

Identity, Morals, and Taboos: Beliefs as Assets

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

Xuhang Fan, Xinyi Xie, Jade Peng, 2022/9/27

Table of Contents

- Background
- Identity, Morals, and Taboos
- Motivating Facts And Puzzles
- Solution: Belief as Assets
- The Model
- Benchmark Cases
- Equilibrium and Welfare: Solving the model
- Taboos and Transgressions
- Conclusion and Discussion

Background

Jean Tirole

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

Research interests

Industrial Organization, Regulation,
Organization Theory, Game Theory, Finance,
Macroeconomics, Psychology



Jean Tirole

Roland J. M. Bénabou

- Jean-Jacques Laffont Prize (2021)
Lecture "Beliefs and Misbeliefs: The Economics of Wishful Thinking"
- Professor at Princeton University (1999-)
- Ph.D., MIT (1981)

Research interests

inequality, growth, social mobility and the political economy of redistribution; education, social interactions and the socioeconomic structure of cities; economics and psychology ("behavioral economics")



Roland J. M. Bénabou

Economics "invade" Psychology

Provide an economic view

This paper provides a extremely flexible and feasible model to explain experimental results from Behavioral Economics and Psychology.

Strength of the model but...

Identity, Morals, and Taboos

Three things cannot be explained by standard model

Identity

- "Who I am"
- Experimental evidence: **People care about Identity**
- People with a strong identity(willpower) can resist the temptation and self-control

Morals

- "Am I a moral person?", based on your own decision
- Experimental evidence: People reject the deceptive but profitable choice(Gneezy, 2005)
- **People's economic decision is constrained by Moral Concern**

How to explain these results?

- In former models:
- 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 because keeping honesty brings happiness
- It is called "**second generation of moral behavior**"

Taboos: Information-Averting Behaviors

- People think it "immoral" to place a monetary value on some "priceless" concepts
- People prohibit themselves from **merely thinking** about taboos
- e.g. Markets for organs, genes, sex, surrogate pregnancy and adoption are widely banned on grounds that they would represent an "*unacceptable commodification*" of human life
- **"More information is not better" under some situations**

Motivating Facts And Puzzles

Unstable Altruism

Positive Side:

People have strong preference for being a good person.

- fairness, cooperation, and honesty in social interactions (anonymous, one-shot)

Negative Side:

People prefer to act selfishly to gain extra money while "feeling moral".

- Excuse-seeking behavior (e.g. Garcia et al., 2020)
 - "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 uncertain in morality, people tend to strategically 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

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

- **How can self-inference happen**

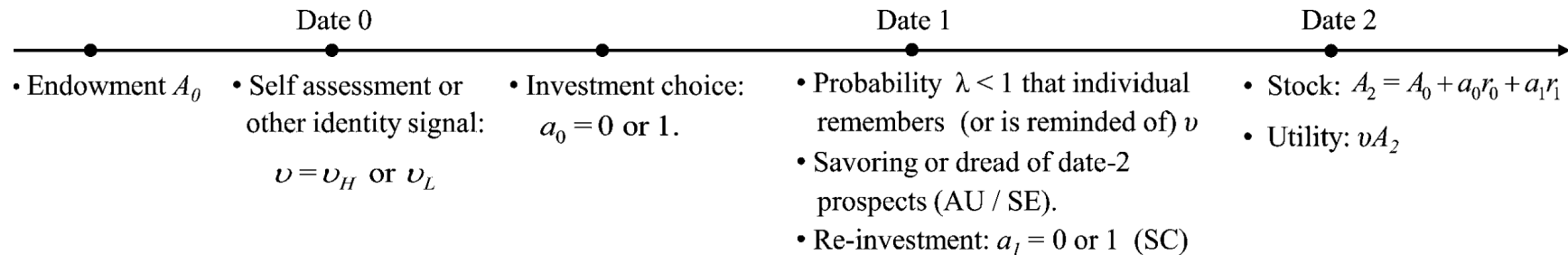
- use imperfect memory or awareness
- because sometimes we are not sure "How good I am", we use self-inference

- **The result of self-inference**

- 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

v is person’s type "good or not"

a_t is investment decision. (= 1 invest in A , and = 0 not)

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 $\rightarrow v$

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 $c_0^H \geq 0$ 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 $\rightarrow \hat{v}$

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 $\rightarrow \hat{v}$

Assumption 3

The value function $V = V(v, \hat{v}, A_1)$ satisfies $V_{\hat{v}} > 0$, $V_{\hat{v}v} \geq 0$ and, if $r_0 > 0$, $V_{vA_1} > 0$.

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

$V_{\hat{v}v} \geq 0$: 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

we do not want: the investment cost is too low so that both types always invest regardless of identity concerns

we assume:

$$V(v_L, \hat{v} = v_L, A_0 + a_0 r_0) - V(v_L, \hat{v} = v_L, A_0) < c_0^L$$

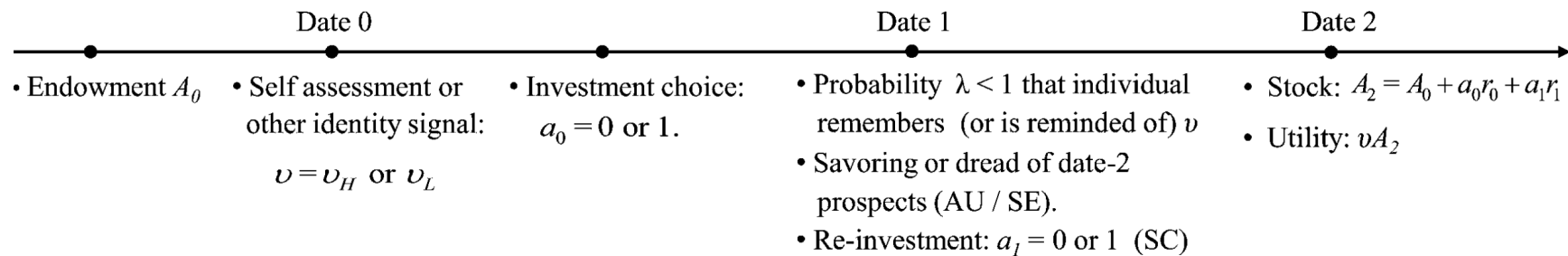
Date 2. Future

vA_2 is long-term value

Benchmark Cases

- We use two benchmark cases to get more interesting results
- Each case has a specific $V(v, \hat{v}, A_t)$ function

reminder of our timing:



Benchmark Cases

Case 1: Anticipatory Utility(self-esteem)

Decision only happens at Date-0

$$a_1 \equiv 0, A_2 = A_1$$

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

Long-term Utility

vA_2 , the expected value of social relationships(family, friends, colleagues, ethnic group, etc.)

Intertemporal utility function

$$V(v, \hat{v}, A_1) = s\hat{v}A_1 + \delta vA_2$$

s : anticipatory feelings or salience

δ : the time discount factor

Self-esteem is a special case of anticipatory utility

SE is AU when $A_t \equiv 1$, $r_t \equiv 0$ and $\delta = 0$, so

$$V(v, \hat{v}, A_1) = s\hat{v}A_1 + \delta vA_1 = s\hat{v}$$

Welfare Analysis

total intertemporal utility

$$W \equiv E[-a_0 c_0 + V]$$

depends on:

1. prior beliefs $v \in \{v_H, v_L\}$, which depends on ρ
2. posterior beliefs $\hat{v} \in \{v, \hat{v}(a_0)\}$, which depends on λ

Case 2: Self-Control

Present bias

At date 1, a myopic person' perceived cost of acting morally is c/β ($\beta < v_L/v_H$)

so β exaggerate present cost c

Investment decision a_0 happens when $t = 0, 1$

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

Case 2: Self-Control (Cont.)

Moral Identity and Self-Restraint

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

Total Intertemporal Utility

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

$\delta v A_1$: default long-run utility

$(\delta v r_1 - c_1)$: extra utility from investment choice

Welfare Analysis

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

β is reversed present bias from Date-0

Equilibrium and Welfare: Solving the model

Utility Maximization

Expected Value Function

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

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

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

Utility depends on $\hat{v}(a_0)$ and a_0

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

where $\hat{\rho}$ relates to ρ, x_K

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

x_H and x_L : probabilities that types H and L behave prosocially at $t = 0$

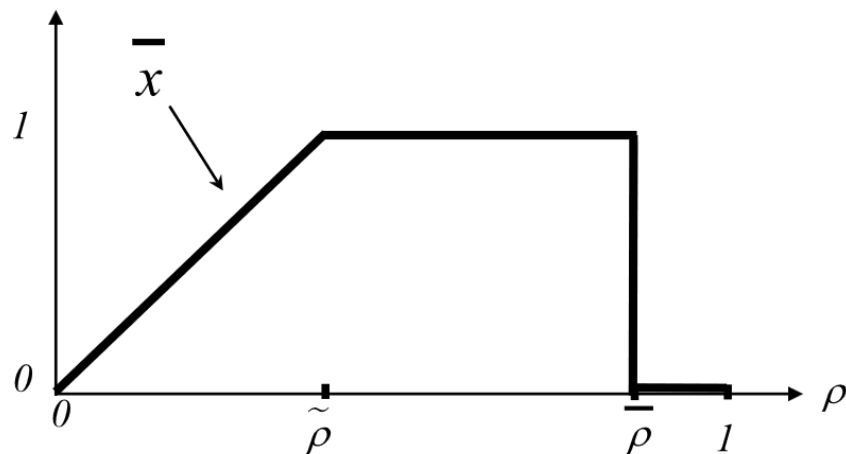
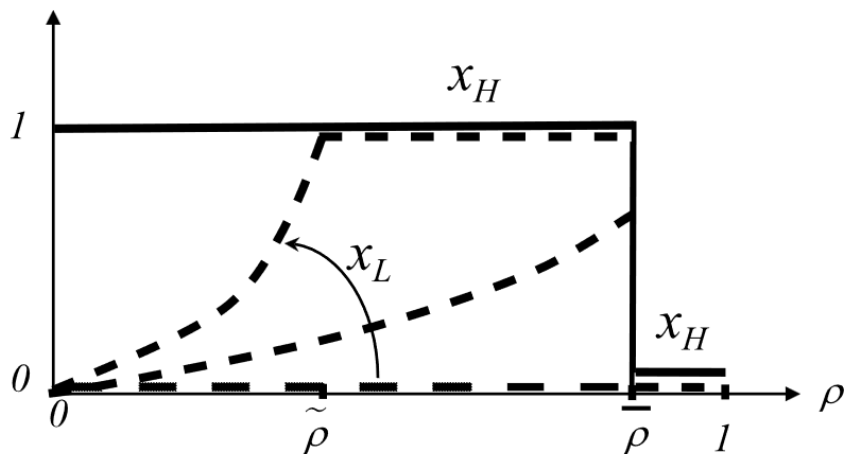
Utility Maximization (Cont.)

Investment choice ($a_0 = 1$) is optimal when:

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

Monotonic Perfect Bayesian Equilibria (Proposition 1.)

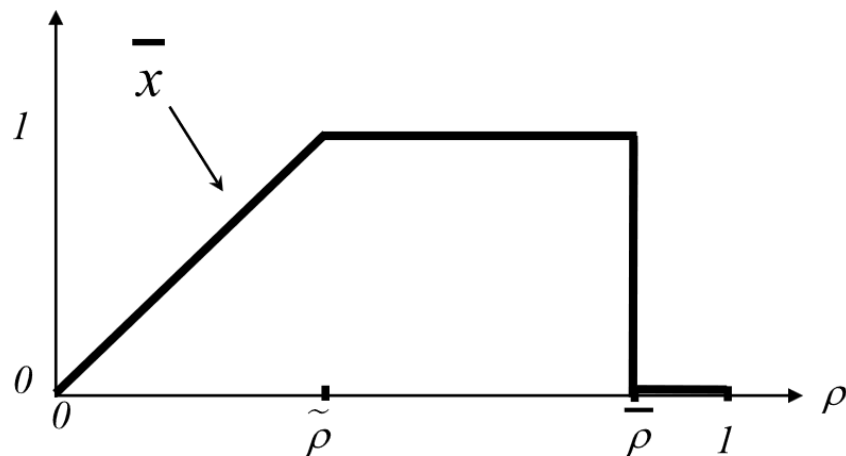
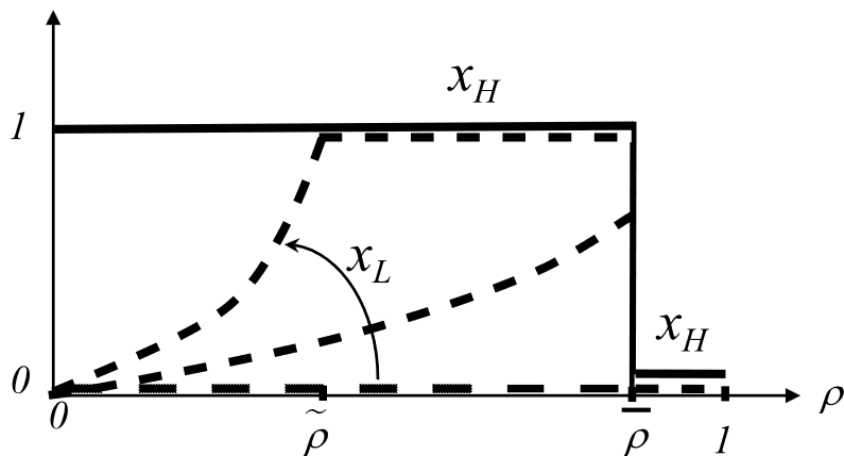
1. $x_H(\rho) = 1$ for $\rho < \bar{\rho}$ and $x_H(\rho) = 0$ for $\rho > \bar{\rho}$
2. $x_L(\rho)$ is non-decreasing on $[0, \tilde{\rho}]$, equal to 1 on $[\tilde{\rho}, \bar{\rho})$ when $\tilde{\rho} < \bar{\rho}$ and equal to 0 on $[\bar{\rho}, 1]$



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.

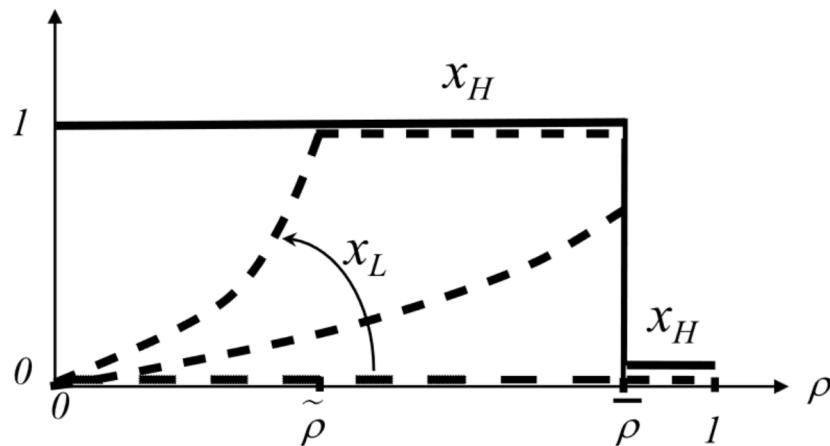
When c_0^L is high, the low-valuation type does not find it worthwhile to invest ($x_L = 0$)

2. Randomization.

For lower values of c_0^L , L type intend to **imitate H type** (but ability of imitation is limited by the prior)

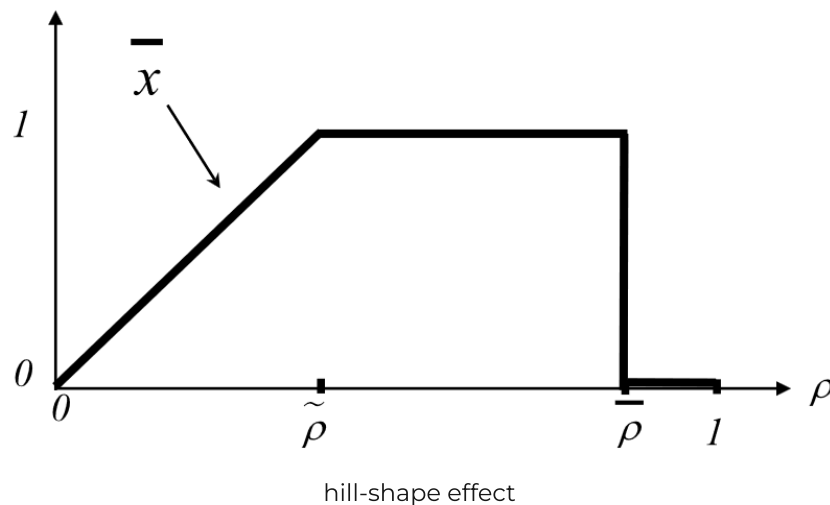
3. Universal Investment.

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



Comparative-Statics Analysis (Proposition 2.)

- An individual invests more in identity if
 1. more malleable self-belief ($\lambda \downarrow$)
 2. lower investment cost ($c_0^L \downarrow$ or $c_0^H \downarrow$)
 3. more salience (SE/AU case, $s \uparrow$)
 4. higher the capital stock A_0 (AU case, $A_0 \uparrow$)
- Initial beliefs have a **non-monotonic, hill-shaped effect** on overall investment



Treadmill Effect (AU/SE case)

Let's examine the no-investment condition:

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

Notice that when A gets sufficiently large, the agent unavoidably chooses to invest, thus reducing his/her lifetime utility.

More broadly speaking (easy to see when $r_0 \approx 0$):

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

The first two terms decrease as identity investment rises.

This leads to more interesting findings.

Proposition 3.

In Anticipatory Utility case:

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

reminder: when r_0 is relatively small, the underlined items are **negative**

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

Commitment Value of Identity (Self-control case)

let assume two cases:

- (a) $\lambda = 1$, neither type behaves prosocially at $t = 0$: $c_0^H > \delta v_H r_0$, so $x_H = x_L = 0$
- (b) $\lambda < 1$, the equilibrium involves mixing: H type cooperates, while L type randomizes.

Difference in intertemporal welfare:

$$\Delta W = W(b) - W(a) = (1 - \rho)x_L (\delta v_L r_0 - \beta c_0^L) + \rho (\delta v_H r_0 - \beta c_0^H) + (1 - \lambda)E[\Delta V]$$

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

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

Proposition 4.

In the self-control case, more malleable beliefs ($\lambda \downarrow$) 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

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

Setting:

$\text{type}(v)$

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

Taboo breaking = Selling decision of Assets

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

price distribution is:

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

Investment choice (a_0)

$$choice = \begin{cases} a_0 = 0, \text{ check the price + consider selling } A_0 \\ a_0 = 1, \text{ keep the taboo, think it priceless} \end{cases}$$

💡 **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”

Selling decision depends on price found

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

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

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

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

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

Taboo holding Condition (in AU and SC case)

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

A special case of former model

where $a_0 r_0 = z$, $c_0 = zp_H$ and initial stock $A'_0 \equiv A_0 - z$

Note.

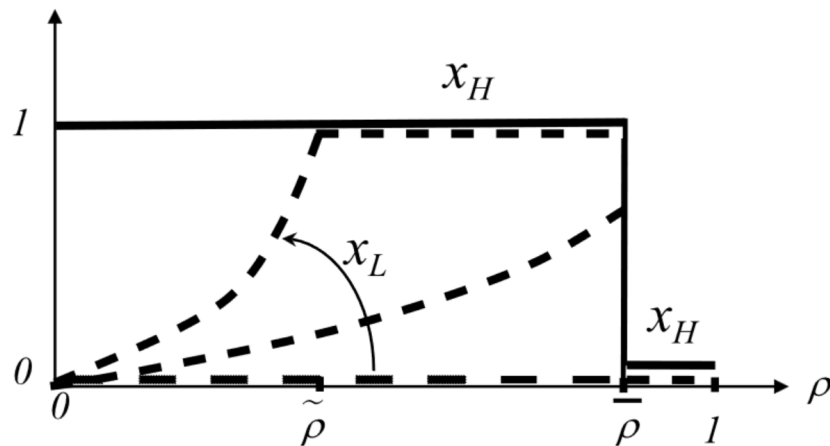
$\mathbf{V}(v, \hat{v}(0), A_0 - z)$ can be written because V is linear in A_1 in AU and SC case

Conclusions

How taboos arise and are sustained

from proposition 1 and 2, it depends on the initial beliefs ρ

1. Full-investment equilibrium
2. More committed (mixing or separating equilibrium)

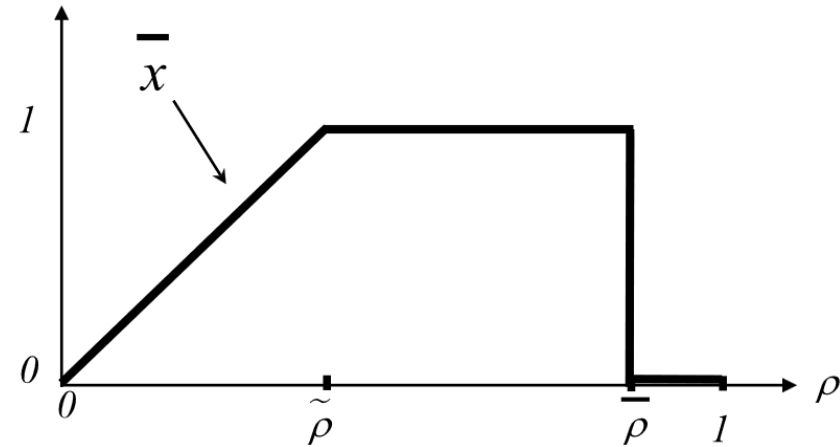


Conclusions (Cont.)

Taboo's Reaffirmation or Collapse

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

1. on the right side
 ρ decrease \rightarrow reaffirmation
2. on the left side
 ρ decrease \rightarrow collapse



Conclusions (Cont.)

Welfare effect of taboos

1. In AU case: upholding taboos generally lowers an individual's ex-ante welfare
2. In Self-control case: it can be beneficial, but only under specific conditions

Note.

Proposition 3. An increase in (per se valuable) capital A_0 can make the individual worse off.

Socially-enforced taboos

Focus: coexistence of social and antisocial punishments

New elements

■ Investment Choice (a_0)

1. $a_0 = 1$: with probability θ , the decision is not socially beneficial; with probability $1 - \theta$, $a_0 = 1$ is socially beneficial, return of relational capital is

$$r_0^k = \xi v_k, \tilde{c}_0^H \geq \tilde{c}_0^L$$

$\xi = 1$ when the action benefit others and $\xi \leq 0$ when not

New elements (Cont.)

- **Ostracism Decision** (y_i)

1. **continue relationship or not**: two agents after observing each other's action, decide whether to continue in the relationship ($y_i = 0$) or to break it ($y_i = 1$)
2. **interactions benefit**: if someone exit, both lose b

- **Agent i utility function**

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

ostracism happens condition: $y \equiv 1 - (1 - y^i)(1 - y^j)$

New Timing

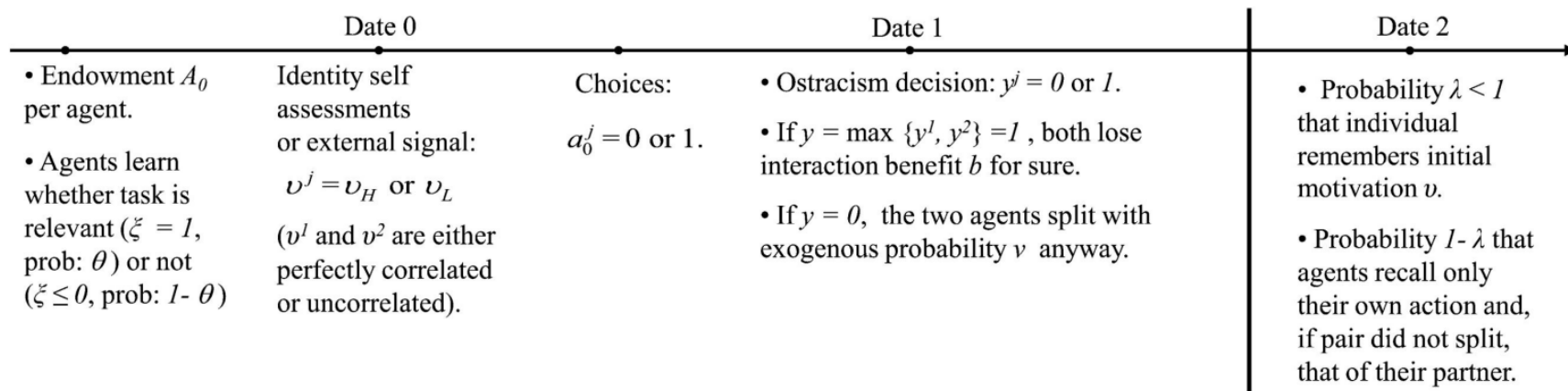


FIGURE III
The Ostracism Game

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)

Two Extreme Benchmark

1. Benchmarking on the Person

- two types are the same $v_1 = v_2 \in \{v_H, v_L\}$
- a_0 is always socially useful ($\xi \equiv 1$)

2. Benchmarking on the Situation

- two types are independent
- Whether a_0 is socially useful is random: $\xi = 1$, with probability θ or $\xi \leq 0$, with $1 - \theta$
- When faced with a given situation agents are able to assess ξ , but later on they recall imperfectly with probability $1 - \lambda$.

Proposition 5.

In an equilibrium such that the H type invests when ($\xi = 1$), let $x \in [0, 1]$ denote the probability of investment by the L type.

- Ostracism occurs **only when actions differ**, i.e. one agent invests and the other not.
 - because each agent has an incentive to exclude those who act differently from him
 - social conformity arises endogenously from self-image concern
- Co-existence of Social punishment and Anti-social punishment

benchmarking is on the person = Social punishment

ostracism comes from the good agent

benchmarking is on the situation = Anti-social punishment

ostracism comes from the bad agent

Proposition 5 (Cont.)

- 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)

$x = 0$ involves no ostracism

2. When benchmarking is on the situation

$x = 0$ is sustained by the ostracism of “do-gooders” (an antisocial norm)

$x = 1$ involves no ostracism

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

Conclusion and Discussion

Conclusion

1. A more general third-generation theory of moral behavior, individual and collective, based on the identity in which people care about “who they are” and infer their own values from past choices.
2. The paper proposed the monotonic Perfect Bayesian equilibria of welfare with three scenarios.
3. Taboos can be formed by internally enforced and socially enforced
4. High endowments trigger escalating commitment and a treadmill effect
5. Competing identities can cause dysfunctional capital destruction

Further Applications

Other Dimensions of Identity

■ Salience of Identity

Messages or cues that make specific components of a person's identity more salient elicit investments along the same dimensions.

Application of salience is advertising, much of which plays up people's desires to achieve or affirm certain identities—raising s with respect to beauty, wealth, or social status. Proposition 3 shows that such messages can be very effective in inducing consumers to purchase ($a_0 = 1$) and yet substantially lower overall welfare.

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

- Uncertain Values and Malleability of Beliefs

People are insecure about “who they are” (p in the middle range) are the most prone to costly identity-affirming behaviors. E.g. adolescents; male subjects with strongly declared homophobia actually showed the most arousal in response to male homoerotic videos.

- Escalating Commitment

someone who has built up enough of some economic or social asset—wealth, career, family, culture, etc.—continues to invest in it even when the marginal return no longer justifies it. This leads to excessive specialization (e.g., work versus family) and persistence in unproductive tasks

e.g. A manager will thus keep throwing good money after bad on a doomed project

Extensions of the Basic Model

Social Signaling. In addition to their self-image \hat{v} , people also care about others' perceptions \hat{v}' of their type, resulting in a continuation value of the form $V(v, \hat{v}, A_1, \hat{v}')$. Since others make inferences from observed behavior, adding a social signaling concern is akin to amplifying the self-image motive, so the entire analysis carries over (see again Appendix II).

The expected value function playing the role of Equation (II) is now

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

Thus, as long as

$$(v, \hat{v}, A_1) \mapsto V(v, \hat{v}, A_1, \hat{v}')$$

satisfies Assumption 3, adding a social signaling concern is akin to amplifying the self-signaling motive (from

$$(1 - \lambda)V_2 \quad \text{to} \quad (1 - \lambda)V_2 + V_4$$

and the whole analysis, positive and normative, carries over.

Questions

- How would you modify this model to incorporate depression and low self-esteem? Would you expect depression to be associated with greater or lower a ? Explain why. Is it consistent with the behavior of Mother Theresa, who suffered from severe depression?

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

Answer. When a person is low in self-esteem, they will not gain not much utility from a higher type, which reveals that s is low. As a result, a low-esteem person will lay more emphasis on self-inferred utility, so they may do some extreme or strange things to strengthen their identity perception.

Questions

- As people age and gain experience, presumably λ increases. However, it is not obvious that young adults are less pro-social than older adults; on the contrary, they may be more earnest and sanctimonious than their elders, who may share the decidedly un-Calvinist sentiments of Cal Smith. What other factors are likely age-related and, ultimately, would you expect λ to rise or fall with age?

Answer. Actually, we have experimental evidence shows that deceptive behavior significantly decreases with age (Glätzle-Rützler & Lergetporer, 2015), and another paper studying on the same relationship not got the significant results, but their adopted experimental paradigm may not ensure the credibility of individual-level data (Buccioli & Piovesan, 2011).

Aging changes people's enjoyment from social capital - gaining a bosom friend at your twentieth is different from knowing awesome friend when you are at the end of your life.

Thanks!

Xuhang Fan, Xinyi Xie, Jade Peng