Identity, Morals, and Taboos: Beliefs as Assets

The Quarterly Journal of Economics, 126(2), 805–855.

Xuhang Fan, Xinyi Xie, Jade Peng, 2022/9/25

Authors

Jean Tirole

- Nobel Prize (2014)
- Honorary Chairman, Toulouse School of Economics (2019)
- Ph.D., MIT (1981)

Jean Tirole

Roland J. M. Bénabou

- Jean-Jacques Laffont Prize (2021)

  "Beliefs and Misbeliefs: The Economics of Wishful Thinking"

  Motivate Belief Theory
- Professor at Princeton University (1999-)
- Ph.D., MIT (1981)

Roland J. M. Bénabou

Identity, Morals, and Taboos

Three things cannot be explained by standard model

Identity

- "Who I am", based on group membership
- Experimental evidence: **People care about Identity**
- People dislike the "free-riders" in the group and punish them(Fehr and Gächter, 2000)

Morals

- "Am I a moral person?"
- Experimental evidence: People reject the deceptive but profitable choice(Gneezy, 2005)
- **People’s economic decision is constrained by moral concern**

Taboos

- Information-Averting Behaviors
- People prohibit themselves from merely thinking about certain "priceless" concepts
- e.g. Markets for organs, genes, sex, surrogate pregnancy andadoption are widely banned on grounds that they would represent an *“unacceptable commodification”* of human life
- **"More information is not better" under some situations**

How to explain these results?

- include **social preference assumptions**
- People’s utility = Economic utility + Social Utility
- e.g. cheating helps to get higher grades, but I still wanna be a honest student
- It is called **"second generation of moral behavior"**

Motivating Facts And Puzzles

Unstable Altruism

Positive Side:

- fairness, cooperation, and honesty in anonymous, one-shot interactions

Negative Side:

- Excuse-seeking behavior(e.g. Garcia et al., 2020)
  - seek an excuse to behave selfishly while "feeling moral"
  - "I do not want to donate money because charities are not reliable "
- Moral Wiggle Room (e.g. Dana et al., 2007)
  - when a decision is vague in moral or immoral, people tend to behave selfishly

Coexistence Of Social And Antisocial Punishments

Social Punishments

- free-riders in public-good games, and violators of social norms more generally, get punished by others

Antisocial Punishments

- who behave too well elicit resentment, derogation, and punishment from their peers
- Such do-gooders always exhibit stronger moral principles or resilience than their peers

Taboo Tradeoffs

- **not make utility trade offs**
  - but considered immoral to place a monetary value on marriage, friendship, or loyalty to a cause;
  - no consideration for markets for organs, genes, sex, surrogate pregnancy andadoption
- **“mere contemplation” effect**
  - when prompted to simply envision or speculate about tradeoffs between sacred and secular values, subjects respond with noncompliance, outrage, and later symbolic acts of moral cleansing
- **Enforcement**
  - people seek to enforce such taboos not only **on others’ behavior**
  - but also **on their own(pre-commitment)**

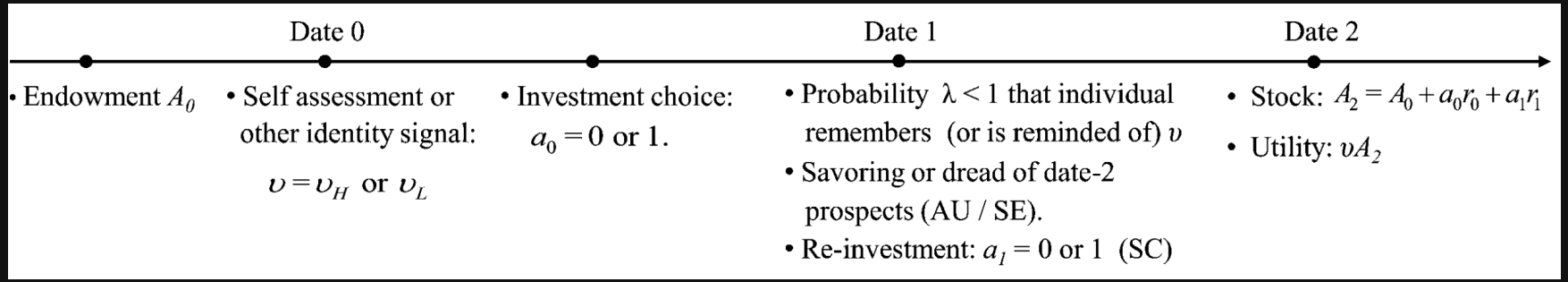
Solution: Belief as Assets

Third-generation theory of Moral Behavior

- **Belief as Assets**
  - let moral identity as beliefs about one's deep "values"
  - holding a positive self-image can increase utility
- **Self-inference Process**
  - judge oneself by own behavior or decisions
  - "Who Am I" partially comes from inference based on former decisions
- **Supply side**
  - use imperfect memory or awareness
  - because sometimes we are not sure "How good I am", we use self-inference
- **Demand side**
  - investment for identity management
  - choose the decision to let the self-inference process produce the positive belief

THE MODEL

Timing of Moves and Actions



Notations

A are “relational assets” and the individual’s long-run utility

υ is the individual’s long-run utility for the benefits flowing from it

a_t is moral/immoral decision. (= 1 if moral, = 0 if immoral)

r_t: the multiplier of moral decision

$A_{t+1} = A_t + \alpha_t r_t$ to measure the relative increase from choosing $a_t = 1$

Date 0. self-assessment

the agent has access to a signal about his type(good or bad)

$$v = \begin{cases} v_H & \text{with probability } \rho \\ v_L & \text{with probability } 1-\rho \end{cases}$$

prior expectation:

$$\bar{v} \equiv \rho v_H + (1-\rho) v_L$$

Assumption 1

The net cost of investment at date 0 is $c_0^H \gtrless 0$ for type H and c_0^L for type L, with $c_0^L \geq c_0^H$

Because a more prosocial individual internalizes more of the benefits accruing to other people, even in one-shot interactions, he finds it (weakly) less costly to act morally—help, refrain from opportunism

Date 1. Self-Inference

Assumption 2. (Self-inference)

the individual is aware of his true valuation v only with probability λ, so with $(1-\lambda)$, he cannot remember and infer his type based on former choice a_0

denote $\hat{\rho}$ as date-1 belief about his type

$$\hat{v} \equiv \hat{\rho} v_H + (1-\hat{\rho}) v_L$$

so with probability λ, $\hat{v}$ is v; and with $1-\lambda$, $\hat{v} = \hat{v}(a_0) \in [v_L, v_H]$

Note:
$(1-\lambda)$ is malleability of beliefs, the probability of information loss thus also reflecting the possibility that deeds may themselves be forgotten or repressed, or be uninformative due to situational factors that can be invoked as plausible excuses

Date 1. Self-Inference

Assumption 3
The value function $V = V(v, \hat{v}, A_1)$ satisfies $V_{\hat{v}} > 0, V_{v\hat{v}} \geq 0$ and, if $r_0 > 0, V_{13} > 0.$

$V_{\hat{v}} > 0$: a “good identity” convention, a moral self-image is better than not

$V_{v\hat{v}} \geq 0$: a sorting condition, when $c_0^H \leq c_0^L$, the investment of H type $\geq$ the investment of L type (behaving more prosocially), so that actions have informational content(type can be identified from the action)

Assumption 4. Exclude the Trivial Case

the investment cost is too low so that both types always invest regardless of identity concerns

we assume:

$$V\left(v_L, \hat{v} = v_L, A_0 + r_0\right) - V\left(v_L, \hat{v} = v_L, A_0\right) < c_0^L$$

Date 2. Future

vA_2 is long-term value

Benchmark Cases

Demand for beliefs 1: self-esteem / anticipatory utility

self-esteem (SE)

preference given by $V = s\hat{v}, \hat{v} \equiv \hat{\rho} v_H + (1 - \hat{\rho}) v_L$

s measures the strength of the self-esteem motive

anticipatory utility (AU)

more consequentialist

what is AU: hopefulness, anxiety, or dread that arise from **contemplating** future and social prospects

define long-term welfare: vA_2, the expected value of social relationships(family, friends, colleagues, ethnic group, etc.)

Date 1: utility=$s\hat{v}A_2$

s reflects both the intensity of such anticipatory feelings and their duration

Another important determinant of s is **salience**

Pure Anticipatory Utility (at Data 1)

no further decision to be made at date 1, $a_1 \equiv 0, A_2 = A_1$

the continuation value (evaluated from t = 0) of entering period 1 with $\hat{v}$ is

$$V(v, \hat{v}, A_1) = s\hat{v}A_1 + \delta v A_1$$

for $s\hat{v}A_1$: Date 1 utility

for $\delta v A_1$: δ is the discount factor between dates 1 and 2, vA_1 is date 2 utility

s/δ reflects also the relative lengths of periods 1 and 2

Assumption 2 is satisfied

$V_{13} = \delta > 0, V_{23} = s > 0$ and $V_{12} = 0$

self-esteem is a special case of *anticipatory utility*

SE is AU when $A_t \equiv 1, r_t \equiv 0$ and $\delta = 0$ (no “day of reckoning”)

so $V(v, \hat{v}, A_1) = s\hat{v}A_1 + \delta v A_1 = s\hat{v}$ the only relationship the agent cares about is with himself

welfare analysis: total intertemporal utility

$$W \equiv E\left[-a_o c_0 + V\right]$$

where the expectation is taken with respect to the prior distribution $(\rho, 1-\rho)$ of values $v \in \{v_H, v_L\}$ and the distribution $(\lambda, 1-\lambda)$ of (endogenous) posterior beliefs $\hat{v} \in \{v, \hat{v}(a_0)\}$.

Demand for beliefs 2: self-control

self-control (SC)

Maintaining a strong, stable sense of identity also has functional value, helping one to make consistent choices and resist harmful temptations

the context of social interactions, which inherently feature a tradeoff between short-term gains from selfishness (or emotional release) and long-run benefits from behaving morally.

long-term welfare: still be vA_2 in date 2

moral decision happen in $t = 0, 1$

assumption for simplicity *(smooth over t = 1 decisions, so as to make V differentiable)*

- Investment at $t = 1$ involves a stochastic cost c_1
- type-independent distribution $F(c_1)$ on R_+ *(can be relax)*

Perception of the cost of acting morally

At date 1, weakness of will make the immediate gains from opportunism more salient than its distant consequences, thus perceives the cost of acting morally as c/β, $\beta < v_L/v_H$

This condition implies that whenever the agent (either H type or both) chooses to behave cooperatively, it is *ex-ante efficient* for him to do so

$u = \delta v r_1$, when $\beta \delta v_H r_1 > c_1$ and $\beta v_H < v_L$, we can get $\delta v_L r_1 > c_1$

moral identity and self-restraint

given a self-view $\hat{v}$, the agent invests when $c_1/\beta \leq \delta \hat{v} r_1$

a threshold cost level that increases with $\hat{v}$

a stronger moral identity generates valuable self-restraint

$$V(v, \hat{v}, A_1) \equiv \delta v A_1 + \int_0^{\beta \delta \hat{v} r_1} (\delta v r_1 - c_1)\, dF(c_1)$$

Assumption 2 are satisfied if $(v - \beta\hat{v})\delta r_1 \geq (v_L - \beta v_H)\,\delta r_1 > 0$

welfare analysis

because the agent will generally have present-biased preferences at date 0, just like at date 1. Thus, if c_0 is the perceived investment cost, the "real" cost, as viewed by an ex-ante self or parent at date "–1", is only βc_0

V is also an *ex-ante* value function, our welfare criterion will be:

$$W = E\left[-\beta a_0 c_0 + V\right]$$

Equilibrium and Welfare: Solving the model

Each type chooses his optimal option a_0, $k = H, L$

$$\max_{a_0 \in \{0,1\}} \left\{ -c_0^k a_0 + \lambda V\left(v_k, v_k, A_0 + a_0 r_0\right) + (1-\lambda) V\left(v_k, \hat{v}\left(a_0\right), A_0 + a_0 r_0\right) \right\}$$

denotation of x_H and x_L

respective probabilities that types H and L behave prosocially at $t = 0$

$\hat{v}\left(a_0\right) \equiv \hat{\rho}\left(a_0\right) v_H + \left[1 - \hat{\rho}\left(a_0\right)\right] v_L$

where

$$\hat{\rho}(1) = \frac{\rho x_H}{\rho x_H + (1-\rho) x_L}, \hat{\rho}(0) = \frac{\rho\left(1 - x_H\right)}{\rho\left(1 - x_H\right) + (1-\rho)\left(1 - x_L\right)}$$

expected value function

$$\mathbf{V}\left(v, \hat{v}, A_1\right) \equiv \lambda V\left(v, v, A_1\right) + (1-\lambda) V\left(v, \hat{v}, A_1\right)$$

which brings together the **demand (preferences)** and **supply (cognition) sides**

inheriting from V all the properties in Assumption 3

Investment(t = 0) is optimal when:

$$\mathbf{V}\left(v_k, \hat{v}(1), A_0 + r_0\right) - \mathbf{V}\left(v_k, \hat{v}(0), A_0\right) - c_0^k \geq 0$$

The Sorting Condition.
net return to "good behavior" is greater for the H type than the L, implying that $\hat{v}(1) \geq \hat{v}(0)$

Reason

1. H type has a lower effective cost $c_0^H \leq c_0^L$
2. when $V_{13} > 0$, the agent attaches greater value to any increment to the capital stock
3. when $V_{12} > 0$, the agent also cares more about having a "strong" identity at date 1, which investing helps achieve if $\hat{v}(1) > \hat{v}(0)$

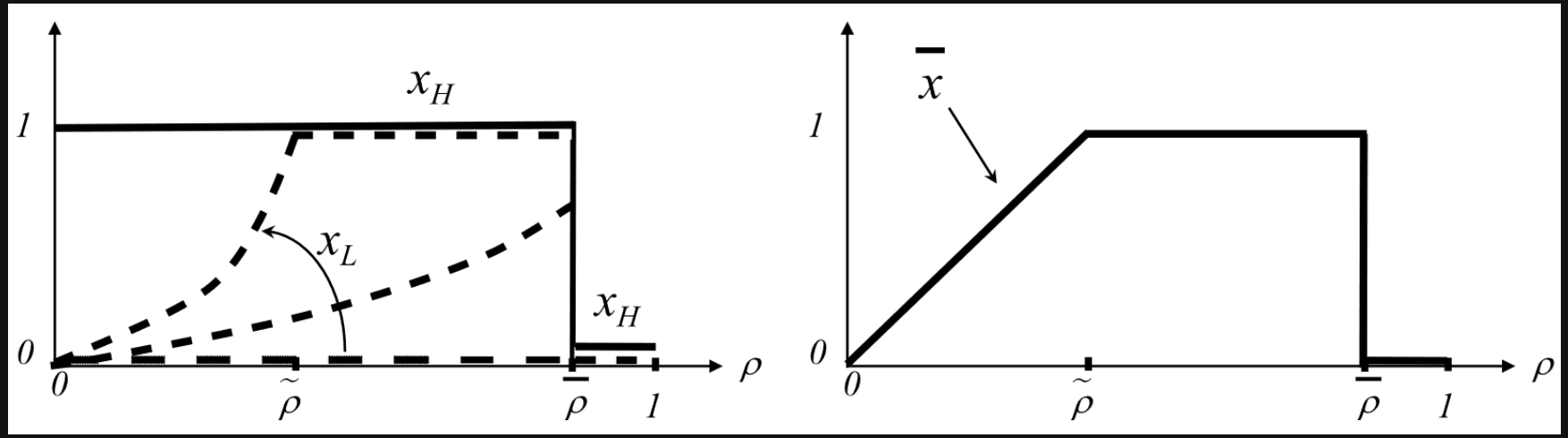
Monotonic Perfect Bayesian Equilibria

1. the H type always invests more: $x_H \geq x_L$, which given $\mathbf{V}\left(v_k, \hat{v}(1), A_0 + r_0\right) - \mathbf{V}\left(v_k, \hat{v}(0), A_0\right) - c_0^k \geq 0$ again means that $x_H = 1$ whenever $x_L > 0$
2. a (stronger) form of monotonicity is also imposed on off-the-equilibrium-path beliefs if $x_H = x_L = 0$, then $\hat{\rho}(1) \equiv 1$ if $x_H = x_L = 1$, then $\hat{\rho}(0) \equiv 0$ This refinement is intuitive and does not affect any qualitative results.
3. over a certain range of parameters there may be multiple (three) monotonic equilibria, among which one is Pareto-dominant and will be selected

Proposition 1.

There exists a unique (monotonic, undominated) equilibrium, characterized by thresholds $\tilde{\rho}$ and $\bar{\rho}$ with $0 < \tilde{\rho} \leq \bar{\rho} \leq 1$ and investment probabilities $x_H(\rho)$ and $x_L(\rho)$ such that:

1. $x_H(\rho) = 1$ for $\rho < \bar{\rho}$ and $x_H(\rho) = 0$ for $\rho > \bar{\rho}$
2. $x_L(\rho)$ is non-decreasing on $[0, \tilde{\rho}]$, equal to 1 on $[\tilde{\rho}, \bar{\rho})$ when $\tilde{\rho} < \bar{\rho}$ and equal to 0 on $[\bar{\rho}, 1]$



Left panel: solid line $= x_H(\rho)$, dashed line $= x_L(\rho)$, for decreasing values of c_0^L

No Investment. $(\rho > \bar{\rho})^{**}$

ρ = initial self-image inference

When initial self-image is good enough, the H type **can afford not to invest,** since the other one also behaves opportunistically the posterior will equal the prior, which is already close to 1 and thus could not be increased much anyway

Investment Cases. $(\rho < \bar{\rho})$

H invest to “stand for his principles” and separate from the L type

1. **Separation**. When c_0^L is sufficiently high, the low-valuation type does not find it worthwhile to invest $(x_L = 0)$, whereas the high-valuation type does.

2. **Randomization**.
   For lower values of c_0^L, L type intend to **imitate the H type**

   but ability of imitation is limited by the prior $(0 < x_L < 1, \tilde{\rho} = \bar{\rho})$

$$\hat{\rho}(1) = \frac{\rho x_H}{\rho x_H + (1-\rho)x_L}, \hat{\rho}(0) = \frac{\rho\left(1-x_H\right)}{\rho\left(1-x_H\right) + (1-\rho)\left(1-x_L\right)}$$

3. **Universal Investment**.

   For c_0^L still lower, even a small gain in self-image is worth pursuing, so $x_L = 1$.

   ρ is above the threshold $\hat{\rho}$ (which increases with c_0^L)

Proposition 2. Comparative-Statics Predictions

- An individual invests more in identity if

  1. the more malleable his beliefs (the lower λ);
  2. the lower the investment cost (the lower c_0^L or c_0^H)
  3. the more salient the identity in the SE/AU case (the higher s)
  4. the higher the capital stock A_0 in the AU case

- Initial beliefs have a **non-monotonic, hill-shaped effect** on overall investment

  $x(\rho)$ increases linearly on $[0, \tilde{\rho})$, equals 1 on $[\tilde{\rho}, \overline{\rho})$, then falls to 0 beyond.

Self-esteem/Anticipatory Utility and the Treadmill Effect

based on $V(v, \hat{v}, A_1) \equiv (s\hat{v} + \delta v)A_1$ and $W \equiv E[-a_0 c_0 + V]$

we get:

$$W = \rho x_H \left[(s+\delta)v_H r_0 - c_0^H\right] + (1-\rho)x_L \left[(s+\delta)v_L r_0 - c_0^L\right] + (s+\delta)\bar{v}A_0$$

- $(s+\delta)\bar{v}A_0$ is constant: although agents actively manage their self-views, this is a zero-sum game across types, by the law of iterated expectations
- $\rho x_H \left[(s+\delta)v_H r_0 - c_0^H\right]$ and $(1-\rho)x_L \left[(s+\delta)v_L r_0 - c_0^L\right]$: always (weakly) decrease as identity investments rise in response to a greater malleability of beliefs, 1−λ

an increase in his capital stock can also make the individual worse off.

the condition for a no-investment equilibrium $(x_H = x_L = 0)$ ceases to hold as A_0 crosses some threshold level. At that point investment **jumps up discretely**, **resulting in a net welfare loss**, by the same reasoning as above

$$\mathbf{V}\left(v_H, v_H, A_0 + r_0\right) - \mathbf{V}\left(v_H, \bar{v}, A_0\right) = \\ (s+\delta) v_H r_0 + (1-\lambda) s \left(v_H - \bar{v}\right) A_0 \leq c_0^H$$

treadmill effect

higher asset levels do no generate much of an increase in life satisfaction, or may even reduce it—and this precisely due to a **self-defeating pursuit of the belief that these assets will ensure happiness, or forestall misery**

a moral treadmill is much less likely than a material one

Diminishing marginal utility of consumption thus makes a **treadmill effect in material pursuits** likely at high wealth levels, but a non-issue for the poor.

Personal relationships and good deeds are arguably less subject to decreasing returns—those may even be increasing, through network effects and the spreading of reputation.

Proposition 3. In the anticipatory utility or self-image case:

1. An increase in the malleability of beliefs $(1-\lambda)$ always reduces welfare.
2. An increase in (per se valuable) capital A_0 can make the individual worse off.
3. An increase in salience s can also lower welfare

Note.

welfare analysis here is for one agent: not consider external costs and benefits on others.

But while costly actions are incurred partly for self-image purposes, their overall impact on it is zero. Therefore even though everyone values identity per se, its social value, positive or negative, must be found entirely in its “side-products.”

Willpower and the Commitment Value of Identity

basic self-control version of the model, A_0 has no behavioral impact

The malleability of beliefs, on the other hand, now affects behavior both at t = 0 and at t = 1

suppose 2 cases:

1. λ = 1, neither type behaves prosocially at t = 0: $c_0^H > \delta v_H r_0$, so $x_H = x_L = 0$
2. for some λ < 1, the equilibrium involves mixing: the more altruistic type always cooperates $(x_H = 1)$, while the more selfish one randomizes $(0 < x_L < 1)$

difference in intertemporal welfare is:

$$\begin{aligned}\Delta W =& (1-\rho)x_L\left(\delta v_L r_0 - \beta c_0^L\right) + \rho\left(\delta v_H r_0 - \beta c_0^H\right) \\ &+ (1-\lambda)E[\Delta V]\end{aligned}$$

while $E[\Delta V]$ reflects the effects of self-image management on date-1 behavior

$$E[\Delta V] = (1-\rho)x_L \int_{\beta\delta v_L r_1}^{\beta\delta\hat{v}(1)r_1} (\delta v_L r_1 - c_1)\, dF(c_1) - \quad \rho \int_{\beta\delta\hat{v}(1)r_1}^{\beta\delta v_H r_1} (\delta v_H r_1 - c_1)\, dF(c_1)$$

first term: how, when the L type invests at t = 0, this strengthens his moral self-regard and thereby raises his subsequent propensity to behave well

such pooling at t = 0 dilutes the identity of the H type, self-doubt increases the likelihood that he will be succumb to opportunism

Since prosocial investment at t = 1, when it occurs, is always ex-ante optimal (by Equation (6)), the first effect leads to a welfare gain, the second to a loss

Proposition 4.

In the self-control case, more malleable beliefs (lower λ) can raise welfare, by improving choices at $t = 1$ (when $E[\Delta V] > 0$) and/or at $t = 0$ when $\Delta W > (1-\lambda)E[\Delta V]$)

Taboos and Transgressions

Taboos and Transgressions

distinguish two complementary ways in which they operate—**ex ante** and **ex post**

1. **self enforced**, aims to avoid dangerous (self-) knowledge that might surface from “cold” analytical contemplation of what short-run tradeoffs might be available or expedient
2. **socially enforced**, is a form of information destruction aimed at repairing the damage to beliefs caused when someone, through his actions or speech, has violated a norm or taboo.

Self-enforced taboos: Information Avoidance

Setting

type(v)

Let $v \in v_H, v_L$ denote the **long-run value of some important asset, relative to** A_t

Selling decision of Assets

Suppose $date = 0$, an agent can find a price p and sell one unit of A_0

price distribution is:

$$p = \begin{cases} p_H \text{ with probability z} \\ p_L \text{ with probability 1-z} \end{cases}$$

Selling decision depends on price found

let p_H be high enough and p_L low enough → transact or not is a signal of type

when $p = p_H$: always sell A_0, implies:

$$p_H > \mathbf{V}\left(v_{H,} v_H, A_0\right) - \mathbf{V}\left(v_H, v_L, A_0 - 1\right)$$

when $p = p_L$: no transaction, implies:

$$p_L < \mathbf{V}\left(v_L, v_H, A_0\right) - \mathbf{V}\left(v_L, v_L, A_0 - 1\right)$$

Choice (a_0)

$$choice = \begin{cases} a_0 = 0, \text{check the price} + \text{consider selling } A_0 \\ a_0 = 1, \text{never to place a price on certain goods} \end{cases}$$

contemplation is done once check
the agent will recall that he contemplated the possibility of a transaction and evaluated whether maintaining his identity or dignity was “worth it”

Assumption: transaction must with price checking

transacting without first finding out the price is either **infeasible**, or else **unprofitable**.

$zp_H + (1-z)p_L$ is too low.

Taboo holding Condition

💡 Taboo is formed when everyone choose $a_0 = 1$

$$V(a_0 = 1) - V(a_0 = 0) = \mathbf{V}\left(v, \hat{v}(1), A_0\right) - \mathbf{V}\left(v, \hat{v}(0), A_0 - z\right) \geq zp_H + (1-z)p_L \approx zp_H$$

Note.

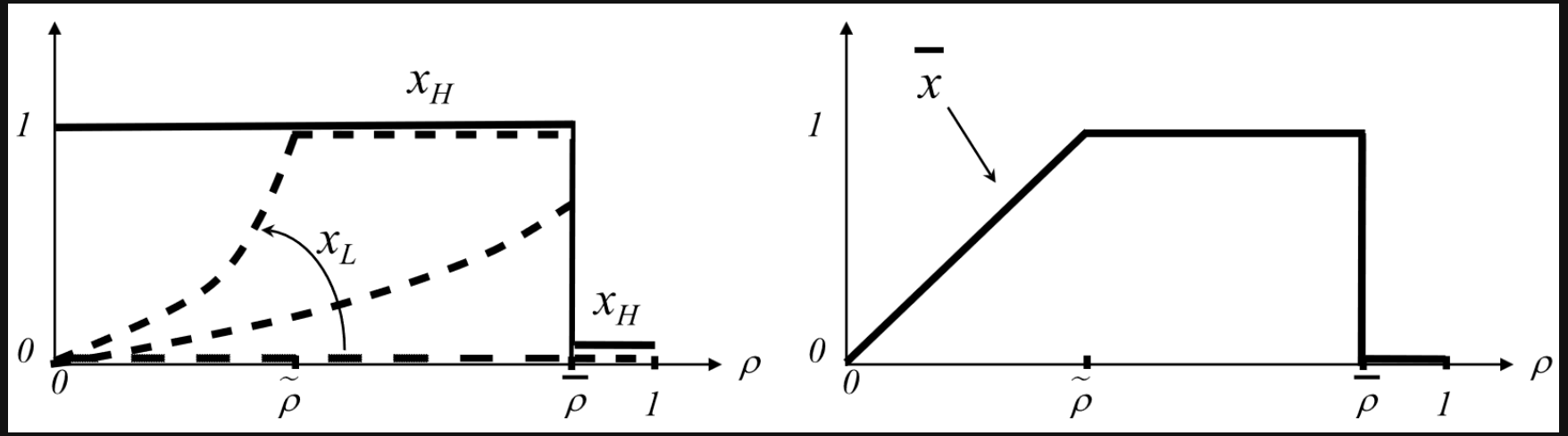
$\mathbf{V}\left(v, \hat{v}(0), A_0 - z\right)$ can be written because V is linear in A_1 in AU and SC case

Conclusions(Positive Side)

How taboos arise and are sustained

special case of former model: where $r_0 = z, c_0 = zp_H$ and initial stock $A_0' \equiv A_0 - z$

full-investment equilibrium or predominantly by the more committed (mixing or separating equilibrium), depends on the initial strength of beliefs ρ

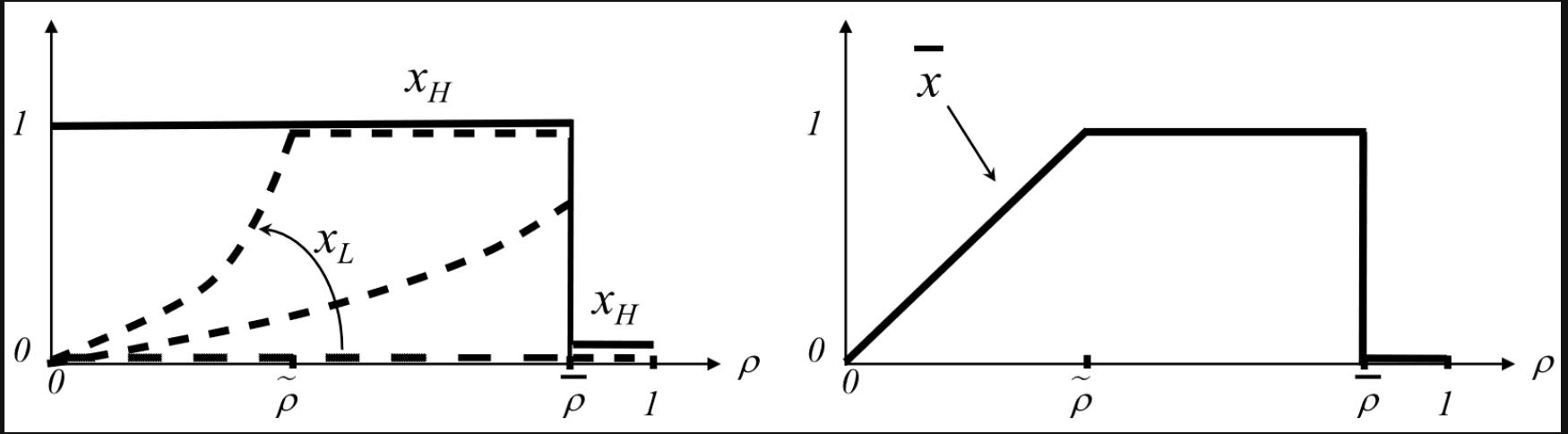


Taboo's Reaffirmation or Collapse

according to which side of the "hill" the induced erosion of ρ occurs on

on the right side: ρ decrease → reaffirmation

on the left side: ρ decrease → collapse



Conclusions(Normative side)

Propositions 3 and 4 show how the *welfare effect of taboos* depends on whether they reflect anticipatory or self-control motives. In the first case, upholding taboos generally lowers an individual's ex-ante welfare. In the latter it can be beneficial, but only under specific conditions involving the severity of the selfcontrol problem.

Note.
Proposition 3.

1. An increase in the malleability of beliefs $(1-\lambda)$ always reduces welfare.
2. An increase in (per se valuable) capital A_0 can make the individual worse off.
3. An increase in salience s can also lower welfare

Sinners and Saints: Information Destruction

socially enforced taboos: the coexistence of both social and antisocial punishments

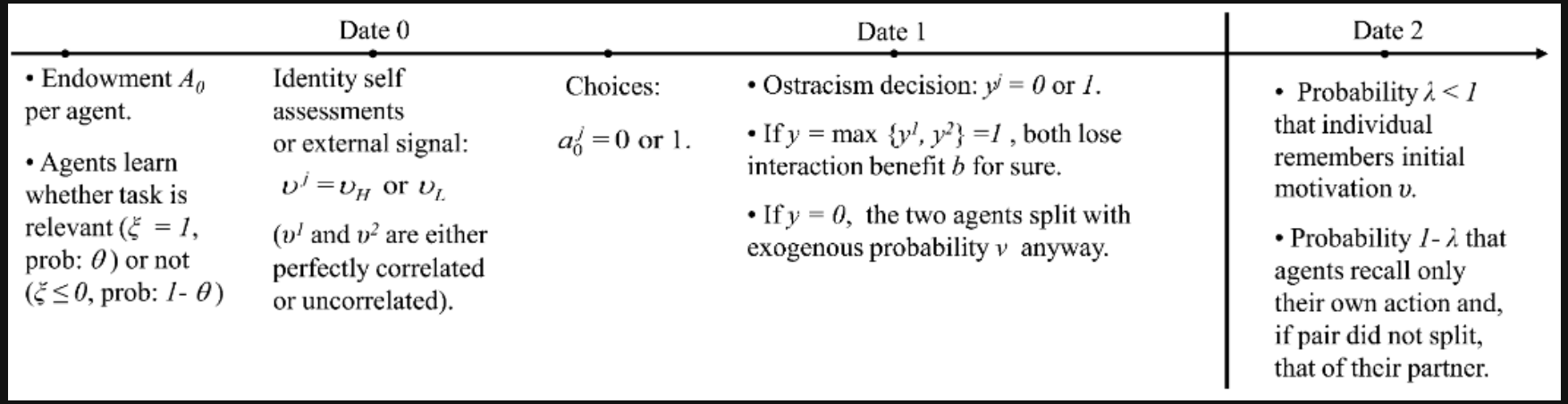
benchmarking idea

people compare themselves to others who they feel are akin to them or face a similar environment while asseing a person's type

- Deviant behavior$(a_0 = 0)$: sends a negative signal about the value of the existing capital stock (anticipatory utility version) or that of motivation-sensitive future investments (imperfect willpower version)
- Good behavior: lapsed individual is oneself, and by peers that is threatening to the self-concept, as it takes away potential excuses involving situational factors or moral ambiguity

In either case, the exclusion of mavericks from the group suppresses the undesirable reminders created by their presence: "out of sight, out of mind." That is, exclusion lowers λ

New elements



Action of benefit society or not(a_0).

Action of exclusion from group or not(y_i)

Agent i's utility function

Action of benefit society or not(a_0).

1. with ex-ante probability θ, agent have an “excuse” for deviate from the group
   choosing $a_0 = 1$ is useless—maybe harmful—to the rest of society, or where the private cost is so high that even the most moral types (H) would choose $a_0 = 0$(in former case)
2. With ex-ante probability 1−θ, the action $a_0 = 1$ is socially beneficial, the return of relational capital is $r_0^k = \xi v_k$ ξ = 1 when the action benefit others and ξ ≤ 0 when not
   $\tilde{c}_0^H \geq \tilde{c}_0^L$, different from before, reflecting the fact that a more prosocial agent is less inclined to engage in a socially harmful action.

Action of exclusion from group or not(y_i)

two agents after observing each other’s action, decide whether to continue in the relationship $(y_i = 0)$ or to break it $(y_i = 1)$
if someone exit, both lose b as future interactions benefit

Agent i's utility function

$y \equiv 1 - \left(1 - y^i\right)\left(1 - y^j\right)$ is the probability that ostracism occurs

$$\left(v^i \xi - c_0^i\right) a_0^i + \mathbf{V}\left(v^i, \widehat{v}^i, A_0 + r_0 a_0^i\right) + (1-\nu)(1-y)b$$

date 1: (no-recall assumptions) each agent always remains aware of his own behavior a_i^0, but he recalls that of his partner **only if they are still together**. If a split occurred, he recalls neither a_0^j nor what caused the separation (extreme and meant only to simplify the derivations)

Benchmarking on the Person

two individuals' values are **perfectly correlated**

date-0 contribution is always socially useful $(\xi \equiv 1)$

two individuals' values are perfectly correlated: $v^1 = v^2 \in \{v_H, v_L\}$

Benchmarking on the Situation

the two types are **independent**

social usefulness ($\xi = 1$, with probability θ) or absence($\xi \leq 0$, with probability $1 - \theta$) of the date-0 contribution is situation-specific, and the same for both agents

When faced with a given situation agents are able to assess ξ, but later on it, too, is subject to imperfect recall (or self-serving memory distortion) with probability $1 - \lambda$.

focus on symmetric equilibria (in undominated strategies) in which the more altruistic type always invests when $\xi = 1$ and no one invests when $\xi \leq 0$ (either ξ is sufficiently negative, or $\xi = 0$ and the value of self-image is low enough).

Proposition 5. In an equilibrium such that the H type invests when it is socially useful (ξ = 1), let x ∈ [0, 1] denote the probability of investment by the L type.

- Ostracism (y = 1) occurs only when actions differ, i.e. one agent invests and the other not.

  shows how a value of social conformity (strategic complementarity) arises endogenously from individual concerns over self -image because each agent has an incentive to exclude those who act differently from him

- Social punishment and Anti-social punishment

  ostracism comes from the virtuous agent $(a_0^j = 1)$ when benchmarking is on the person

  ostracism comes from the unvirtuous one $(a_0^i = 0)$ when benchmarking is on the situation.

  correspond to experimental eveidence: free-riders in public-good games get punished, but also who exhibit stronger moral principles or contribute “too much” to public goods are get punished

- With both the AU/SE and SC specifications and under either type of benchmarking, there exists a (positive -measure) range of parameters such that both $x = 1$ and $x = 0$ are equilibria:

  1. When benchmarking is on the person, x = 1 is sustained by the ostracism of “sinners” (a prosocial norm), while x = 0 involves no ostracism
  2. When benchmarking is on the situation, x = 0 is sustained by the ostracism of “do-gooders” (an antisocial norm), while x = 1 involves no ostracism

  shows *cross-society-differences* in civic norms and how they are enforced

Further Applications

Questions

- why not use a genral price distribution?

there may be two signals of an agent's type: whether check the price and, if so, whether he transacted or not, given the price.

to isolate the effects("priceless" effect and "mere-contemplation" effect)

Thanks!

Xuhang Fan, Xinyi Xie, Jade Peng