Identity, Morals, and Taboos: Beliefs as Assets

Introduction

“Third-generation” theory of Moral Behavior

- What is second-generation

  Second-generation: induce social preference, but cannot explain all the thing

  e.g. people want help others, but dislike helping-too-much behavior
- Based on a general model of identity management and self-inference mechanism
- Model of motivated belief

Model

- Let moral identity and similar concepts as beliefs about one’s deep “values”
- Emphasizes the **self-inference process** through which they operate
- The needs served by particular beliefs are linked to more basic aspects of preferences
  - “Demand side”: **affective benefits** (hedonic value of self-esteem, or anticipatory utility from one’s economicand social assets), **functional ones** (a strong moral sense of self that helps resist temptations), or both
  - “Supply side”: use imperfect memory or awareness, which emphasizes *identity investments* as self-signals = judge oneself by own behavior or decisions

Motivating Facts And Puzzles

Unstable Altruism

Individual-level

- Fairness, cooperation, and honesty in anonymous, one-shot interactions
- Excuse seeking, moral wiggle room,
- History dependence:
    - **Foot-in the door effect** : an initial request for a small favor (which most people accept) raises the probability of accepting costlier ones later on; similarly, a large initial request (which most people reject) reduces later willingness togrant a smaller one (see DeJong 1979).
    - **Moral credentialing**: acting as if an initial good behavior (again, exogenously induced) provided a license to misbehave later on
- Nonmonotonicity

Coexistence Of Social And Antisocial Punishments

Group-level

Free-riders in public-good games, and violators of social norms more generally, get punished by others (e.g., Fehr and G¨achter 2000)

Who behave too well—exhibiting stronger moral principles or resilience than their peers (objectors to injustice, vegetarians, and whistle-blowers) or contributing “excessively” to public goods—also elicit resentment, derogation, and punishment from their peers

Taboo Tradeoffs

- Not make utility tradepoffs; 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. people seek to enforce such taboos not only on others’ behavior (which could be accounted for by standard externalities) or **even on their** own(precommitment), **but even on their own thoughts and cognitions.**

Other Social Phenomena

- Prosocial behavior is the main focus of our paper, many other social phenomena involve beliefs which people treat as valuable assets. Religion is the most obvious one.
- Escalating commitments: 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. e.g. a higher stock raises the stakes on viewing the asset as beneficial to one's long-run welfare, and the way to reassure oneself of its value is to keep investing. This leads to excessive specialization (e.g., work versus family) and persistence in unproductive tasks
- Treadmill effect: increases in wealth, social status, or professional achievement induce a self-defeating pursuit of the belief that happiness lies in the accumulation of those same assets.
- Oppositional behaviors: When two identities are likely to compete later on for time or resources, investing in one depreciates the perceived value of the other. An agent with substantial capital vested in an insecure, hard-to-measure identity (e.g., cultural attachments) may therefore refrain from profitable investments in others (education, labor market integration), and even destroy valuable assets, ending up worse off.

THE MODEL

Timing of Moves and Actions

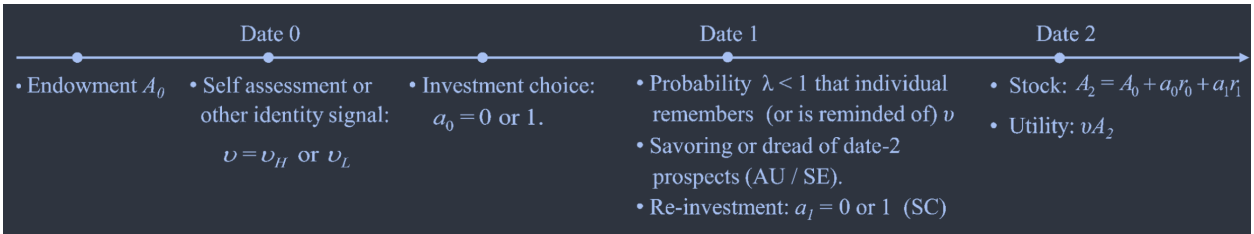


FIGURE I
Timing of Moves and Actions (for AU/SE and SC Specifications)

Preferences and Beliefs

A: accumulated assets

A is "relational assets" (human capital, wealth, status, religion specific good deeds, knowledge of a culture, etc.) and the individual's long-run utility for the benefits flowing from it

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

v: altruism or prosocial orientation: the extent to which the agent internalizes the welfare of others ; the individual's long-run utility for the benefits flowing from it

a_t: moral/immoral decision

$a_t = 1$ moral, prosocial or cooperative behavior

$a_t = 0$ immoral, selfish, or opportunistic behavior

r_t: the multiplier of moral decision， which measure the return to the individual's efforts in raising the welfare of those he cares about

One-shot interaction(no use to build social network): $r_t = 0$

But the behavior still release a signal of what he is (for self-inference)

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. Supply side of Motivated Belief

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 $1-\lambda$, $\hat{v} = \hat{v}(a_0) \in [v_L, v_H]$

💡 $(1-\lambda)$ 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

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, used for following AU case

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

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

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

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

long-term welfare still be vA_2 in date 2

Moral decision

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

continuation value increases in $\hat{v}$

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

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:

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

Mixed Case

Determine when anticipatory emotions alleviate or worsen the self-discipline problem

will be examined in Section VI

Interpreting the Model*

Tell us it is more applicable than the text suggest

Identity as Multidimensional

Identity is single dimentional in model

The model can represent a tradeoff between two dimensions A and B, such as morality and wealth, or family and career, linked by uncertainty over their relative value $v_A - v_B$ and a resource or time constraint on total investment.

A second type of identity conflict, arising from rivalry in consumption rather than investment, is analyzed in Section VI.C

Identity as a Social Object

In our main illustration, A_t corresponds to relationships with others and v to altruism or public-spiritedness. Other social aspects of identity may include agents' prior beliefs (ρ) and, critically, information flows within a reference group. Section VI.B will thus study people's responses to both norm-violators who fail to uphold a valued identity and “do-gooders” who uphold it too well

Self-Knowledge and Affirmation of Values

The assumption that people have imperfect insights into their own values and motives admits several formally equivalent interpretations:

1. A **moral sentiments view**, in which people experience guilt or pride not only when actually observed by others, but also from the **virtual judgements of "imagined spectators"** (Smith 1759).
2. An **ego-superego view**, in which **v is simultaneously known at the subconscious level and not known at the conscious level** (Bodner and Prelec 2003). This corresponds in the model to a limiting case of "instantaneous forgetting."
3. **Intergenerational transmission**. In this polar case "forgetting" takes a generation, so the date-0 agent is a parent and the date-1 agent his child. Parents have experience with the value of certain assets, such as the life satisfaction derived from social bonds versus money and career, or the benefits that religion might yield. Children start less informed and learn (with probability 1 − λ) from the example that their parents set, or from what they force them to do (a0) . Parents strive, altruistically or selfishly, to inculcate in their children "values" (beliefs $\hat{v}$) that will enrich their lifetime experience or lead them to take desirable actions.

Equilibrium And Welfare: Solving the model

Behavior

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

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

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 one, 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 $\hat{v}(1) > \hat{v}(0)$

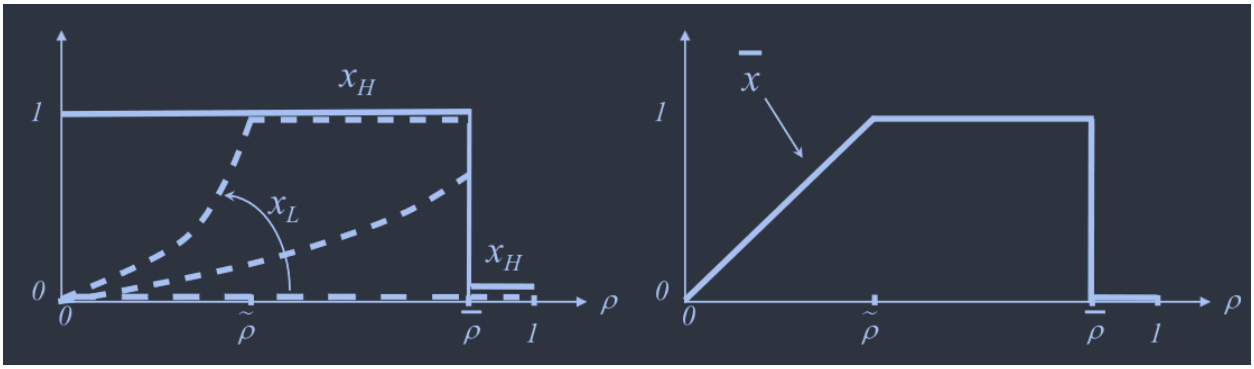
Monotonic Perfect Bayesian equilibria

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

Right panel: average investment $\bar{x}(\rho)$

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 $(0 < x_L < 1, \tilde{\rho} = \bar{\rho})$

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

Proof of Proposition 1.

The difference between the two types' incentives to invest in Equation (12) is

$$(24)\Delta \equiv \int_{v_L}^{v_H} \left[\int_{A_0}^{A_0+r_0} \mathbf{V}_{13}(x, \widehat{v}(1), z)dz + \int_{\widehat{v}(0)}^{\widehat{v}(1)} \mathbf{V}_{12}\left(x, y, A_0\right) dy \right] dx + c_0^L - c_0^H$$

If $V_{12} = 0$ (as with anticipatory utility) then $\Delta > 0$, so any equilibrium must have $x_L(1-x_H) = 0$. When $V_{12} > 0$ the same holds provided $\hat{v}(1) \geq \hat{v}(0)$, but since those beliefs are endogenous we must make monotonicity a requirement. The possible equilibrium configurations are then:

1. No investment: $x_H = x_L = 0$, hence, $\hat{v}(0) = \bar{v}$and $\hat{v}(1) = v_H$,with

$$(25)V(v_H, v, A_0) \geq V(v_H, v_H, A_0 + r_0) - c_0^H.$$

2. Randomization by v_H: $1 > x_H > x_L = 0$, hence, $\hat{v}(1) = v_H$ and $v_L < \hat{v}(0) < \bar{v}$, with

$$V(v_H, \hat{v}(0), A_0) = V(v_H, v_H, A_0 + r_0) - c_0^H.$$

3. Separation: $1 = x_H > x_L = 0$, hence, $\hat{v}(1) = v_H$ and $\hat{v}(0) = v_L$,with

$$(26)V(v_H, v_L, A_0) \leq V(v_H, v_H, A_0 + r_0) - c_0^H.$$

$$(27)V(v_L, v_L, A_0) \geq V(v_L, v_H, A_0 + r_0) - c_0^L.$$

4. Mixing by v_L: $1 = x_H > x_L > 0$, hence, $\hat{v}(0) = v_L$ and $\bar{v} < \hat{v}(1) < v_H$, with

$$(28)V(v_L, v_L, A_0) = V(v_L, \hat{v}(1), A_0 + r_0) - c_0^L.$$

5. Full investment $x_H = x_L = 1$, hence, $\hat{v}(0) = v_L$ and $\hat{v}(1) = \bar{v}$,with

$$(29)V(v_L, v_L, A_0) \leq V(v_L, \bar{v}, A_0 + r_0) - c_0^L.$$

We now show that there exists a unique equilibrium, which involves no investment when Equation (25) holds and, when this condition fails, separation, randomization by v_L or full investment, depending respectively on whether Equations (26)–(27), (28), or (29) hold.

If $V(v_H, v_L, A_0) \geq V(v_H, v_H, A_0 + r_0) - c_0^H$, it is a dominant strategy for both types not to invest, so $x_H = x_L = 0$ for all ρ, or equivalently $\bar{\rho} \equiv 0$.

Assume now that $V(v_H, v_L, A_0) < V(v_H, v_H, A_0 + r_0) - c_0^H$. Because $\bar{v} \simeq v_L$ for ρ small, the no-investment regime (1) cannot prevail for ρ small. More generally, it obtains if and only if $\rho \geq \bar{\rho}$,where $\bar{\rho} > 0$ is defined by

$$(30) V(v_H, \bar{\rho} v_H + (1 - \bar{\rho}) v_L, A_0) \equiv V(v_H, v_H, A_0 + r_0) - c_0^H$$

if this equation has a solution in $(0, 1)$ and to 1 otherwise. For $\rho < \bar{\rho}$ we have $x_H = 1$ from the previous taxonomy and the Pareto dominance assumption.

If Equation (27) holds, the equilibrium is separating: $x_H = 1$ and $x_L = 0$. By contrast, if $V(v_L, v_L, A_0) < V(v_L, v_H, A_0 + r_0) - c_0^L$,the L type must invest with positive probability. If Equation (29) holds there can be nosolution toEquation (28) with $x_L < 1$, sothe only equilibrium is full investment on $[0, \bar{\rho})$. If Equation (29) is reversed, on the other hand, it involves mixing: by Equation (10),

$$(31) \hat{v}(1) = \frac{\rho}{\rho + (1 - \rho) x_L} v_H + \frac{(1 - \rho) x_L}{\rho + (1 - \rho) x_L} v_L,$$

and by Equation (28) this expression must be independent of ρ.

Thus, $x_L = (\gamma - 1)/(1/\rho - 1)$, where $\gamma \equiv 1/\hat{\rho}(1) > 1$ is also a constant. If $(\gamma - 1)/(1/\bar{\rho} - 1) < 1$, then the L type mixes over all of $[0, \bar{\rho}]$; if $(\gamma - 1)/(1/\bar{\rho} - 1) \geq 1$, define $\tilde{\rho}$ by $(\gamma - 1)\tilde{\rho}/(1 - \tilde{\rho}) \equiv 1$ or, equivalently,

$$(32) V(v_L, v_L, A_0) = V(v_L, \tilde{\rho} v_H + (1 - \tilde{\rho}) v_L, A_0 + r_0) - c_0^L,$$

implying $\tilde{\rho} > 0$ by Assumption 4. Then $x_L \in (0, 1)$ for $0 < \rho < \tilde{\rho}$ and $x_L = 1$ for $\rho \geq \tilde{\rho}$.

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.

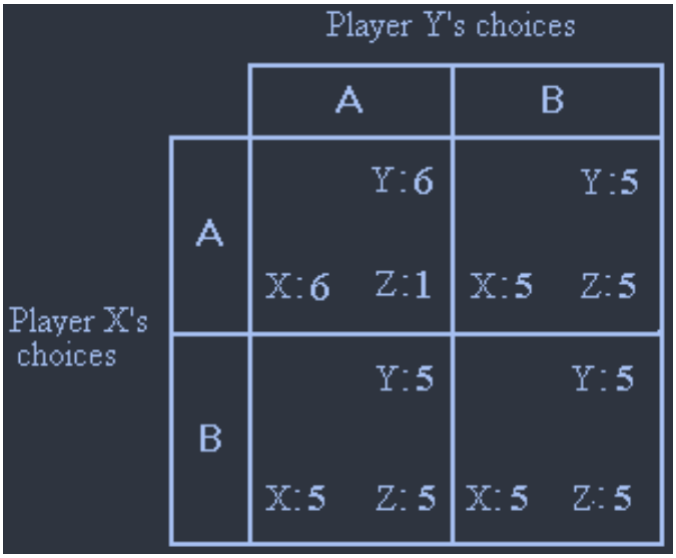
Implications and Evidence on Moral Identity and Behavior

Malleability of Beliefs

If the individual have better ability to know his true preferences and motives(lower λ), self-identity management is more likely to occur.

moral wriggle room / self-concept maintenence theory / Delegation

= presence and power of self-deception



Salience of Identity

Mazar, Amir, and Ariely (2008). honesty promoted after read the Ten Commandments

consumption on “symbolic” goods such as carbon offsets, green products, largely spurred by advertising campaigns that manipulate the salience of people’ self (and social) image.

The fact that most of the same households vote against environmental taxes, together with experiments documenting the moral-licensing effects of green purchases (Mazar and Zhong

2010), provides further support for the idea that such expenditures are in large part identity investments.

Uncertain Values

- Overall (ex-ante) probability of investment $\overline{x}$ is hill-shaped with respect to ρ

  investing in self-reputation has a low return when the prior is low, and is not needed when it is already high

1. Identity-affirming behaviors are characteristic of people with unsettled preferences and values; hence the moral zeal of the newconvert (religious or political), or the exacerbated nationalism of the recent immigrant
2. The predicted hill-shape of behavior with respect to ρ can help reconcile two contradictory sets of experimental findings on people's responses to manipulations of their self-image
   a. *Threats to a strongly held identity*. "transgression-compliance" effct: subjects who are led to believe that they have harmed someone (by administering painful electric shocks, or carelessly ruining some of her work) show an increased willingness to later on accept requests to perform a good action
   b. *Manipulating weaker aspects of identity*. when ρ changes marginally, starting from below ˜

      "foot in the door" effect

Identity and Welfare: Treadmill Effect or Empowerment

Self-esteem/Anticipatory Utility and the Treadmill Effect

Based on $V\left(v,\hat{v},A_1\right)\equiv(s\hat{v}+\delta v)A_1$ and $W\equiv E\left[-a_0c_0+V\right]$

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

> 💡 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

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

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

Turning now to the direct contribution of date-0 behavior to intertemporal welfare, if β is low enough that (say) the first two terms in Equation (15) are positive, ex-ante efficient investments fail to occur in period 0 if λ = 1: from the very start, the agent behaves too opportunistically for his own good. The ability to affect his self-image (λ < 1) provides additional motivation for acting prosocially at t = 0, which then directly raises ΔW. When the first two terms in Equation (15) are positive, conversely, such good behavior entails a net cost, which only pays off in terms of improved self-restraint at t = 1 if E[ΔV] sufficiently positive.

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

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

> "To compare is to destroy." (Fiske andTetlock 1997)

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: he 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 $a_0 = 1$

$V(a_0 = 1) = \mathbf{V}\left(v, \hat{v}(1), A_0\right)$

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

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

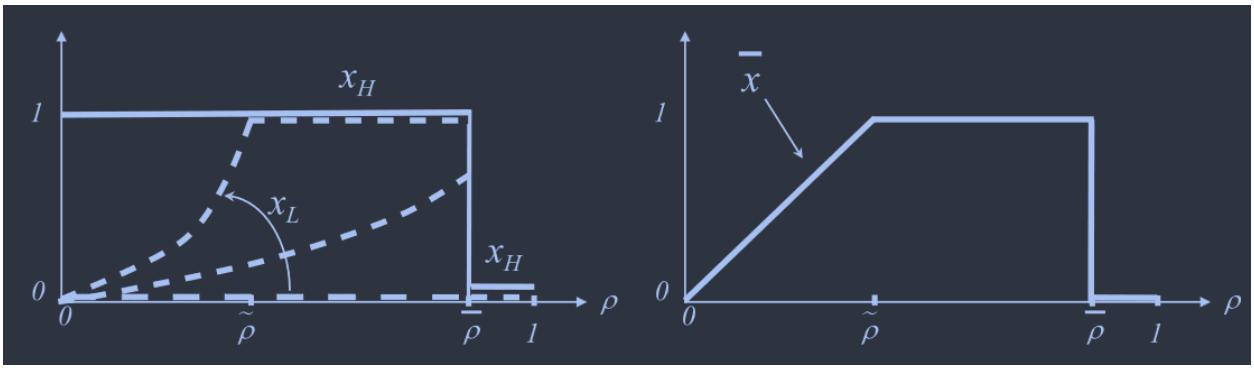
$\mathbf{V}\left(v, \hat{v}(0), A_0 - z\right)$ can be written because V is linear in A_1 in AU and SC case

Positive Side of Conclusions

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" (Figure II, right panel) the induced erosion of ρ occurs on

on the right side: ρ decrease → reaffirmation

on the left side: ρ decrease → collapse

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.

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

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."

Dealing with Sinners and Saints: Information Destruction

Socially enforced taboos: understand the *coexistence of both social and antisocial punishments*

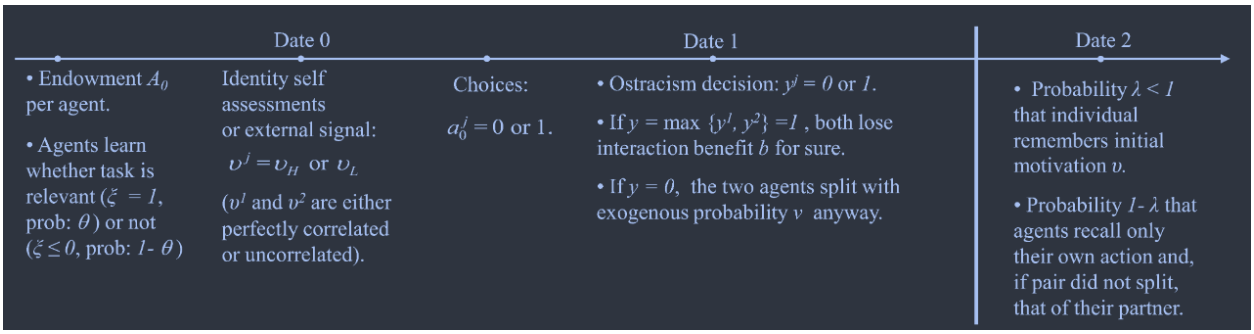
benchmarking idea

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

The Person and the Situation



New elements:

1. **Action of benefit society or not**(a_0).
    a. with ex-ante probability θ, agent have an "excuse" for deviate from the group

    in former case = 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 $a_0 = 0$

    b. 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.

2. **Action of exclusion from group or not**$\left(y_i\right)$

   two agents after observing each other's action, decide whether to continue in the relationship $\left(y_i = 0\right)$ or to break it $\left(y_i = 1\right)$

   if someone exit, both lose b as future interactions benefit

3. **Agent i's utility function**

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

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

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 $\left(\xi \equiv 1\right)$

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

Benchmarking on the Situation

The two types are **independent**

Social usefulness (ξ = 1, with probability θ) or absence(ξ ≤ 0, with probability 1 − θ) 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 − λ.

Focus on symmetric equilibria (in undominated strategies) in which the more altruistic type always invests when ξ = 1 and no one invests when ξ ≤ 0 (either ξ is sufficiently negative, or ξ = 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.

1. 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

2. Social punishment and Anti-social punishment

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

   ostracism comes from the unvirtuous one $\left(a_0^i = 0\right)$ 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

3. 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:

   a. When benchmarking is on the person, x = 1 is sustained by the ostracism of “sinners” (a prosocial norm), while x = 0 involves no ostracism

   b. 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

Other Dimensions of Identity

1、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

2、Uncertain Values and Malleability of Beliefs

People are insecure about “who they are” (ρ 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.

3、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 $\hat{v}\prime$ of their type, resulting in a continuation value of the form $V(v,\hat{v},A_1,\hat{v}\prime)$ 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 (11) is now

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

Thus, as long as

$$(v,\hat{v},A_1)\longmapsto V(v,\hat{v},A_1,\hat{v})$$

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

$$(1-\lambda)V_2 to (1-\lambda)V_2+V_4)$$

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

Competing Identities and Dysfunctional Behavior

Tradeoff between the future benefits from two identities, investing in one (say, B) inevitably damages the other (A), as it suggests that the individual may not value it that much.

1、Resistance to Structural Change

The transition, which is risky and requires new skills and lifestyles, will be resisted if it is seen as de-valuing the old (rural, extended-family, blue-collar, etc.) identity

2、Resistance to Assimilation

Immigrants and their descendents experience strong tensions between integrating into Western societies and preserving their specific culture.

3、Destructive Identity, Discrimination, and Communitarianism

Not investing in B in order to safeguard beliefs about the value of A can also mean actively destroying productive B capital. E.g young rioters attacked and destroyed a number of schools, nursery schools and cars in their own communities.

Conclusion

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

The paper proposed the monotonic Perfect Bayesian equilibria of welfare with three scenarios.

Taboos can be formed by internally enforced and socially enforced

High endowments trigger escalating commitment and a treadmill effect

Competing identities can cause dysfunctional capital destruction

Questions

How might you modify this model to incorporate？

Banerjee herd behavior problem. Taboo enforcement may be strengthened by herd behavior. Principal-agent problems.

Expert Recommandation problem. A expert persuade investors to buy a risky but high outcome lottery because he will earn from the company, but actually expert self-deceived that everybody is risk-seeking(Gneezy et al., 2020). We should add a_3 or modify the a_3.

As people age and gain experience, presumably λ increases. However, it is not obvious that young adults are less pro-social than older adults.

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 (Bucciol & Piovesan, 2011).