:

(i) *less prosocial*; non-contributors also worse, however

(ii) *more greedy*

⇒ reputational value adjusts, $r(0) \rightsquigarrow r(y)$



II. THE OVERJUSTIFICATION EFFECT AND CROWDING OUT

- Continuous a, cost $C(a) = ka^2/2$.
- Both $\boldsymbol{v}$ and $\boldsymbol{\mu}$ may vary across individuals; private information.
- Normal distribution:

$$\mathbf{v} \equiv \begin{pmatrix} v_a \\ v_y \end{pmatrix} \sim N\left(\begin{pmatrix} \bar{v}_a \\ \bar{v}_y \end{pmatrix}, \begin{bmatrix} \sigma_a^2 & \sigma_{ay} \\ \sigma_{ay} & \sigma_y^2 \end{bmatrix} \right), \quad \bar{v}_a \gtrless 0, \quad \bar{v}_y > 0,$$

$$\boldsymbol{\mu} \equiv \begin{pmatrix} \mu_a \\ \mu_y \end{pmatrix} \sim N\left(\begin{pmatrix} \bar{\mu}_a \\ \bar{\mu}_y \end{pmatrix}, \begin{bmatrix} \omega_a^2 & \omega_{ay} \\ \omega_{ay} & \omega_y^2 \end{bmatrix} \right), \quad \bar{\mu}_a \geq 0, \quad \bar{\mu}_y \geq 0,$$

with $(\boldsymbol{v},\boldsymbol{\mu})$ independent. Optimal decision for agent with type $(\boldsymbol{v},\boldsymbol{\mu})$:

$$C'(a) = v_a + v_y y + r(a, y; \mu),$$

where

$$r(a, y; \mu) \equiv \mu_a \frac{\partial E(v_a \mid a; y)}{\partial a} - \mu_y \frac{\partial E(v_y \mid a; y)}{\partial a}.$$

External or internal observer's inference problem: from behavior a, knows only the *sum of the three sources of motivation.* Signal extraction.

MATERIAL REWARDS

Take identical reputational concerns (same$(\bar{\mu}_a, \bar{\mu}_y)$) for all agents

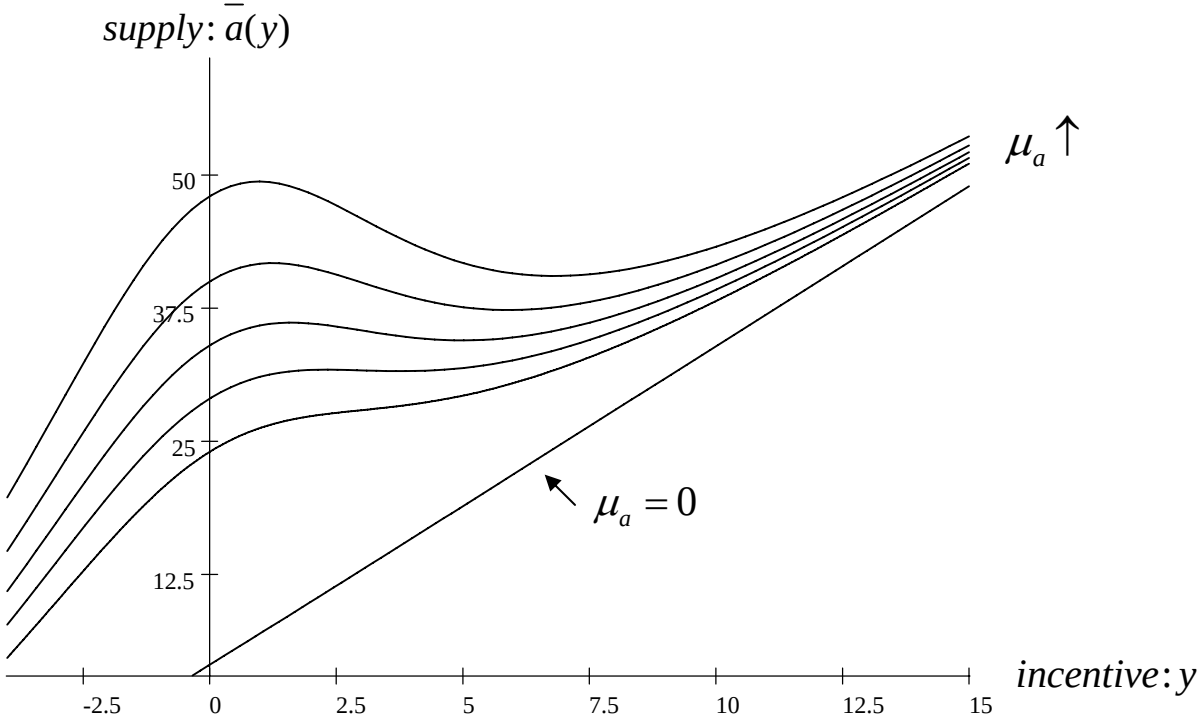
Proposition 1 *Equilibrium behavior (with* $\sigma_{ay} = 0$ *) is:*

$$a = \underbrace{\frac{v_a + v_y y}{k}}_{\text{direct incentive (intrinsic + extrinsic)}} + \underbrace{\bar{\mu}_a \left(\frac{1}{1 + y^2 \sigma_y^2 / \sigma_a^2} \right) - \bar{\mu}_y \left(\frac{y \sigma_y^2 / \sigma_a^2}{1 + y^2 \sigma_y^2 / \sigma_a^2} \right)}_{\text{reputational incentive } r(y)}$$

- Reward *y* has usual direct effect but also acts like an increase in signal-to-noise ratio, depressing *r(y)*. *Overjustification effect*.
- Multidimensional heterogeneity / signaling is key for the result.
- Aggregate supply: summing *a* across individuals $\Rightarrow \bar{a}(y)$.

Overjustification effect and crowding out

Proposition 2 *Let* $\sigma_{ay} = 0$. *When the concern for prosocial reputation* μ_a *is above some threshold* μ_a^*, *there exists a range* $[y_1, y_2]$ *over which incentives are counterproductive: a higher reward y reduces the total amount of prosocial behavior,* $\bar{a}(y)$



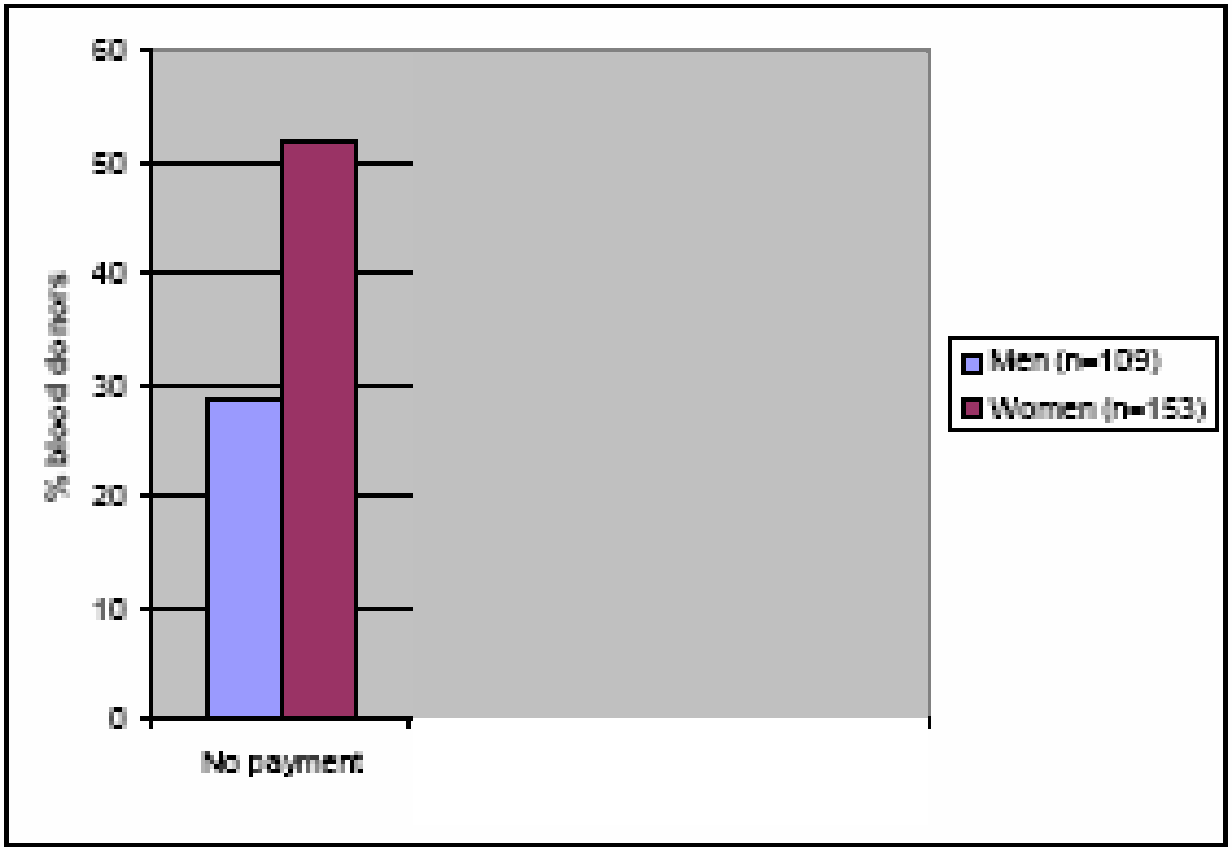


Figure 1. The supply of blood donors among men and women in the three experimental treatments.

Mellström and Johannesson (2005)

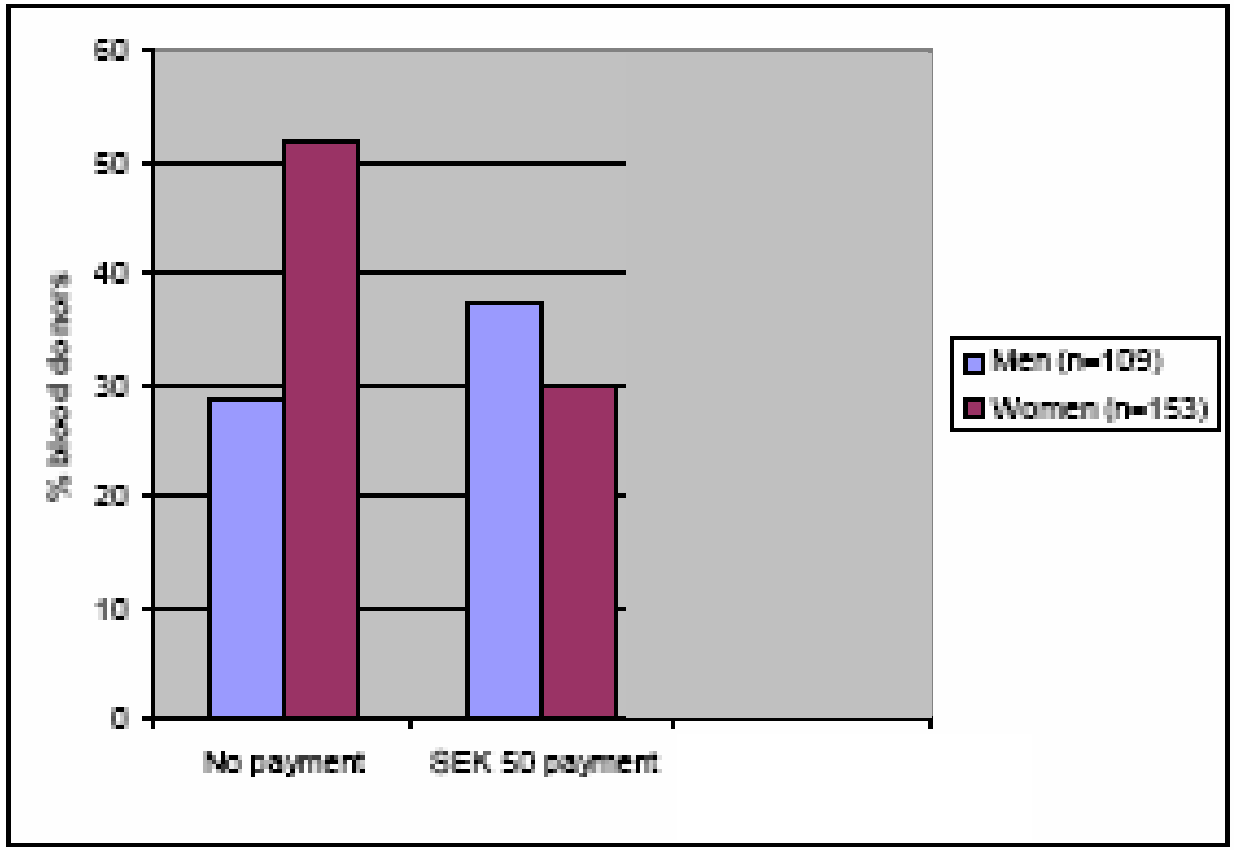


Figure 1. The supply of blood donors among men and women in the three experimental treatments.

Mellström and Johannesson (2005)

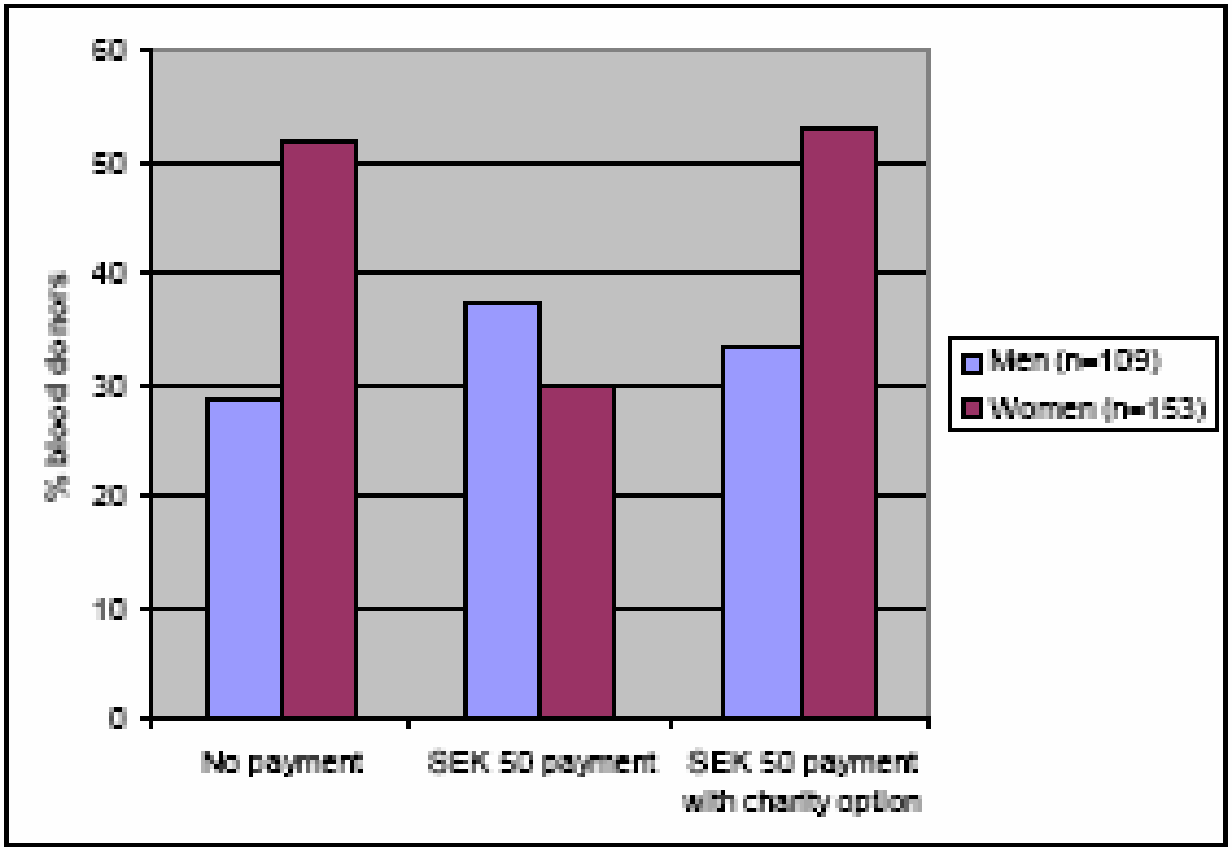


Figure 1. The supply of blood donors among men and women in the three experimental treatments.

Mellström and Johannesson (2005)

IMAGE REWARDS (praise, shame, etc.)

Agents now also differ in their image concerns (μ_a , μ_y): normally distributed

- Behavior is noisy measure of true preferences (v_a ,v_y), with "noise", $r(a,y;\mu)$, i.e., contribution of image motivation to behavior, now endogenous.

  Proposition 4 *An individual with preferences* (v_a ,v_y ; μ_a , μ_y) *contributes*

$$a = \frac{v_a + yv_y}{k} + \mu_a \rho(y) - \chi(y)\mu_y,$$

  with $\varrho(y)$ *and* $\chi(y)$ *now given by a fixed point equation.*

- Policies based on publicity, prominence, memorability

  [Medals, titles, named buildings, public praise and shame, televised arrests, e-registry, pillory…]

  Scale up the reputation weights (μ_a, μ_y) by a factor $x \Rightarrow$

$$C'(a) = v_a + v_y y + x \cdot r(a, y, x; \mu)$$

  - Direct impact: increases incentive to behave well
  - Dampening effect: as $x \uparrow$, *observers increasingly ascribe behavior to image concerns.* New form of overjustification effect.
  - Aggregate supply $\bar{a}(x)$ grows only as $x \cdot \bar{r}(a; y, x) \sim x^{1/3}$

III. HONOR, STIGMA, AND SOCIAL NORMS

- What makes a behavior socially or morally unacceptable is often the very fact that "it is just not done". But in other times, other places: "everyone does it".

  [choosing surrender over death, not going to church, not voting, divorce, welfare dependency, minor tax evasion, conspicuous modes of consumption,....]

- People contribute more when they know / see that others do
  [public goods, fundraising, voting; helping strangers, Salvation Army…]

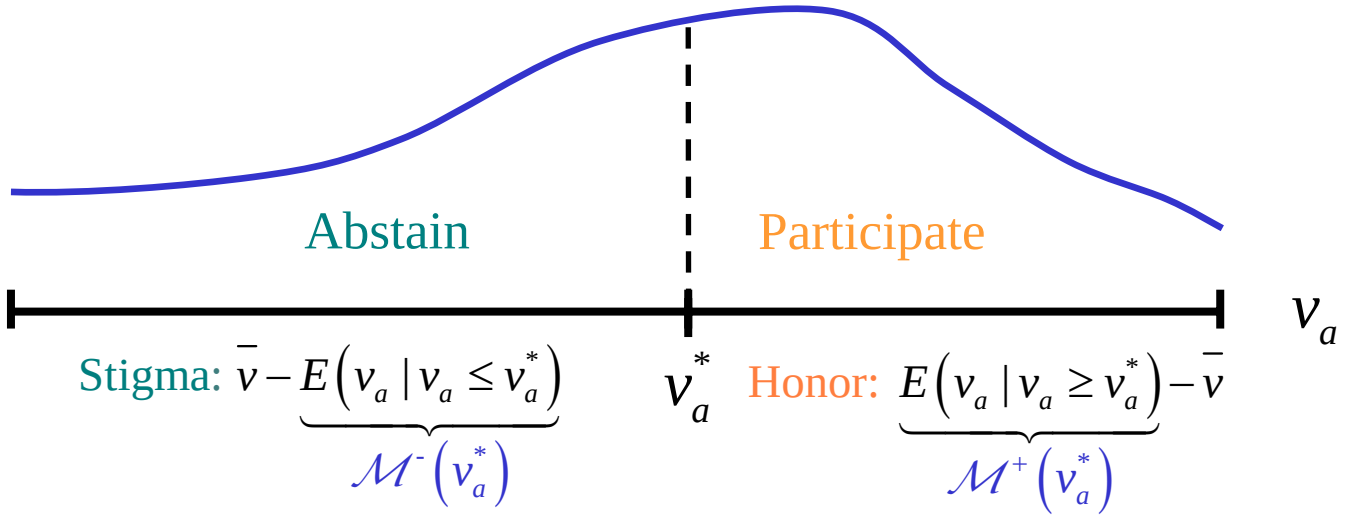
- Often explained / modeled by complementarity in preferences ($\bar{a}$ raises v_a): general "norm", untargeted "reciprocity".

- In fact: complementarities arise *endogenously* from the interplay of *honor and shame.*

- More generally: when does the fact that others contribute more *increase* or *decrease* the pressure (social, moral) on me to do so? (Strategic complements / substitutes). Policy implications.

Honor and Stigma

- Discrete decision, $a = 0, 1$. Focus on reputation for social orientation: v_a is unknown: distributed on $[v_a^-, v_a^+]$. Money orientation $v_y \equiv 1$ is known.

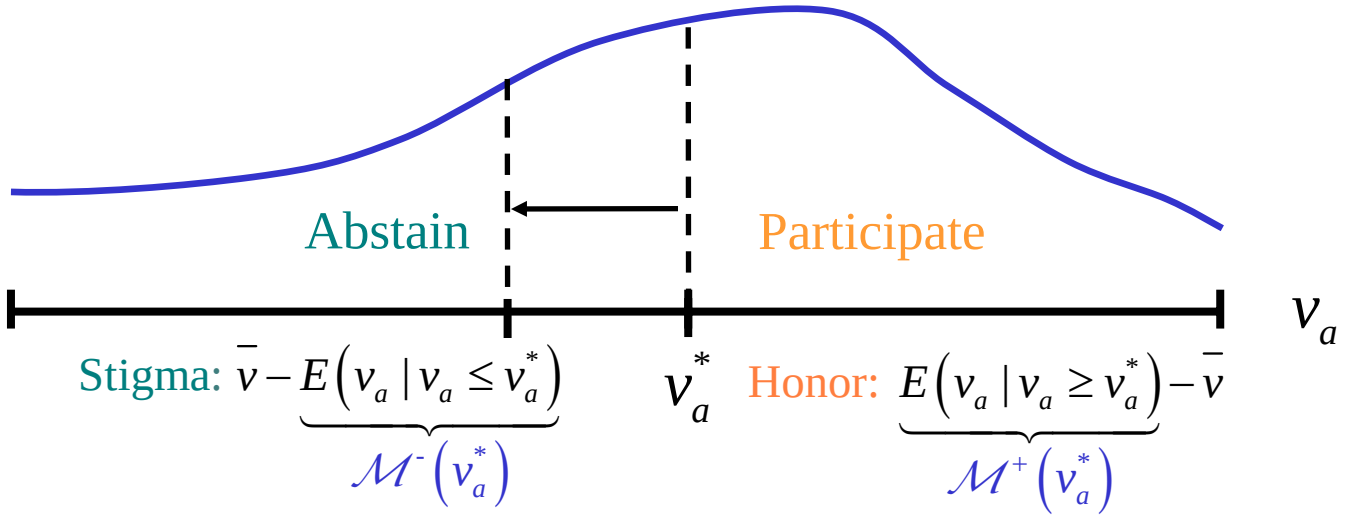


- Who participates? Those above the cutoff are those with

$$v_a + y + \mu_a \left[\mathcal{M}^+(v_a^*) - \mathcal{M}^-(v_a^*) \right] \ge c.$$

Honor and Stigma

- Discrete decision, $a = 0, 1$. Focus on reputation for social orientation: v_a is unknown: distributed on $[v_a^-, v_a^+]$. Money orientation $v_y \equiv 1$ is known.



- Who participates? Those above the cutoff are those with

$$v_a + y + \mu_a \left[\mathcal{M}^+(v_a^*) - \mathcal{M}^-(v_a^*) \right] \ge c.$$

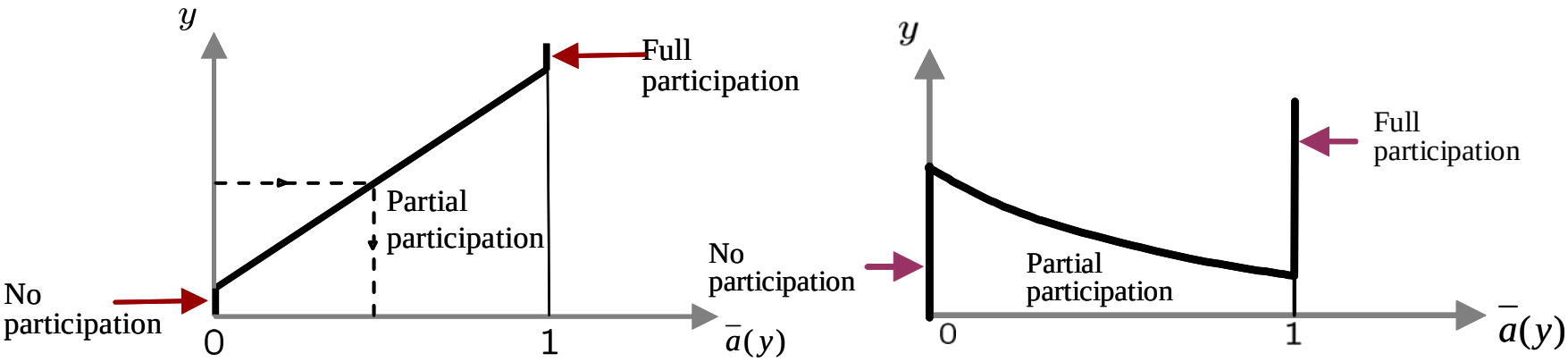
- When more people participate (cutoff v_a^* falls), social / moral pressure to participate decreases / increases depending on sign of $\left(\mathcal{M}^+ - \mathcal{M}^-\right)'$

  $\Rightarrow$ is concern for *honor* or that for *stigma* predominant (at the margin)?

Proposition 5 *If the function*

$$\Psi(v_a) \equiv v_a + \mu_a \left[\mathcal{M}^+(v_a) - \mathcal{M}^-(v_a) \right]$$

is increasing, the equilibrium is unique. If Ψ is decreasing, there is a range of rewards y under which multiple norms are self-sustaining.



- Actions are substitutes, or weak complements
- $\Delta y \rightarrow$ partial crowding out, or crowding in

- Actions are strong complements
- Multiple norms coexist
- Small $\Delta y \rightarrow$ large effects

$\Rightarrow$What features of "market" create complementarity/substitutability?

Intuition: factors that *accentuate stigma* or *dampen glory* facilitate the emergence of complementarity / imitation / endogenous social norms. Vice-versa for substitutability

Intuition: factors that *accentuate stigma* or *dampen glory* facilitate the emergence of complementarity / imitation / endogenous social norms. Vice-versa for substitutability

➢ Distribution of preferences: how many true altruists / egoists?

- Avoiding stigma = dominant concern when only a few "bad apples" with v_a far below that of "most people" (mode of the distribution). Example: serious crime.

- Pursuing honor = dominant concern when only a few "saintly" types with v_a well above that of most people. Examples: heroism, organ donation.

Proposition 6 (based on Jewitt (2004)). *If* v_a *has increasing density,* $\mathcal{M}^+ - \mathcal{M}^-$ *is decreasing (⇒ complements). Conversely,* $\mathcal{M}^+ - \mathcal{M}^-$ *is increasing if the density of* v_a *is decreasing (⇒ substitutes) .*

Applications / Related results / Extensions

- Material incentives (prizes, law) not very effective to spur "*admirable*" prosocial behaviors (honor driven): $y \nearrow$ weakens motivation for the upper tail (partial crowding out).

- More effective to strengthen "*respectable*" ones (stigma-driven): $y \nearrow$ strenghtens social pressure in the lower tail (crowding in). Implications for optimal incentives: see later.

- Small changes in incentives can have very big effects, shift social norms: Continental Airlines $50 bonus program based on *company-wide* overall performance for the month (Knez and Simester (2001)).

- *Nonlinear pricing*. Dur (2006): policy of zero tolerance for minor offences can have a "*double dividend*": it increases their signaling value for being tough (drives out the wimps) and may thus result in a decrease in serious crimes as well.

  [Model with $a = 0,1,2$; reward 1, may get more 2 as well].

- *Nonlinear payoffs.* Corneo (1997): uniqueness vs. multiplicity when esteem is a *convex* ("elitist") vs. *concave* ("conformist") function of an individual's perceived rank (uniformly distribution) in the population.

V. SPONSORS, WELFARE AND COMPETITION

Detail a bit more the public-good aspects of agents' contributions.

- There are n agents (small group or large population). If total supply of contributions is $na \Rightarrow$ public-good na/n^{κ}, valued by an individual at

$$w_a(n\bar{a}/n^{\kappa})$$

- Individual's intrinsic motivation to contribute now consists of

$$v_a = \underbrace{u_a}_{\text{joy of giving}} + \underbrace{w_a/n^{\kappa}}_{\text{concern for public good}}$$

  - Each contribution thus has external benefit to society of

$$\bar{e} \equiv (n-1)(\bar{w}_a/n^{\kappa})$$

and yields a gain in (self) esteem of $r(y)$ to the contributor

.

- Pursuit of esteem is a *zero-sum game:* average reputation in society remains fixed. But participation decision is based on private reputational return rather social one (zero) => inflicts an externality on others. How? Brings down quality of the pools of contributors *and* non-contributors.

- The net *social* return to contributing is thus

$$\underbrace{(n-1)(\bar{w}_a / n^{\kappa})}_{\bar{e}} - \underbrace{\mu_a(\mathcal{M}^{+} - \mathcal{M}^{-})(v_a^{*})}_{r(y)}$$

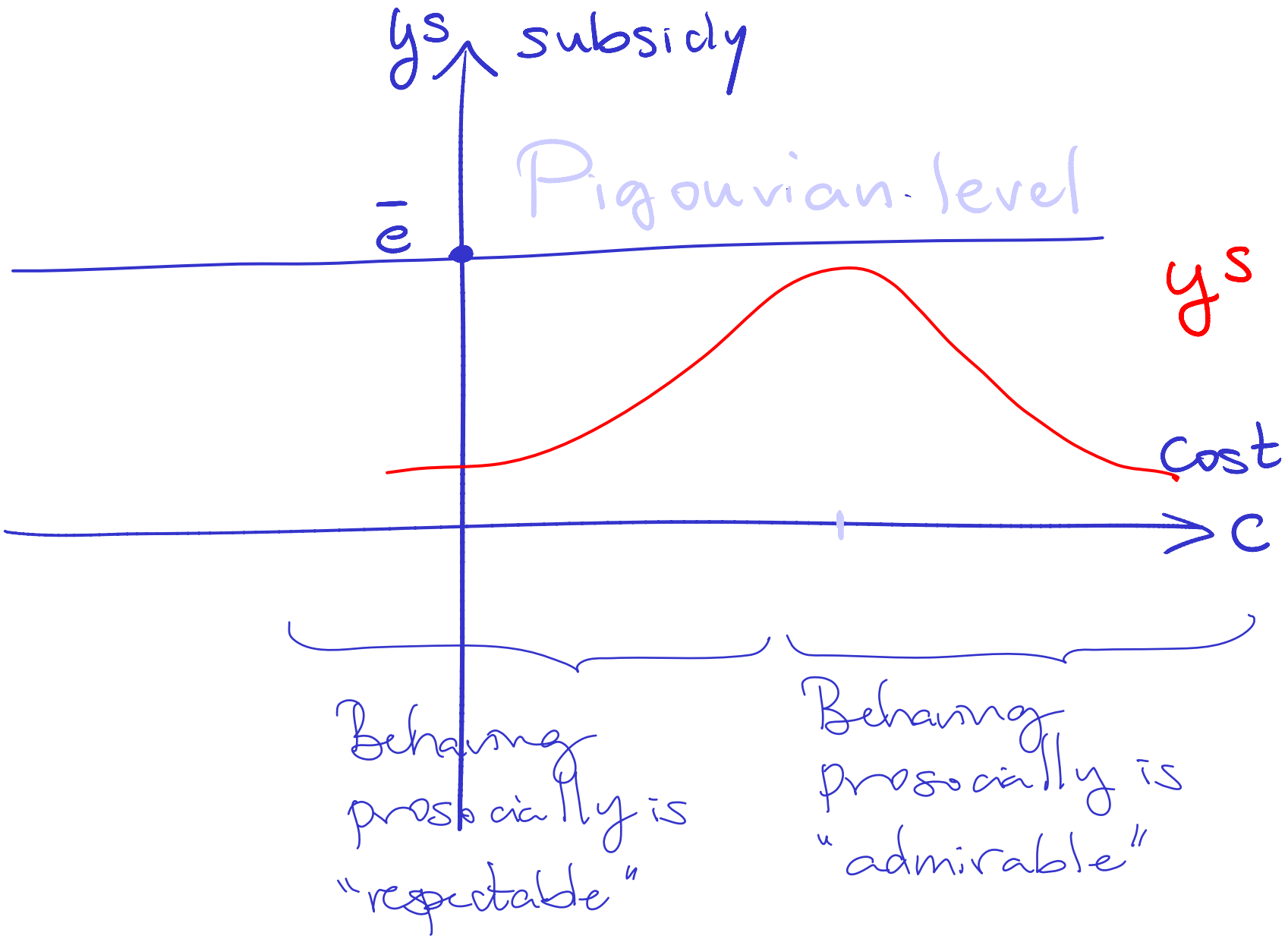
Difference between *free-riding* effect and *reputation-stealing* effects

Proposition 8 *1) The socially optimal incentive rate y^s, is strictly less than the standard Pigouvian subsidy that leads agents to internalize the full public-good value of their contribution. In basic case:*

$$y^s = \bar{e} - \mu_a(\mathcal{M}^{+} - \mathcal{M}^{-})(c - y^s)$$

Implications:

- Tax treatment of charitable contributions: deduction rate may be low, or may actually want to tax them. See also Blumkin-Sadka (2006) in a model of wealth signaling.
- Same for ethical funds, fair trade products?
- Optimal subsidy / tax / very different for “admirable” vs “respectable” prosocial behaviors. In particular, depends on *c*, nonmonotonically.


y^s subsidy
Pigouvian level
$\bar{e}$
y^s
Cost
c
Behaving prosocially is "respectable"
Behaving prosocially is "admirable"

Sponsors' choice of incentives and the social optimum

Look at sponsors who set y to maximize:

$$\bar{W}(y) \equiv \alpha \underbrace{\bar{U}(v_a^*(y); y)}_{\text{contributors' welfare}} + \pi(y), \text{ where } \alpha \in [0,1] \text{ and}$$

$$\pi(y) = n\bar{a}(y)\left(B - y\right).$$

- Social planner who internalizes the ex-ante utility of the n agents and has access to lump-sum taxes: $\alpha = 1$ and $B = 0$. But $B \geq 0$ could reflect a different discounting of the welfare of future generations, and $\alpha \leq 1$ a shadow cost of public funds

- For other actors such as NGO's or specialized government agencies, B reflects the purely private benefits (material or reputational) that the sponsor derives from contributions transiting through it and $\alpha < 1$ the weight it places on social welfare, both normalized by the sponsors' own opportunity cost of funds.

Proposition 8 *2) A monopoly sponsor may offer contributors a reward* y^m *that is too generous (or, require of them too low a donation) from the point of view of social welfare, resulting in excess participation. This is true even when the private benefits it derives from agents' participation coincide exactly with the gap between their social and private contributions to the public good.*

3) Competition between sponsors(e.g., nonprofits, NGO's) increases rewards (or, reduces required monetary contributions. In this case, competition reduces social welfare.

Intuitions

1) The optimal incentive scheme should include *a tax that corrects for the reputation-seeking motive* to contribute, which in itself is socially wasteful.

2) A monopolist setting y^m does not internalize the reputational losses of inframarginal agents to the same extent as a planner would. This gives it an incentive to attract too many customers, which works against the standard monopolistic tendency to serve too few.

When reputational concerns are important enough, this informational externality can dominate, *making the monopolist too generous* or not demanding enough in the standards it sets for monetary donations.

3) *Sponsor competition* further *exacerbates this inefficiency*, because each firm now has a much higher incentive to raise its offer than a monopolist (it takes the whole market), but still inflicts the same reputational cost on all inframarginal non-contributors.

VI. "HOLIER THAN THOU" SPONSOR COMPETITION

Introduce non-price competition. Show that free entry may reduce welfare, as leads to sponsors screening contributors in inefficient ways.

Illustrations: religions and sects competing in asceticism, rituals, sacrifice. Marathons, walks, etc. associated to fundraising for worthy causes.

- Two types: $v_a = v_a^H$ (prob. ρ) or $v_a = v_a^L$ (prob. 1-ρ)

- Non-monetary cost of contributing is c, unless sponsor demands a (verifiable) "sacrifice" $\Rightarrow$ becomes, for low and high type respectively

$$c^L >> c^H > c$$

- Sacrifice = *pure deadweight loss.* Only benefit for the sponsor is to help screen the agents, because it is less costly for the more motivated.

Proposition 1) *A monopoly sponsor who wants both types to contribute does not screen contributors inefficiently, but offers two prices.*

2) By contrast, competing sponsors may require high- valuation individuals to make costly sacrifices that represent pure deadweight losses. In this case, competition reduces social welfare.

Intuition: non-price screening imposes a negative externality on low-type agents, the cost of which a monopolist serving whole market (with two prices) must fully bear, but which competitive sponsors do not internalize.

SUMMARY / CONCLUSIONS

- Three basic motives for prosocial behaviors:
  - altruistic / public spirited
  - material self-interest
  - social or self image concerns.

Key: interactions between them + response to info. environment

- Altering incentives or visibility *changes the meaning* attached by observers (or self) to prosocial or antisocial behavior, and hence the reputational value of engaging in it.

- Very simple model, but many results. Sheds some light on:
  - Individual and aggregate contributions: crowding out (or in), complementarity/substitutability, multiple social norms…
  - Strategic behavior of public or private sponsors seeking to foster or capture prosocial contributions: disclosure / secrecy, equilibrium vs. socially optimal incentives, inefficiency of competition.

❑ Possible extensions / Future work

- Optimal mix of material and publicity incentives (*x* and *y*). Links to law and economics literature on legal vs. social sanctions (Kaplow and Shavell (2001), Rasmusen (2005)).

- Sponsors: deeper analysis of their objectives and behavior, including own reputational concerns. Ethical funds, socially responsible firms, fair trade products. Tax treatment.

- Self-image / self-perception version: agents judges themselves / assess their own preferences from their past behavior. Conversely, behave so as to maintain certain beliefs about "who they are".

  Not just for prosociality / greed (as here) but also: attachment to work or family, culture, ethnic background, politics, religion, etc. => basis for cognitive, psychologically founded model of identity.