Identity, Morals and Taboos:

Beliefs as Assets[1]

Roland Bénabou[2] Jean Tirole[3]

This version: June 2010

Forthcoming in the *Quarterly Journal of Economics.*

[1]An earlier version of this paper was titled "Identity, Dignity and Taboos: Beliefs as Assets". We are grateful for helpful comments and suggestions on this project to Andrew Caplin, Ed Glaeser,Yuemei Ji, Robert Oxoby, Philipp Sadowski and Glen Weyl, as well as to three referees and participants at several conferences and seminars. Bénabou gratefully acknowledges support from the National Science Foundation and the Canadian Institute for Advanced Research.

[2]Princeton University, NBER, CEPR, CIFAR and IZA.

[3]Toulouse School of Economics

Abstract

We develop a theory of moral behavior, individual and collective, based on a general model of identity in which people care about "who they are" and infer their own values from past choices. The model sheds light on many empirical puzzles inconsistent with earlier approaches. Identity investments respond nonmonotonically to recent acts or threats, and taboos on mere thoughts arise to protect beliefs about the "priceless" value of certain social assets. High endowments trigger escalating commitment and a treadmill effect, while competing identities can cause dysfunctional capital destruction. Social interactions induce both social and antisocial norms of contribution, sustained by respectively shunning free riders or do-gooders.

Keywords: identity, morals, prosocial behavior, taboos, self-image, wishful thinking, memory, anticipatory utility, self control, hedonic treadmill, religion.

JEL numbers: D81, D91, Z13.

"Man naturally desires... not only praise, but praiseworthiness; or to be that thing which, though it should be praised by nobody, is, however, the natural and proper object of praise. He dreads, not only blame, but blame-worthiness; or to be that thing which, though it should be blamed by nobody, is, however, the natural and proper object of blame..."

"When I endeavour to examine my own conduct, when I endeavour to pass sentence upon it, and either to approve or condemn it, it is evident that, in all such cases, I divide myself, as it were, into two persons: and that I, the examiner and judge, represent a different character from that other I, the person whose conduct is examined and judged of.... The first is the spectator, whose sentiments with regard to my own conduct I endeavour to enter into, by placing myself in his situation, and by considering how it would appear to me, when seen from that particular point of view. The second is the agent, the person whom I properly call myself, an of whose conduct, under the character of the spectator, I was endeavouring to form some opinion."

(Adam Smith, *The Theory of Moral Sentiments,* p. 151-152)

Introduction

From charitable donations to experimental games, there is by now ample evidence that people often behave "morally" even in anonymous, one-shot interactions. This has justifiably dispelled an excessively narrow view of economic man, but the standard replacement –human beings endowed with various forms of "social preferences"– is still a highly unreliable guide for understanding the vicissitudes of (im)moral behavior. If good deeds stem from altruistic tastes, why do the same people often seize upon (even actively seek) the most transparent change in framing, the thinnest of veils to revert to selfishness? If punishing cheaters and free-riders reflects a taste for fairness or reciprocity, why do we also see people turning on those who behave too well? This "flickering" nature of moral behavior (see Section I for detailed evidence) makes clear that more is at work than socially enriched utility functions. No such preferences, moreover, can account for information-averting behaviors, such as people prohibiting *themselves* from merely *thinking* about certain "taboo tradeoffs".

We develop in this paper a "third-generation" theory of moral behavior, based on a general model of identity management. The theory is cognitive, in that it explicitly models moral identity and similar concepts as *beliefs* about one's deep "values" and emphasizes the *self-inference* process through which they operate. At the same time, the *needs served* by particular beliefs are linked to more basic aspects of preferences. This "demand side" can reflect a quest for affective benefits (hedonic value of self-esteem, or anticipatory utility from one's economic and social assets), functional ones (a strong moral sense of self that helps resist temptations), or both.[1] On the "supply side" of motivated beliefs, the pivotal role is played by imperfect

[1] The "demand side" of our framework thus unifies models based on a consumption value of beliefs (Akerlof

memory or awareness, which naturally gives rise to *identity investments* as self-signals: because people have better, more objective access to the record of their conduct than to the exact mix of motivations driving them, they are led to judge themselves by what they do.[2] When contemplating choices, they then take into account what kind of a person each alternative would "make them" and the desirability of those self-views –a form of rational cognitive dissonance reduction.[3]

Many puzzling aspects of (im)moral behavior become much easier to understand from a self-reputational perspective. First, being linked to imperfect self-knowledge, identity-enhancing behaviors are more likely when objective information about deep preferences is scarce (true generosity, loyalty or faith) and they are easily affected by minor manipulations of salience such as cues, reminders and transparent veils of personal responsibility (see Section I). Most importantly, investments in one's self-view are hill-shaped with respect to prior confidence in being a moral person. This implies *history-dependence* in behavior with a distinctive, *non-monotonic* pattern of responses to manipulations that helps reconcile many divergent experimental findings. We thus show that whereas challenges to a weakly held identity (low prior) elicit *conformity* effects, effective challenges to a strongly held one (high prior) elicit forceful *counterreactions* aimed at restoring the threatened beliefs.

Second, since the preferences and prospects of similar individuals are likely to be correlated, "deviant" behavior by peers –violating norms and taboos, fraternizing with outsiders, etc.– conveys bad news about the value of existing social assets (anticipatory-utility version) or that of future investments in them (imperfect self-control version). If the morally dubious action was one's own, on the other hand, it is good behavior by peers that is now threatening to the self-concept, as it takes away potential excuses involving situational factors or moral ambiguity. In both cases, ostracizing mavericks suppresses the undesirable reminders created by their presence. Thus, depending on the perceived situational uncertainty and correlation of individual values, the same agents will act prosocially and shun free riders, or act selfishly and shun moral exemplars.

Finally, our cognitive model naturally generates *taboo tradeoffs* and an aversion to engage in even the *mere contemplation* of such choices. We show that upholding certain (endogenously) valuable beliefs or illusions concerning the "incommensurable" value of certain goods, or the

and Dickens [1982], Loewenstein [1987], Rabin [1995], Caplin and Leahy [2001], Landier [2000], Brunnermeier and Parker [2005]), Köszegi [2004]) and those in which they serve a more instrumental role (Carrillo and Mariotti [2000], Bénabou and Tirole [2002, 2006a], Battaglini et al. [2005], Dessi [2008]). In particular, we show that these two classes of models lead to very similar behaviors but potentially opposite welfare consequences.

[2] See, e.g., Festinger and Carlsmith [1959] on cognitive dissonance and, especially, Bem [1972] on self-perception. On the self-manipulation of "diagnostic" actions see Quattrone and Tversky [1984].

[3] The idea of self-signaling or self-reputation makes the paper most closely related to Bodner and Prelec [2003] and Bénabou and Tirole [2004]. Young [2006] and Dal Bo and Terviö [2008] extend the single-agent analysis to infinite horizons and steady-states, and Battaglini et al. [2005] and Bernheim and Thomadsen [2005] to one-shot, strategic interactions with simultaneous moves. None of these papers deals with the empirical puzzles discussed in Section I, nor with taboos, group norms or (anti)social sanctions.

things one "would never do" (various forms of selling out) can require shunning any evaluation, in act or in thought, that might reveal what terms of trade could be obtained or would be accepted.

While 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, but significant resources are similarly invested to build up and defend national, cultural and even professional identities. Our model therefore provides a unifying framework for the study of *identity,* and in the last section of the paper we demonstrate its applicability across a wide range of behaviors.[4]

The model thus explains *escalating commitments,* in which 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. Intuitively, 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. Most strikingly, one can even be made worse off by a higher capital stock, as the escalating-commitment mechanism leads to a *treadmill effect* in which 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. The model also sheds light on *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.

While the model's positive results are quite general, the welfare consequences of the quest for moral identity and other self-views, in contrast, depend importantly on whether the "demand" side reflects mental-consumption motives (self-esteem, anticipatory utility) or instrumental ones (self-discipline, sense of direction). In the first case, identity investments reduce an individual's ex-ante welfare, being *in fine* a form of wasteful signaling. As a consequence, he is worse off with malleable beliefs or memory than with non-manipulable ones. When identity serves a commitment purpose, by contrast, more malleable beliefs and the resulting ability to shape them through actions can, under specific conditions, increase welfare.

[4]The paper naturally relates to the growing literature on the economics of identity. In an influential set of papers, Akerlof and Kranton [2000, 2002, 2005] emphasize how, in a wide range of contexts, agents' preferences are structured by their choices of a social category (see also Shayo [2009] on redistributive politics and Basu [2006] on development). Greif [2009] models moral behavior based on similar self-categorization and preference externalities, but with standards of conduct now strategically defined by "moral authorities" external to the group. In Rabin [1994], Konow [2000] and Oxoby [2003, 2004], agents engage in "dissonance reduction", again represented by costly adjustments in utility parameters not tied to an information structure. By explicitly modeling the value and management of beliefs our model endogenizes the identity prescriptions, payoffs and cognitive costs in this broad class of models. This also leads to distinctive results such as non-monotonicities, information-aversion and the fact that being able to manage his self-image can make a person worse off.

I Motivating facts and puzzles

Decisions on contributing to a public good, cooperating with others or enforcing a collective norm –in short, moral behavior– exhibit important inconsistencies with standard models of socially-minded behavior. These recurrent patterns can be categorized into three main puzzles: unstable altruism, coexistence of social and antisocial punishments, and taboo tradeoffs.

Unstable altruism. Prosocial behaviors in anonymous, one-shot interactions, where concerns for social reputation are inoperative, are often taken to directly reflect the extent of altruistic, reciprocal or other fairness-valuing preferences in a population (e.g. Fehr and Schmidt [1999]). More recent findings however, show that it takes remarkably little to turn such behaviors on or off. The slightest change in framing, the thinnest of veils as to the moral implications of their choices suffices for many people to revert to self-interest. In fact, they will not just seize upon such excuses and superficial ambiguity but actively seek them, foregoing to do so both material payoffs and decision-relevant information. For instance, when decision-makers can avoid finding out whether taking a high payoff for themselves will hurt or benefit someone else, over half take advantage of this "moral wriggle room" to behave selfishly (Dana et al. [2007]). Similarly, many subjects will take \$9 rather than having \$10 to freely allocate between themselves and an anonymous recipient, and the more likely to use such costly exit options are in fact those who share the most when no opt-out is possible (Dana et al. [2006], Lazear et al. [2009]).[5] Conversely, trivial cues making morality more salient, such as paying for performance in hard cash rather than tokens redeemable for money, or reading the Ten Commandments at the start of an experiment, dramatically increase cooperation and decrease cheating (Mazar et al. [2008]).

Two related forms of behavioral instability are history-dependence and non-monotonicity. When a person has been induced to behave prosocially or selfishly, or just provided with signals presumed to be informative about his morality, his choices in subsequent, unrelated interactions are significantly affected. Moreover, this reaction sometimes amplifies the original manipulation, and is sometimes in opposition to it. The well-known "foot-in the door" effect, for instance, documents how 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 to grant a smaller one (see DeJong [1979]). Yet, in different settings the same subject pools display "moral credentialing", acting as if an initial good behavior (again, exogenously induced) provided a license to misbehave later on (Monin and Miller [2001]). Similarly, people offered an opportunity to purchase "green" products tend to respond positively, but those who do buy are later on less likely to share in a dictator game, and more likely to cheat on a task to increase their gains (Mazar and Zhong [2010], Zhong et al. [2010]). Such reversals mirror earlier findings on the "transgression-compliance" effect, in which people

[5] When given the opportunity, many people will also delegate a sharing decision to a third party likely to be biased in their favor rather than do the "dirty deed" themself (Hamman et al. [2009]).

led to believe that they have harmed someone show later on an increased willingness to perform unrelated good deeds (Carlsmith and Gross [1969]).

Social and antisocial punishments. Turning from single-agent settings to groups, one encounters similar inconsistencies in behavior and judgement. On the one hand, it is well established that free-riders in public-good games, and violators of social norms more generally, get punished by others (e.g., Fehr and Gächter [2000]). On the other, there is growing evidence that those who behave too well –exhibiting stronger moral principles or resilience than their peers (objectors to injustice, vegetarians, whistle-blowers) or contributing "excessively" to public goods– also elicit resentment, derogation and punishment from their peers (Monin [2001], Jordan and Monin [2008], Monin et al. [2008]). Such moral relativism is not confined to the laboratory but also reflected in cross-society-differences in civic norms. In countries where subjects in public-goods experiments engage in more social (antisocial) punishment, surveys also show citizens to be less (more) tolerant of cheating behaviors such as tax evasion or welfare fraud and more (less) trusting in other people's adherence to the rule of law (Herrman et al. [2008]).

Taboos thoughts and tradeoffs. Whereas economics views all goods as fungible, that is, subject to tradeoffs, most societies and cultures hold certain ones to be "priceless" or sacred: life, justice, liberty, honor, love, religious faith, etc. (e.g., Durkheim [1925]). It is thus considered highly immoral to place a monetary value on marriage, friendship or loyalty to a cause. Markets for organs, genes, sex, surrogate pregnancy and adoption are widely banned on grounds that they would represent an unacceptable "commodification" of human life. Admittedly, such rules are often observed in the breach, and the boundaries between the secular and the sacred are evolving ones, as demonstrated by changing attitudes toward life insurance (Zelizer [1999]), pollution permits, or, in certain places, legalized prostitution. Nonetheless, taboos often do bind, removing a number of activities from the traditional economic sphere or confining them to black markets (see, e.g., Kanbur [2004], Roth [2007]).

Most puzzling is the fact that people seek to enforce such taboos not only on others' behavior (which could be accounted for by standard externalities) or even on their own (which might reflect a desire for precommitment), but even on *their own thoughts* and cognitions. Many experiments document this "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 (Fiske and Tetlock [1997], Tetlock et al. [2000]). Their view, and that of the law in most countries, is that certain transactions are so "contrary to human dignity" that they would affect even those who simply know or speculate about them, by inviting "morally corrosive" thoughts of fungibility. Yet what exactly is being corroded by placing a hypothetical monetary value on certain goods or activities, and how this damage occurs, is never really spelled out.

The patterns of behavior described in this section are not easily accounted for by existing models. The choices of agents with altruistic, joy-of-giving, reciprocity or fairness concerns will

(under anonymity) consistently reflect these stable preferences, and not exhibit the "Jekyll and Hyde" reversals and path dependencies commonly seen in both lab and field. Similarly, agents with social-identity motivations or other group-based preferences may engage in costly acts of repair when their identity has been challenged or damaged, but will typically not show "licence" to misbehave when it has been affirmed, nor the kind of amplification exemplified by the "foot-in-the-door" effect. They may ostracize and punish those who violate group norms but not turn against those who are the best exemplars of their chosen identities. Finally, in none of these models will agents ever exhibit information avoidance, such as is consistently observed in "moral wriggle room" experiments and in the phenomenon of taboo thought.

The body of the paper is organized as follows. Section II presents the model and Section III the main propositions, with moral identity and decisions as the leading application. Section IV analyzes sacred values and taboo tradeoffs, then turns to the mechanism underlying both social and antisocial norms. Section V gathers extensions of the model and applications to other realms of behavior. Section VI offers directions for further research. Proofs are gathered in Appendix A, details on the model's extensions in Appendix B.

II The Model

"An identity is a definition, an interpretation, of the self... People who have problems with identity are generally struggling with the difficult aspects of defining the self, such as the establishing of long-term goals, major affiliations, and basic values." (Baumeister [1986]).

A Preferences and beliefs

There are three periods, $t = 0, 1, 2$, as illustrated in Figure I. An individual starts with initial endowment A_0 of some asset; from the final stock A_2 he will derive long-run welfare vA_2. In our main application, A corresponds to social relationships and v reflects the extent to which the agent internalizes the welfare of others.[6] We shall accordingly refer to A as "relational capital" and to v as altruism or prosocial orientation.[7]

[6] More generally, A can be any asset (human capital, wealth, status, religion-specific good deeds, knowledge of a culture, etc.) and v the individual's long-run utility for the benefits flowing from it. See Section V on these other dimensions of identity.

[7] The specific form of the interactions involved is inessential for our purposes, but examples might be useful. In the simplest one, A_t equals the agent's cumulated contributions to other people's welfare and v is how much he cares about their well-being. Alternatively, let A_t be the number of people willing to engage with him in a repeated-prisoner's-dilemma type of interaction that starts at $t = 2$. Someone who places more weight on others' payoffs is less likely to cheat down the road, causing a break-up of the relationship and a loss of its benefits to both parties; he will thus value each social bond more highly. Partners' investment in a relationship could also be strategic complements, as in Rotemberg (1994). All that matters for our study of moral identity is that social relationships be representable as assets (fixed or accumulable, see (1)) with a continuation value, such as vA_2, that is higher for more prosocial individuals (see Assumption 3 more generally).

At dates $t = 0, 1$, the individual can "invest" ($a_t = 1$), with return $r_t \geq 0$, or "not invest" ($a_t = 0$), so that

$$A_{t+1} = A_t + a_t r_t. \tag{1}$$

Thus, by helping others, cooperating and contributing to public goods, he can enhance existing relationships (raise the utility of people he cares about) or establish new ones (make friends, gain productive partners). By behaving badly he will fail to increase A_t, or could even erode it; we normalize (1) to measure the relative increase from choosing $a_t = 1$. We shall refer to $a_t = 1$ interchangeably as moral, prosocial or cooperative behavior, and to $a_t = 0$ as immoral, selfish or opportunistic behavior.[8]

The "investment" action will in fact play a dual role. The first is standard accumulation, when $r_t > 0$. The second is informational: even when the current decision has no impact on relational capital ($r_t = 0$ for one-time encounters, tipping, etc.), the individual's behavior will constitute a signal of his altruism.

The central ingredient in the model is indeed that people are, at times, unsure of their own deep preferences: moral standards, concern for others, strength of faith, etc. Such uncertainty over "long-term goals, major affiliations, and basic values" (Baumeister) means that the capital stock A_2 from which an individual will eventually derive benefits may prove to be very important to his long-run welfare, or not that meaningful.

• *Date 0*. At the start of period 0 the agent has access to a signal about his type, which may be one of high or low altruism, H or L. Through an instinctive feeling of empathy, a temptation to cheat or a conscious self-assessment, he obtains a momentary insight into his true nature,

$$v = \begin{cases} v_H & \text{with probability } \rho \\ v_L & \text{with probability } 1-\rho \end{cases}, \tag{2}$$

with $v_H > v_L$ and $\bar{v} \equiv \rho v_H + (1-\rho) v_L$ denoting the prior expectation. 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, etc.

Assumption 1 *The net cost of investment at date* 0 *is* $c_0^H \gtrless 0$ *for type* H *and* c_0^H *for type* L, *with* $c_0^L \geq c_0^H$.

• *Date 1.* The standard assumption in economics is that people gain, through experience, better knowledge of their preferences. For a person's past actions to define his sense of identity, however, it must be that he no longer has direct access to the deep motives and feelings that

[8] The specific interaction involved is, again, not essential. Thus, r_t can measure the return to the individual's efforts in raising the welfare of those he cares about. Alternatively, it can capture, in reduced form, the average propensity of partners to stay in the relationship, depending on how he has treated them.

gave rise to these choices –an information *loss.* Otherwise, past behavior conveys no useful information, so there is no sense in which one can make (or claim to make) choices intended to "be true to myself," "maintain my integrity," "keep my self-respect", "not betray my values", "be able to look at myself in the mirror," and the like. There is indeed extensive evidence that people's recall of their past feelings and true motives is highly imperfect and self-serving, that they judge themselves by their actions and that many decisions are shaped by a concern to achieve or maintain a desirable self-view, particularly in the moral domain.[9]

Assumption 2 *(Self-inference). At date* 1, *the individual is aware (or reminded) of his true valuation* v *only with probability* λ. *With probability* $1-\lambda$, *he no longer recalls (has access to) it and uses instead his past choice of* a_0 *to infer his type.*

Let us denote by $\hat{\rho}$ the individual's date-1 belief about "what kind of a person" he is and by

$$\hat{v} \equiv \hat{\rho} v_H + (1-\hat{\rho}) v_L \tag{3}$$

the corresponding expected valuation of A_2, either of which defines his (subjective) "sense of identity" at $t=1$. With probability λ the posterior $\hat{v}$ is thus equal to the original signal v, and with probability $1-\lambda$ it is equal to the conditional expectation $\hat{v}(a_0) \in [v_L,\, v_H]$ formed on the basis of previous behavior. More generally, $1-\lambda$ should be thought of as the *malleability of beliefs through actions,* and 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.[10]

This process of *self-inference* can be thought of as the "supply side" of motivated beliefs in the model. We next turn to the "demand side," which encompasses most mechanisms that make certain self-views more desirable than others. These include pure self-regard, anticipatory utility and imperfect self-control, all of which can be represented by a continuation value $V(v, \hat{v}, A_1)$, evaluated at $t=0$, of entering period 1 with beliefs $\hat{v}$ and capital A_1.

Assumption 3 *The value function* $V = V(v, \hat{v}, A_1)$ *satisfies* $V_2 > 0$, $V_{12} \geq 0$ *and, if* $r_0 > 0$, $V_{13} > 0$.

---

[9] On imperfect retrospective and prospective access to feelings and desires, see Kahneman et al. [1997] and Loewenstein and Schkade [1999]. On self-perception and self-signaling, see footnote 2, Bodner and Prelec [2003] and Bénabou and Tirole [2004]. Decisions problems with (exogenously) imperfect recall but no demand for motivated beliefs were first studied in Piccione and Rubinstein [1993].

[10] If an action is uninformative with probability ν, the posterior $\hat{v}$ equals v, $\bar{v}$ or $\hat{v}(a_0)$ with respective probabilities λ, $(1-\lambda)\nu$ and $(1-\lambda)(1-\nu)$, so the effect on signaling incentives is similar to that of a decrease in $1-\lambda$. For a model of self-reputation with misremembered actions and excuses, see Bénabou and Tirole [2004]. The recall or awareness probability could also be different for good and bad signals, $\lambda_H \geq \lambda_L$, whether exogenously or endogenously (see Bénabou and Tirole [2002]). We focus here on the case in which $\lambda_H = \lambda_L$, both for simplicity and to highlight the role of self-inference, which seems most relevant to "identity".

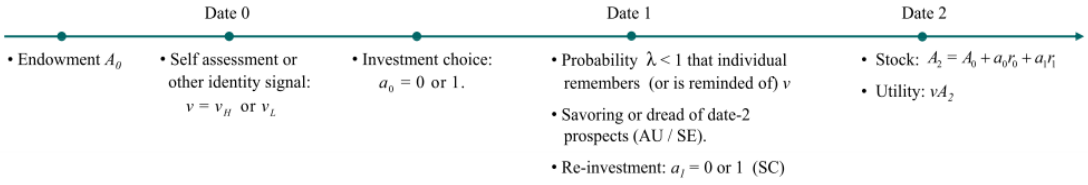

Figure I: Timing of Moves and Actions (for AU/SE and SC specifications)

The first condition is mainly a "good identity" convention: thus, a self-image of high morals and concern for others is better than one of low morals and selfishness.[11] The cross-partial restrictions, together with $c_0^H \leq c_0^L$, will generate a sorting condition leading the H type to always invest at least as much as the L one (behaving more prosocially), so that actions have informational content. We exclude the trivial case where both types always invest, regardless of identity concerns:

Assumption 4 $V(v_L, v_L, A_0 + r_0) - V(v_L, v_L, A_0) < c_0^L$.

The two "canonical" examples of preferences leading to motivated beliefs are discussed below and summarized in Figure I.

• *Demand for beliefs 1*: *self-esteem (SE) or anticipatory utility (AU).*

Someone who cares intrinsically about being, or having been over their lifetime, "the natural and proper object of praise [or] blame" (Smith [1759]) has preferences given by $V = s\hat{v}$, where s measures the strength of the self-esteem motive and $\hat{v}$ is given by (3).[12]

Closely related to self-esteem but more consequentialist in nature are anticipatory emotions –feelings of hopefulness, anxiety or dread that arise from contemplating one's future material and social prospects. Let long-term welfare be vA_2, the expected value of social relationships: family, friends, colleagues, ethnic group, etc. During period 1, an individual with subjective prosocial identity $\hat{v}$ experiences a utility flow $s\hat{v}A_2$, where the "savoring" parameter s reflects both the intensity of such anticipatory feelings and their duration.[13] Another important determinant of s is *salience* –the extent to which the individual thinks (perhaps prompted by an experimenter or advertiser) about the contribution of A_2 to his future welfare**,** and how it depends on where his true values really lie.

[11] Furthermore, it will only be used to select the Pareto-dominant equilibrium in the case of multiplicity.

[12] Formally equivalent is a Calvinistic concern for being among the "chosen" who are predestined for salvation and given the ability to engage in virtuous work, rather than among the "reprobates" who are irredeemably abandoned to sin and damnation.

[13] The hedonic value of period-1 beliefs could also be nonlinear in probabilities (equivalently, in $\hat{v}$). Our *positive* results (Propositions 1 and 2) apply unchanged such cases, as long as Assumption 3 is satisfied. Propositions 3 and 4 show, on the other hand, that *normative* conclusions do depend on linearity or the specific form of nonlinearity.

For simplicity, we focus here on pure anticipatory utility, in which there is no further decision to be made at date 1.[14] Thus $a_1 \equiv 0$, $A_2 = A_1$ and the continuation value (evaluated from $t = 0$) of entering period 1 with subjective identity $\hat{v}$ is

$$V(v, \hat{v}, A_1) \equiv (s\hat{v} + \delta v) A_1, \tag{4}$$

where δ is the discount factor between dates 1 and 2.[15] Assumption 2 is clearly satisfied, with $V_{13} > 0$, $V_{23} > 0$ and $V_{12} = 0$.

Note also that *self-esteem is a special case* of anticipatory utility with $A_t \equiv 1$ (the only relationship the agent cares about is with himself), $r_t \equiv 0$ and $\delta = 0$ (no "day of reckoning"). Accordingly, we shall study and refer to them together as the SE/AU case.

For welfare analysis, our criterion will be total intertemporal utility

$$W \equiv E[-a_o c_0 + V], \tag{5}$$

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 (SC)*

People with self-esteem concerns or anticipatory emotions about the value of their social assets want to hold certain beliefs for purely *affective* reasons. Maintaining a strong, stable sense of identity also has *functional* value, helping one to make consistent choices and resist harmful temptations. This adaptive role, equally stressed by psychologists, leads to our second benchmark case. It is particularly important in 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.

Let long-term welfare still be given by vA_2, but with moral decisions now taking place both at $t = 0$ and at $t = 1$. Investment at $t = 1$ involves a stochastic cost c_1, with type-independent distribution $F(c_1)$ on R_+.[16] At date 1, moreover, weakness of will can make the immediate gains from opportunism more salient than its distant consequences. The individual's "Self 1" thus perceives the cost of acting morally as c/β, where

$$\beta < v_L/v_H. \tag{6}$$

[14] This restriction is relaxed in Appendix B.

[15] Note also that s/δ reflects also the relative lengths of periods 1 and 2. Any discounting between periods 0 and 1 is implicitly embodied as a common factor in s and δ.

[16] Both assumptions are made for simplicity. The role of uncertainty over c_1 is only to smooth over $t = 1$ decisions, so as to make V differentiable. The type-independence of $F(\cdot)$ can also be relaxed, as long as realizations of c_1 are imperfectly informative about v; equivalently, the agent could need to make the $t = 1$ investment decision before having experienced its full cost.

This condition implies that whenever the agent (either the H type only, or both) chooses to behave cooperatively, it is *ex-ante efficient* for him to do so: if $\beta\delta v_H r_1 > c_1$, then $\delta v_L r_1 > c_1$. Sometimes, however, he will cave in to temptation and cheat or free-ride, thereby damaging his own long-term interests (e.g., poorly raised child, broken marriage, criminal prosecution or other forms of social retaliation).

Given a self-view $\hat{v}$, the agent invests when $c_1 \leq \beta\delta\hat{v}r_1$, defining a threshold cost level that decreases with $\hat{v}$. *Thus, a stronger moral identity generates valuable self-restraint.* This is also reflected in the continuation value

$$V(v, \hat{v}, A_1) \equiv \delta v A_1 + \int_0^{\beta\delta\hat{v}r_1} (\delta v r_1 - c_1)\, dF(c_1), \tag{7}$$

which increases in $\hat{v}$, since $(v - \beta\hat{v})\,\delta r_1 \geq (v_L - \beta v_H)\,\delta r_1 > 0$. The other conditions in Assumption 2 are satisfied as well.

With regard to welfare analysis, it is no longer appropriate to just add up $-c_0 a_0$ and $E[V]$, since 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. Recalling that V is also an ex-ante value function, our welfare criterion will be:

$$W = E[-\beta a_0 c_0 + V]. \tag{8}$$

The two benchmark cases (AU/SE and SC) presented above can also be combined, so as to determine when anticipatory emotions alleviate or worsen the self-discipline problem. This "mixed" case, more relevant for dimensions of identity other than the moral one, will be examined in Section V, together with other extensions of the basic framework.

B Interpreting the model

Before proceeding to solve the model, we point out three important ways in which it is more broadly applicable than a literal reading might suggest. Readers wishing to skip this discussion can proceed directly to the analysis in the next section.

• *Identity as multidimensional.* We focus the exposition on a single dimension of identity (stock A and associated value v), using moral self-image as a running example. The model can, however, equally 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. The analysis is identical, with everything now interpreted in a "differential" sense, in terms of A relative to B (see Appendix B for details). A second type of identity conflict, arising from rivalry in consumption rather than investment, is analyzed in Section V.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 V.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.[17]

• *Self-knowledge and affirmation of values.* The assumption that people have imperfect insights into their own values and motives admits several formally equivalent interpretations:

(i) 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]).

(ii) 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".

(iii) *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 $1-\lambda$) from the example that their parents set, or from what they force them to do (a_0). 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.

III Equilibrium and Welfare

A Behavior

At date 0, each type chooses his action optimally, taking into account the impact that may result for his self-concept at date 1 and the affective and/or functional payoffs that flow from it. Thus an agent with type $k = H,\ L$, solves

$$\max_{a_0 \in \{0,1\}} \left\{ -c_0^k\ a_0 + \lambda V\left(v_k, v, A_0 + a_0 r_0\right) + (1-\lambda) V\left(v_k, \hat{v}\left(a_0\right), A_0 + a_0 r_0\right) \right\}, \tag{9}$$

where the posterior beliefs $\hat{v}\left(a_0\right)$ in case of self-inference are derived from Bayes' rule.[18] Denoting by x_H and x_L the respective probabilities that types H and L behave prosocially at $t = 0$, this

[17] Different social aspects of identity are explored by Fryer and Jackson [2003], who show optimal categorization can lead to ethnic stereotypes, and by Fang and Loury [2005], who model group identity as a shared convention (akin to a language) for the transmission of information.

[18] By modeling agents as Bayesian, and thus aware that they sometimes make decisions seeking to maintain or enhance a valued identity, we are treating them as fairly sophisticated. Relaxing this "metacognition" assumption (e.g., Bénabou and Tirole [2002]) would make the model's positive results only stronger, but lead in certain cases to different welfare implications (see footnote 27). Note also that while our model has beliefs entering agents' utility functions, as in "psychological games" (Geanakoplos et al. [1989]), these beliefs are about *types,* not actions –whether by others or oneself. As a result, standard equilibrium concepts and refinements for games of imperfect information remain directly applicable.

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

$$\hat{\rho}(1) = \frac{\rho x_H}{\rho x_H + (1-\rho) x_L} \quad \text{and} \quad \hat{\rho}(0) = \frac{\rho(1-x_H)}{\rho(1-x_H) + (1-\rho)(1-x_L)} \tag{10}$$

for all (x_H, x_L) not equal to $(0,0)$ and $(1,1)$ respectively. To lighten the notation, let us define the expected value function

$$\mathbf{V}(v, \hat{v}, A_1) \equiv \lambda V(v, v, A_1) + (1-\lambda) V(v, \hat{v}, A_1), \tag{11}$$

which brings together the *demand* (preferences) and *supply* (cognition) sides of the model, inheriting from V all the properties in Assumption 3. Investing at $t = 0$ is thus an optimal strategy for type $k = H, L$ if

$$\mathbf{V}(v_k, \hat{v}(1), A_0 + r_0) - \mathbf{V}(v_k, \hat{v}(0), A_0) - c_0^k \geq 0. \tag{12}$$

• *The sorting condition.* There are three reasons why this net return to "good behavior" is greater for the H type than the L one, implying that $\hat{v}(1) \geq \hat{v}(0)$ on the equilibrium path. First, the H type has a lower effective cost, $c_0^H \leq c_0^L$. Second, when $V_{13} > 0$, he attaches greater value to any increment to the capital stock. Finally, if $V_{12} > 0$ he also cares more about having a "strong" identity at date 1, which investing helps achieve if $\hat{v}(1) > \hat{v}(0)$.

From now on, we shall restrict attention to *monotonic* Perfect Bayesian equilibria, defined as those in which: (a) the high-value type always invests more: $x_H \geq x_L$, which given (12) again means that $x_H = 1$ whenever $x_L > 0$; (b) a (stronger) form of monotonicity is also imposed on off-the-equilibrium-path beliefs: if $x_H = x_L = 0$, then $\hat{\rho}(1) \equiv 1$; symmetrically, if $x_H = x_L = 1$, then $\hat{\rho}(0) \equiv 0$. This refinement is intuitive and does not affect any qualitative results.[19]

Finally, over a certain range of parameters there may be multiple (three) monotonic equilibria, among which one is Pareto-dominant and will be selected.[20]

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 $[\bar{\rho}, 1]$.

The equilibrium is illustrated in Figure II, for $0 < \bar{\rho} < 1$ and for decreasing values of c_0^L, keeping c_0^H fixed.

[19] It is implied for instance by the Never a Weak Best Response (NWBR) criterion if $V_{12} = 0$ (as is the case for the *SE/AU* specification of the model).

[20] An equilibrium Pareto dominates another one if it yields a weakly higher payoff to both types and a strictly higher payoff to at least one of them.

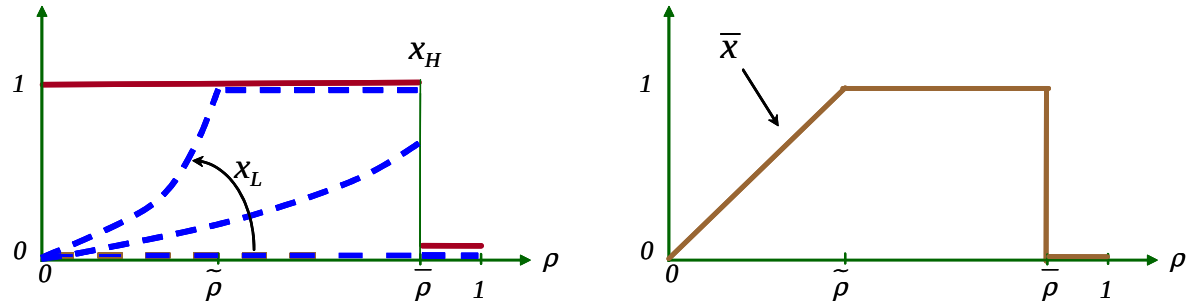

Figure II: Equilibrium as a function of ρ. Left panel: solid line $= x_H(\rho)$, dashed line $= x_L(\rho)$, for decreasing values of c_0^L. Right panel: average investment $\overline{x}(\rho)$.

(i) No investment: when ρ is high enough ($\rho > \bar{\rho}$), 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.[21]

When initial self-image is below the threshold $\bar{\rho}$, on the other hand, the H type needs to invest in order to "stand for his principles" and separate from the less moral L type. Turning now to the latter's behavior, one of three cases arises.

(ii) 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.

(iii) Randomization by v_L : for lower values of c_0^L, it becomes desirable for the L type to imitate the H type, but his ability to do so profitably is limited by the prior ($0 < x_L < 1, \tilde{\rho} = \bar{\rho}$). The lower is ρ, the more truthful (low x_L) his strategy must be in order for investment to signal a high value with sufficient credibility; see (10).

(iv) Universal investment: for c_0^L still lower, even a small gain in self-image is worth pursuing, so the low-valuation type pools completely with the other one ($x_L = 1$), provided ρ is above the threshold $\tilde{\rho}$ (which increases with c_0^L).

Having fully characterized equilibrium behavior, we now derive comparative-statics predictions and relate them to experimental evidence. We shall say that an individual invests more in identity –in our example, behaves more prosocially– when both x_H and x_L (weakly) increase. The total probability of investment, $\bar{x} \equiv \rho x_H + (1-\rho)\, x_L$, also rises as a consequence.[22]

Proposition 2 *(1) An individual invests more in identity:*
(i) the more malleable his beliefs (the lower λ);

[21] The case $\bar{\rho} < 1$ arises only when investment is not so intrinsically desirable for the high-value agent (H type) to engage in for its own sake, without any signaling motive.

[22] Given Proposition 1, the fact that (for all ρ) x_H increases also means that $\bar{\rho}$ increases, and the fact that (for all ρ) x_L increases also means that either $x_L(\bar{\rho})$ increases or $x_L(\bar{\rho}) = 1$ and $\tilde{\rho}$ decreases.

(ii) the lower the investment cost (the lower c_0^L or c_0^H);
(iii) the more salient the identity in the SE/AU case (higher s);
(iv) the higher the capital stock A_0 in the AU case.
(2) Initial beliefs have a nonmonotonic, hill-shaped, effect on overall investment: $\bar{x}$ increases linearly in ρ on $[0, \tilde{\rho})$, equals 1 on $[\tilde{\rho}, \bar{\rho})$, then falls to 0 beyond.

B Implications and evidence on moral identity and behavior

These results can help understand a broad range of empirical phenomena. While some of those admit alternative explanations (learning by doing, habit formation or unstable preferences), a different story would have to be invoked in each case. We aim instead to provide a unified account, which also extends to other evidence considered later on in the paper.

1) Malleability of beliefs. An increase in the ex-ante probability λ that the individual will remain aware, or be reminded of, his true preferences and motives, reduces investment. Identity-management is thus most likely to occur in domains where verifiable, hard information about deep values is scarce (e.g., morality, love, religion). A second, more operationalizable source of variation in λ is the extent to which actions are informative about one's underlying "character" or could instead be attributed to mistakes, rationalized by situational factors, etc.[23]

Dana et al. [2007] document the importance of such inferential *"wriggle room"* for altruistic self-image. When subjects in a dictator-like game did not know whether their payoff and that of the recipient were positively or negatively related, but could find out at no cost, over half of them chose not to know and proceeded to make the self-serving choice; when faced with an explicit tradeoff, by contrast, two-thirds chose a "fair" allocation. Mazar et al. [2008] document a similar effect of attributional ambiguity on self-imposed honesty: when subjects whose payment was based on their self-reported, unverifiable performance on a task earned their compensation in the form of tokens that would later on be exchanged for money (at a known rate), the overinflating of claims (assessed relative to a verifiable-performance benchmark) was 50% higher than when they had to lie for cash directly.

Delegation is another "veil" that commonly allows people to make selfish decisions while protecting their moral self-image. Hamman et al. [2010] show that recipients in dictator games receive much less when principals have the option to delegate the sharing decision to a third party. These agents compete to be "hired" by developing a reputation for favoring principals, and principals systematically seek those known to be the least generous with recipients.

In all these experiments, the fact that such a thin veil allows drastic increases in actual selfishness is also a clear indication of the presence and power of *self-deception.*

2) Salience of identity. In Mazar et al. [2008], making the issue of personal honesty more

[23] See footnote 10 for a formal correspondence between potential excuses and decrease in $1 - \lambda$.

salient (increasing s) by having subjects read the Ten Commandments or a university's honor code before performing tasks in which they could cheat on their claimed performance without risk of detection led to significant decreases in claims inflation.

In the marketplace, an important instance of the same mechanism is consumers' fast-growing expenditure on "symbolic" goods such as carbon offsets, green products and the like, 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.

3) Uncertain values. The overall (ex-ante) probability of investment $\bar{x}$ is hill-shaped with respect to ρ : intuitively, investing in self-reputation has a low return when the prior is low, and is not needed when it is already high (provided $\bar{\rho} < 1$).[24] This means, first, that identity-affirming behaviors are characteristic of people with unsettled preferences and values; hence the moral zeal of the new convert (religious or political), or the exacerbated nationalism of the recent immigrant. Second, 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.

(i) Threats to a strongly held identity (e.g., being a decent, moral person) trigger large opposing responses aimed at restoring the damaged self-image –as occurs in the model when ρ is caused to fall below $\bar{\rho}$. A good example is the *"transgression-compliance"* effect (e.g., Carlsmith and Gross [1969]): 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, even though the requester is not their "victim" and does not even know about their "misdeed". Religions understand well, and make frequent use of, this demand for atonement (e.g., Kuran [1996], Cassone and Marchese [1999]). Symmetrically, subjects with freshly acquired "moral credentials" as non-prejudiced persons show a greater willingness to subsequently express politically incorrect opinions and make employment recommendations that conform to ethnic or racial stereotypes (Monin and Miller [2001]).[25]

(ii) Manipulating weaker aspects of identity (e.g., being helpful, kindhearted), on the other hand, tend to induce confirmatory rather than fighting responses –as occurs in the model when ρ

[24] This non-monotonicity is the general and robust insight from Proposition 2, rather than the specific piecewise-linear response illustrated in the right panel of Figure II. Thus, if small amounts of individual heterogeneity are introduced in the parameters that affect $\tilde{\rho}$ and $\bar{\rho}$, aggregation will result in a "smoothed" version of $\bar{x}(\rho)$ that first increases, then decreases.

[25] To rule out a social-signaling explanation, the two rounds of choices were also administered as ostensibly different experiments, inducing subjects to believe that the second experimenter would not know of their previously established "credentials", or lack thereof. The results were statistically unchanged.

changes marginally, starting from below $\tilde{\rho}$. Such is the case with the *"foot in the door"* effect (e.g., DeJong [1979]), in which freely accepting an initial request for a small favor raises the probability of accepting a more costly one in the future. Conversely, an initial costly request, which most people turn down, decreases the probability of accepting a smaller one later on.[26]Identity and welfare: treadmill effect or empowerment?

While the model's equilibrium-behavior and comparative-statics results are very general, relying only on Assumptions 1 to 3, the implications of belief management for an individual's welfare depend on whether it reflects a demand for "consumable thoughts" or instrumental concerns.

1) Self-esteem / anticipatory utility and the treadmill effect.

Equations (4)-(5) lead to

$$W = \rho x_H \left[(s+\delta)\, v_H r_0 - c_0^H \right] + (1-\rho)\, x_L \left[(s+\delta)\, v_L r_0 - c_0^L \right] + (s+\delta)\, \overline{v} A_0. \tag{13}$$

The last term is constant: although agents actively manage their self-views, this is a zero-sum game across types, by the law of iterated expectations.[27] As to the first two terms, they always (weakly) decrease as identity investments rise in response to a greater malleability of beliefs, $1-\lambda$. This is immediate to see when self-regard is the sole motive underlying moral behavior, and more generally when the identity-related asset is fixed: with $r_0 = 0$, there remains only a net loss of $-\rho x_H c_0^H - (1-\rho) x_L c_0^L$.[28] The result –a form of wasteful signaling induced by imperfect self knowledge– applies equally when identity capital can be accumulated.

More strikingly, an increase in his capital stock can also make the individual worse off. Indeed, the condition for a no-investment equilibrium ($x_H = x_L = 0$),

$$\mathbf{V}\left(v_H, v_H, A_0 + r_0\right) - \mathbf{V}\left(v_H, \overline{v}, A_0\right) = (s+\delta) v_H r_0 + (1-\lambda)\, s\, (v_H - \bar{v})\, A_0 \leq c_0^H, \tag{14}$$

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.[29]

The model thus yields a type of *treadmill effect:* higher asset levels do no generate much of

[26] In neither case are the results due to self-selection, since the probabilities being compared are the average compliance rates between the members of an experimental group (who get two requests) and those of a control group (who get only the second request).

[27] For welfare gains to arise, it must thus be that either: (a) agents' updating is at least partially *naïve*: when $a_0 = 1$, they do not properly correct for pooling by the L type, resulting in a departure from the martingale property of Bayesian beliefs. This additional form of malleability could easily be incorporated into the model (e.g., Bénabou and Tirole [2002]); or (b) the consumption value of beliefs is *nonlinear* (and thus not purely anticipatory in the standard sense), as in Rabin [1995], Caplin and Leahy [2001] and Köszegi [2009].

[28] Each product $x_k c_0^k$, $k = H, L$, is (weakly) decreasing in λ if $c_0^k > 0$, or constant if $c_0^k \leq 0$, since (12) then implies $x_k = 1$.

[29] Equation (14) leads to a loss when $c_0^H > (s+\delta) v_H r_0$. Otherwise, $x_H \equiv 1$ and a loss arises from the counterpart of (14) that focuses on how x_L rises with A_0.

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.[30] In the moral realm, one can point to religious and political zealotry (all the way to self-mortification), or the compulsive internalization of honor and shame. The most economically relevant applications of the result, however, concern assets such as wealth or prestige. Studies of "hedonic forecasting" thus suggest that people tend to overestimate the contribution of material or status goods to their long-term life-satisfaction, relative to time spent in personal relationships or doing good, such as volunteering (see, e.g., Stutzer and Frey [2007] for a survey).

The model also sheds some light on these differences: a treadmill effect is more likely in activities that are subject to decreasing returns, which cause the material return $r_0 v$ to fall relative to the savoring motive, $sA_0(\hat{v} - \bar{v})$.[31] 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. Consequently, a moral treadmill is much less likely than a material one.

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.*

An important caveat is that the welfare analysis is conducted here from the perspective of one agent, and thus abstracts from the external costs and benefits that his behavior generates for others. Even when considering social welfare, however, the point remains that 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".[32]

2) Willpower and the commitment value of identity

[30] Ours is thus a different mechanism for treadmill effects from the traditional one, which is based on preferences or "aspirations" adapting to changes in consumption levels.

[31] Let utility from a long-run stock A_2 be $v\phi(A_2)$ instead of vA_2, where ϕ is concave. The total return to investing at $t = 0$ is then $(s\hat{v}(1) + \delta)\phi(A_0 + r_0) - [s\hat{v}(0) + \delta\phi(A_0)] \approx s[\hat{v}(1) - \hat{v}(0)]\phi(A_0) + r_0\phi'(A_0)[s\hat{v}(1) + \delta)]$, provided r_0/A_0 is not too large. Decreasing marginal utility from A, leading to a low $\phi'(A_0)$, is thus equivalent to a low r_0 in the linear specification (or, also, to a technology where r_0 falls with A_0).

[32] Sen [1985] discusses identity as personal "commitments" –distinct from any kind of altruistic utility function or "goal"– which individuals feel bound to respect even though this may lower their own welfare, while benefiting others. Our model shows how such a notion can be formalized, in a way consistent with consequentialist rationality. Thus, a rule to "always" cooperate ($a_0 = 1$) because "that is what a good person does, and this is who I am", can: (i) be self-enforcing although it lowers U_0^i when taking others' actions as fixed; (ii) in general equilibrium, yield a Pareto improvement. In the self-control version of the model, such cognitive commitments can also serve the individual's own long-term "goals", while running against (and constraining) his short-term preferences.

In the basic self-control version of the model, A_0 has no behavioral impact, as seen from (7).[33] The malleability of beliefs, on the other hand, now affects behavior both at $t = 0$ and at $t = 1$. Suppose for instance that: (a) for $\lambda = 1$, neither type behaves prosocially at $t = 0 : c_0^H > \delta v_H r_0$, so $x_H = x_L = 0$; (b) for some $\lambda < 1$, on the contrary, the equilibrium involves mixing: the more altruistic type always cooperates ($x_H = 1$), while the more selfish one randomizes ($0 < x_L < 1$).[34] The difference in intertemporal welfare $W = E\left[-\beta a_0 c_0 + V\right]$ between these two cases is then

$$\Delta W = (1-\rho)x_L\left(\delta v_L r_0 - \beta c_0^L\right) + \rho\left(\delta v_H r_0 - \beta c_0^H\right) + (1-\lambda)\,E\left[\Delta V\right], \tag{15}$$

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

$$E\left[\Delta V\right] = (1-\rho)x_L \int_{\beta\delta v_L r_1}^{\beta\delta\hat{v}(1)r_1} \left(\delta v_L r_1 - c_1\right)\; dF\left(c_1\right) - \rho \int_{\beta\delta\hat{v}(1)r_1}^{\beta\delta v_H r_1} \left(\delta v_L r_1 - c_1\right)\; dF\left(c_1\right). \tag{16}$$

The first term in (16) shows how, when the L type invests at $t = 0$, this *strengthens* his moral self-regard and thereby raises his subsequent propensity to behave well. At the same time, such pooling at $t = 0$ *dilutes* the identity of the H type, and this 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 (6)), the first effect leads to a welfare gain, the second to a loss.[35]

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 (15) are positive, ex-ante efficient investments fail to occur in period 0 if $\lambda = 1$: from the very start, the agent behaves too opportunistically for his own good. The ability to affect his self-image ($\lambda < 1$) provides additional motivation for acting prosocially at $t = 0$, which then directly raises ΔW. When the first two terms in (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 $E[\Delta V]$ sufficiently positive.

Proposition 4 *In the self control case, more malleable beliefs (a lower λ) can raise welfare, by improving choices at $t = 1$ (when $E\left[\Delta V\right] > 0$) and/or at $t = 0$ (when $\Delta W > (1-\lambda)\,E[\Delta V]$).*

[33] Unless A_0 affects the return r_0 or cost c_0, which would be easy to incorporate. There could thus be decreasing returns to social capital; or, on the contrary, a transgression ($a_0 = 0$) could destroy not just a fixed number r_0 of long-term relationships but a fixed fraction of the existing stock A_0.

[34] This is without loss of generality: a similar reasoning applies for complete pooling (whether on 0 or on 1), with $\hat{v}(1)$ simply replaced by $\bar{v}$. Of course, the nature of the equilibrium, including the value of $\hat{v}(1)$, is endogenous and depends on the distribution $F(c_1)$. The proof of Proposition 4 takes this fixed-point aspect into account.

[35] More specifically, when $F(\cdot)$ is such that the support of $c_1/\beta\delta r_1$ is mostly concentrated in the interval $[v_L, \hat{v}(1)]$, meaning that the (opportunity) cost of good behavior and the magnitude of the self-control problem are relatively moderate, there is a net gain from malleability. When they are more severe, so that the support is mostly concentrated in $[\hat{v}(1), v_H]$, there is a net loss.

IV Taboos and Transgressions

Taboos and sacred values are closely related to identity, in the sense of *protecting certain beliefs (or illusions)*, deemed vital for the individual or for society, concerning things one "would never do" and the "incommensurable" value of certain goods. We distinguish two complementary ways in which they operate –ex ante and ex post. The first, *internally* enforced, aims to *avoid* dangerous (self-) knowledge that might surface from "cold" analytical contemplation of what short-run tradeoffs might be available or expedient. The second one, *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.

A Sacred values, taboo tradeoffs and markets: information avoidance

"To compare is to destroy" (Fiske and Tetlock [1997])

Let $v \in \{v_H, v_L\}$ continue to denote the long-run value of some important asset, with associated capital A_t. In the social-moral realm this again could be family, friends, clan, country, or religion. In the personal one it could be health, bodily integrity, or personal freedom. We saw how, for motives of either *anticipatory utility* (possibly extending to an afterlife) or *self control* (temptations to erode A_t for short-term gains), people will often want to be optimistic about v, resulting in a value function $V(v, \hat{v}, A_1)$ satisfying Assumption 3.

Suppose now that, at $t = 0$, an agent can find out the "sellout" price p at which he could exchange one unit of A_0 against money or other goods of known consumption value. A priori, the price could be high or low,

$$p = \begin{cases} p_H & \text{with probability} \quad z \\ p_L & \text{with probability} \quad 1-z \end{cases} . \tag{17}$$

Depending on the context, the actual value may be learned by checking what is being offered on a formal or informal market (for loyalties, votes, crime, organs, sex, children, etc.) or by simply engaging in deliberate, "coldhearted" calculations about the personal costs and benefits of different courses of action.

To simplify the problem, let p_H be high enough and p_L low enough that, if an agent does ascertain the price ($a_0 = 0$), he will always transact when $p = p_H$, reducing A_0 by one unit, and not transact when $p = p_L$.[36] Even in the latter case, he will later recall that he *contemplated* the possibility of a transaction and *evaluated* whether maintaining his identity or dignity was "worth

[36] Formally, this is a dominant strategy for both types $k = H, L$, provided that $p_H > \mathbf{V}(v_H, v_H, A_0) - \mathbf{V}(v_H, v_L, A_0 - 1)$ and $p_L < \mathbf{V}(v_L, v_H, A_0) - \mathbf{V}(v_L, v_L, A_0 - 1)$. In the absence of such conditions, or with a more general price distribution, there may be two signals of an agent's type: whether he looked into the price and, if so, whether he transacted or not, given the price. We isolate here the first effect, which is the relevant one for the idea that certain things should remain "priceless" and the presence of a "mere contemplation effect".

it" or not. From this fact he will then have to draw (with probability $1-\lambda$) the appropriate inference about where his true values lie.

Investing in moral identity ($a_0 = 1$) thus consists here in upholding a rule *never to place a price* on certain goods –staying away from markets where such transactions occur, not entertaining for a second any "indecent proposal" one may receive, and *forbidding oneself even mere thoughts of commensurability.* The cost of doing so is the option value of the potential transactions foregone, so an individual with value $v = v_H, v_L$ will uphold the taboo if

$$\mathbf{V}(v,\hat{v}(1), A_0) - \mathbf{V}(v,\hat{v}(0), A_0 - z) \geq z p_H, \tag{18}$$

with the same notation as usual.[37]

This is clearly a special case of our model, with $r_0 = z$, $c_0 = z p_H$ and initial stock $A_0' \equiv A_0 - z$. Therefore, all previous results apply directly:

(1) On the positive side, Propositions 1 and 2 show *how taboos arise and are sustained,* either universally (full-investment equilibrium) or predominantly by the more committed (mixing or separating equilibrium); how this depends on the initial strength of beliefs, ρ; and how challenges to taboos or transgressions by others can lead to reaffirmation or collapse, according to which side of the "hill" (Figure II, right panel) the induced erosion of ρ occurs on.[38]

(2) On the normative side, Propositions 3 and 4 show how the *welfare effect of taboos* depends importantly on whether they reflect anticipatory or self-control motives. In the first case, upholding taboos generally lowers an individual's ex-ante welfare.[39] In the latter it can be beneficial, but only under specific conditions involving the severity of the self-control problem.[40]

B Dealing with sinners and saints: information destruction

"Anyone who has violated a taboo becomes taboo himself, because he possesses the dangerous quality

[37] We assume that transacting without first finding out the price is either infeasible, or else unprofitable (due to the average "auction" price $z p_H + (1-z) p_L$ being too low). In writing the second term in (18) we took advantage of the linearity of $\mathbf{V}$ in A_1 under both the AU and the SC models (and their combination in Example 3). More generally, it would be $z\mathbf{V}(v,\hat{v}(0), A_0 - 1) + (1-z)\mathbf{V}(v,\hat{v}(0), A_0)$, which leaves all the results unchanged.

[38] Because they involve the avoidance of normally valuable information, taboos are related to the strategic ignorance in Carrillo and Mariotti [2000] and Carrillo [2005], and especially to the rule-based behavior in Bénabou and Tirole [2004]. There are, however, two important differences. On the demand side, imperfect willpower is here only one of several potential sources of motivated beliefs. On the supply side, it is the mere act of exploring the price to be gained from certain transactions, rather than the price thus revealed or whether the transaction is actually "consumed", that destroys the valued belief. In Fershtman et al., [2008], the benefit of taboo contemplation is also the option value of finding out that one might benefit (on net) from engaging in a socially reprehensible action. The cost deterring from taboo contemplation, on the other hand, is an exogenously assumed function of how many other people are believed to be refraining from it.

[39] Unless agents are sufficiently non-Bayesian, or the consumption value of beliefs appropriately nonlinear.

[40] The impact of taboos on individual and on social welfare will of course differ when agents' actions have direct externalities on others (in addition to the informational spillovers on which we focus here). Dessi [2008] analyzes the role of indoctrination by a benevolent principal in such public-goods contexts, and Bénabou [2008] the contagiousness of beliefs, as well as responses to dissent.

of tempting others to follow his example: why should he be allowed to do what is forbidden to others? Thus he is truly contagious in that every example encourages imitation, and for that reason he himself must be shunned". (S. Freud, Totem and Taboo, p.86).

Consider now a situation where someone has behaved "immorally" –exhibiting selfishness or, through his words or actions, breaking a taboo. How will others respond, and how will the violator himself react to such lapses? We seek to understand in particular the *coexistence of social and antisocial punishments:* in some cases people ostracize (or even incur personal costs to hurt) the less virtuous, and in others the more virtuous members of their group (Monin [2001], Jordan and Monin [2008], Monin et al. [2008], Herrmann et al. [2008]).

Our analysis builds on a simple *benchmarking* idea: in assessing what kind of a person they are, people compare themselves to others whom they feel are akin to them or face a similar environment. "Deviant" behavior by peers ($a_0 = 0$) 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). Thus, members of a religious, ethnic or national community who are not fully supportive of its positions or mingle with outsiders undermine others' sense of commitment to (belief in) the common value. If the lapsed individual is oneself, on the other hand, it is *good* behavior by peers that is now 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.*[41] Consider the two-agent generalization of the basic model that is illustrated on Figure III. The first new element is that, with ex-ante probability θ, there is a valid "excuse" for not behaving in an identity-congruent manner. To keep with our main example, this corresponds to a situation where choosing $a_0 = 1$ is useless –perhaps even harmful– to the rest of society, or where the private cost is so high that even the most moral types (H) would choose $a_0 = 0$. With probability $1 - \theta$, on the other hand, the action $a_0 = 1$ is socially beneficial, so performing it can be a sign of valuing others. Formally, for an individual with altruism type $k = H, L$, the return in terms of relational capital is now $r_0^k = \xi v_k$, where $\xi = 1$ when the action is useful to others and $\xi \leq 0$ when it is not. In the former case the net costs involved are still $c_0^H < c_0^L$; in the latter, they are $\tilde{c}_0^H \geq \tilde{c}_0^L$, reflecting the fact that a more prosocial agent is less inclined to engage in a socially harmful action.

The second new element is that the 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 either leaves, both lose the benefit b of future interactions, which we take to be symmetric and type-independent for simplicity. To allow for ex-post rationalizations as to why the separation occurred, we assume

[41] We borrow here from the title of Ross and Nisbett's [1991] classical book in social psychology.

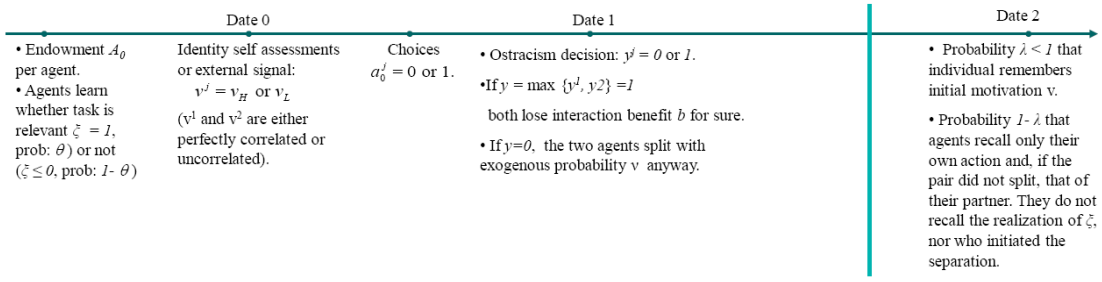

Figure III: The ostracism game

that matches also dissolve for independent reasons, with exogenous probability ν. Agent i's utility function is thus:

$$(v^i\xi - c_0^i)a_0^i + \mathbf{V}(v^i, \widehat{v}^i, A_0 + r_0 a_0^i) + (1-\nu)(1-y)b, \tag{19}$$

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

At date 1, each agent always remains aware of his own behavior a_0^i, but he recalls (or is reminded of) 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. The idea is that, whereas what one did is "hard" data that is relatively easy to remember and verify, the past behavior of someone with whom contact has ceased and the true reasons why the split occurred represents softer, less verifiable information. And it is always more pleasant, ceteris paribus, to "recall" whatever scenario is the most kind to one's self-image. These no-recall assumptions are extreme and meant only to simplify the derivations. All that is needed for the results is that ostracism reduces the probability that people are reminded of the bad news conveyed by their peers' behavior.

We allow for two polar forms of benchmarking:

- *Benchmarking on the person*: in this version the date-0 contribution is always socially useful ($\xi \equiv 1$), as has been assumed so far; equivalently, there is never any excuse for not supplying it. Moreover, the two individuals' values are perfectly correlated: $v^1 = v^2 \in \{v_H, v_L\}$.

- *Benchmarking on the situation*: in this version, the two types are independent, but the social usefulness ($\xi = 1$, with probability θ) or absence thereof ($\xi \leq 0$, 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 $1-\lambda$.

We shall 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).[42]

Proposition 5 *(1) In an equilibrium such that the H type invests when it is socially useful ($\xi = 1$), let $x \in [0,1]$ denote the probability of investment by the L type.*
(i) Ostracism ($y = 1$) occurs only when actions differ, i.e. one agent invests and the other not.
(ii) Ostracism comes from the virtuous agent ($a_0^j = 1$) when benchmarking is on the person and from the unvirtuous one ($a_0^i = 0$) when benchmarking is on the situation.
(2) 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:
(i) When benchmarking is on the person, $x = 1$ is sustained by the ostracism of "sinners" (a prosocial norm), while $x = 0$ involves no ostracism.
(ii) 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.

The first result shows how a value of *social conformity* (strategic complementarity) arises endogenously from individual concerns over *self*-image, as these give each agent an incentive to exclude those who act differently from him. The second one captures the idea that discordant actions are more threatening to a person's self-concept, the more similar the individuals involved are to him, either in their personal characteristics (religion, ethnicity, occupation, etc.) or in the social environments they face (e.g., country and its level of development). Accordingly, the harshest moral condemnations and punishments are reserved for deviant "insiders".[43]

The second result also allows us to understand not only the standard findings that free-riders in public-good games get punished (e.g., Fehr and Gächter [2000]), but also the more puzzling phenomenon of derogation, resentment and even punishment of those who exhibit stronger moral principles or contribute "too much" to public goods (e.g., Monin [2001] and other previously cited references). The mechanism involved, moreover, is very much in line with Monin's interpretation of these behaviors as ego-defense mechanisms in response to threatening social comparisons in the moral domain.

The proposition's third result, finally, sheds light on *cross-society-differences in civic norms* and how they are enforced. Herrmann et al. [2008] find a significant positive (respectively, negative) correlation across countries between the extent of social (respectively, antisocial) punishment by subjects in public-goods experiments, and national survey responses on both: (a) citizens' lack of tolerance for tax evasion, welfare fraud and other cheating behaviors; (b) their trust in other people's adherence to the rule of law and in local institutions' enforcement of it.

[42] These two restrictions on the set of equilibria we consider are respectively ensured by conditions (C.22) and (C.11) in Appendix A.

[43] For instance, the Catholic Church long imposed excommunication on apostates, and tortured and executed heretics. The Sharia still prescribes that apostates should be put to death, lose their children and their property. For an experimental study of the "black sheep effect", see, e.g., Branscombe et al. (1993).

V Further Applications

A Other dimensions of identity

Having so far focused on prosocial identity as the main example, we now relate the model's results to evidence on how other dimensions of people's self-concept affect their behavior. As in Section III.B, the different paragraphs each correspond to a specific result in Proposition 2 for the positive ones, or Propositions 3-4 for the normative ones.

1) Salience of identity. Messages or cues that make specific components of a person's identity more salient elicit investments along the same dimensions. LeBoeuf and Shafir [2010] thus find that even minor manipulations emphasizing alternative aspects of subjects' self-concept, such as scholar versus socialite, or ethnic Chinese versus American citizen, trigger identity-consistent expressions of consumption preferences. In experiments with monetary stakes, Benjamin et al. [2010] similarly find that priming subjects to their ethnic identity causes Asian-Americans to make more patient choices, Whites to makes choices that are both more patient and less risk averse, and African-Americans to make more risk-averse ones.

A direct economic application of salience manipulation 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.

2) Uncertain values and malleability of beliefs. People who, deep down, are insecure about "who they are" (ρ in the middle range) are the most prone to costly identity-affirming behaviors; adolescents are perhaps the prime example. Individuals with stable self-knowledge, by contrast, invest only if r_0 is large enough to justify the cost. In line with this "uncertainty principle," Adams et al. [1996] found that male subjects with strongly declared homophobia actually showed the most arousal in response to male homoerotic videos (with no difference from others subjects for heterosexual or female homoerotic materials).

3) Escalating commitment. The more identity-relevant capital they have, the more identity-affirming investment people will make, thereby raising the stock even further.[44] Intuitively, someone with more A_0 has a greater vested interest in viewing this asset as valuable, and further investment is the way to demonstrate such beliefs –in accordance with the psychology

[44] This result is not due to any increasing returns in investment: in our model, r_0 and c_0 are independent of A_0. Instead, it reflects the fact that people with more at stake have a higher demand for optimistic beliefs (implying $V_{23} > 0$), an idea that has substantial empirical support. Pyszczinsky [1982] found that lottery participants rated the prize as more desirable, the greater their perceived chance of winning it. Kay et al. [2002] found similar outcomes among political partisans for electoral outcomes and among students for changes in tuition. In the health domain, Kunda [1987] found that a purported scientific article about the increased risk of breast cancer from coffee was judged less credible (among female but not male subjects) by coffee-drinkers than by non-coffee drinkers. Best known is the "Stockholm syndrome", in which hostages come to see their captors in a favorable light, most plausibly so as to maintain hope that they will not harm them.

literature on self-justification. A manager will thus keep throwing good money after bad on a doomed project, as in the experiments of Staw [1976]. A farmer faced with adverse market signals may obstinately refuse to quit rather than admit that his efforts and sacrifices (or those of his parents) have been in vain**.** Others will keep accumulating wealth, professional achievements, political or religious activism, not so much for the marginal product of the later investments but to preserve the perceived value of earlier ones –that is, to safeguard the belief, true or false, that these assets will bring happiness (or forestall misery) over the course of their lifetime, or a favorable fate in some hereafter.

4) Responses to identity threats and boosts. Because equilibrium investment is (qualitatively) hill-shaped in the strength of identity ρ (see right panel of Figure II), manipulations of strong and weak self-images in the same direction tend to have opposite effects.

(a) Threatening a strong identity triggers a fighting response. In Maas et al. [2003], male subjects who were told by the experimenters that their score on a personality test was so atypical as to place them squarely in the female part of the distribution were subsequently much more likely than the control group to harass a female (but not a male) chat-line user, by sending her pornographic images. This effect was further accentuated when she (a confederate) had previously described herself as a professionally ambitious feminist rather than a meek, family oriented traditionalist. It was also more pronounced, the more the subjects had initially self-rated themselves as masculine.

(b) Questioning a weak identity tends to induce confirmatory rather than opposing responses. Consider, for instance, the debilitating impact of *"stereotype threat"* on test performance (Steele and Aronson [1995]). A stereotype of female or African-American students as having a lower distribution of comparative mathematical abilities than their male or White and Asian counterparts means precisely that society places a lower probability on their being a high type (with v_k now representing ability rather than taste, or a combination of both). Making gender or race subtly more salient before a test reminds these subjects of this widespread statistical perception and thus (consciously or unconsciously) lowers their self-confidence. The equilibrium response to this decrease in ρ is (on average) to discourage academic-identity investment –in this case, effort and motivation to perform on the test.

B Extensions of the model

• *Wishful thinking and procrastination: entrepreneurial versus precautionary behaviors.* When does the desire to indulge in hopeful thoughts and avoid frightening ones aggravate the self-control problem, and when does it alleviate it? To answer this question, we combine the AU and SC specifications and allow for type-dependent returns $r_1(v)$ in investment. For an agent

with self-view $\hat{\rho} \in [0,1]$, the marginal expected utility of investment at $t=1$ is then

$$\hat{\rho} v_H r_1(v_H) + (1-\hat{\rho}) v_L r_1(v_L). \tag{20}$$

More optimistic beliefs enhance savoring of the existing stock (raising $\hat{v}A_1$), but whether they induce higher investment or "coasting" hinges on whether $z_1(v) \equiv v r_1(v)$ rises or falls with v, bringing to light an important dichotomy between situations in which identity and effort are *complements or substitutes.*[45]

(a) Wealth accumulation, status-seeking, and other entrepreneurial behaviors (complementarity). When $z_1(v)$ is increasing, wishful thinking (raising $\hat{\rho}$) alleviates the motivation problem, if there is one; otherwise, it only results in excessive activism. This case occurs for instance if r_1 is type-independent (financial assets), or if v corresponds to some ability that raises both the probability of winning in a competitive situation and the expected value of the prize. Unrealistic dreams of riches and glory –and of how enjoyable those will be– thus propel entrepreneurs, explorers and athletes to sacrifices and persistence in the pursuit of long-term endeavors.

(b) Health investments, safe driving and other risk-prevention behaviors (substitutability). In those cases $z_1(v)$ is decreasing. A favorable genetic endowment thus protects from disease and makes taking care of one's health less of a necessity; good driving skills and reflexes permit driving at faster speeds. Wishful thinking –understating the likelihood of illness, accident or death– then makes the present more enjoyable but further encourages negligent and risky behaviors that are precisely those to which weakness of will already makes one too tempted to succumb.

• *Disappointment aversion.* Beyond self-esteem and anticipatory utility, many other forms of "mental consumptions" (Schelling [1985]) can be incorporated into the model. For instance, while anticipatory utility creates a demand for optimistic beliefs, the fear of being disappointed when final outcomes are realized generates an opposing incentive to maintain low expectations ("defensive pessimism"). This corresponds to a period-2 payoff of the form $D((v-\hat{v})A_1)$, where D is increasing, concave and satisfies auxiliary assumptions listed in Appendix B.

• *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 B).[46]

[45] See Appendix B for details, including the value function corresponding to (20).

[46] On social signaling see, e.g. Bernheim [1994] and Bénabou and Tirole [2006b].

C Competing identities and dysfunctional behavior

We saw in Section II.B how the single-asset model can be interpreted as representing a tradeoff between two identity dimensions, A and B, whose relative value is uncertain and which are subject to resource or time rivalry at the *investment* stage. We analyze here a different kind of identity conflict, *consumption* rivalry, and show it can lead to highly dysfunctional behaviors.

When time, geographical, legal or other exclusivity constraints (such as national or religious affiliation) create a potential tradeoff between reaping 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. If he has substantial capital vested in A but the ultimate value of this identity is less "secure" than that of B, he may then refrain from even highly desirable investments in B and end up worse off as a result. We demonstrate this mechanism using anticipatory utility or self-image, then discuss the more general case. We also make simplifying assumptions under which A can be interpreted as the "traditional identity" and B as the "modern" one –for instance, in the context of farmers and workers faced with globalization and technical change, or that of immigrants confronting the issue of integration.

(a) Modern identity. At $t = 0$, the agent decides whether to invest in B ($b_0 = 1$), at a cost c_B, type-independent for simplicity: acquiring new skills, mastering a new language and culture, socializing with an unfamiliar group, etc. The investment succeeds with probability $\pi \in (0, 1)$, in which case B_0 rises to $B_1 = B_0 + b_0 r_B$; it fails with probability $1 - \pi$ ($B_1 = B_0$), for instance because this is a new activity to which the agent may not be suited. The (per unit) value of B capital, on the other hand, is a known v_B. For instance, the monetary benefits of successfully integrating into the formal, majority-dominated labor market, of acquiring a degree or working in the more dynamic sectors of the economy, are relatively easy to assess.

(b) Traditional identity. There is no possibility of investment in A at $t = 0$. Thus A_0 corresponds either to a fixed trait (e.g., ethnicity) or to an asset that was accumulated in the past but can no longer be significantly augmented: long-held skills, connections to "the old country", etc. Furthermore, the hedonic value of this stock is uncertain, since its benefits are of a more subjective and less quantifiable nature than, say, those of a wage premium: strength of personal values and commitments, long-run utility from family, morals, culture, religion, etc. Thus v_A equals v_H or v_L, with probabilities ρ and $1 - \rho$.

The timing is the same as before. At date 0, the agent receives the signal v_A, then chooses $b_0 \in \{0, 1\}$. At date 1, he recalls v_A with probability λ ($\hat{v}_A = v_A$), and otherwise looks to his past actions to form his sense of identity ($\hat{v}_A = \hat{v}(b_0)$). At date 2, he is aware of v_A (one could allow for uncertainty here as well) and, assuming full rivalry, chooses optimally between consuming either A or B, thus achieving $\max\{v_A A_2, v_B B_2\}$. To focus on the interesting case, suppose that:

(a) *ex post,* the agent will consume B only if he had successfully invested in it,

$$v_B B_0 < v_L A_0 < v_H A_0 < v_B (B_0 + r_B), \tag{21}$$

so that A serves as a "fallback" or insurance option;

(b) *ex ante,* the expected return from investing in B is sufficiently high that, when beliefs are not malleable (the "objective" case where $\lambda = 1$), such investment is optimal even for agents who value A the most:

$$\pi (s + \delta) [v_B (B_0 + r_B) - v_H A_0] > c_B. \tag{22}$$

When self-perception concerns are operative, however, *both* types will fail to make this efficient investment, as long as

$$\pi (s + \delta) [v_B (B_0 + r_B) - v_L A_0] - (1 - \pi) s (1 - \lambda) (\bar{v} - v_L) A_0 < c_B. \tag{23}$$

The first term is the standard economic return to investing, for an agent with relatively low valuation for A. The second term represents the *loss of identity* that is incurred (by either type) when doing so: with probability $1 - \lambda$ such "betrayals" will signify to the individual that he does not care that much about A, and therefore has only grim prospects to look forward to in case his investment in B does not work out.

On average, such affect-driven identity management ends up lowering personal welfare, as in the single-identity case. While the value function is now nonlinear, making the analysis more complicated, one can exploit the basic intuition that *not* investing in B is effectively like investing in A, to show that all the preceding results apply here as well.

Proposition 6 *Assume the anticipatory-utility (AU) specification, with (21)-(22).*
1) The individual invests (weakly) less in a known identity (B) when it will compete in the future with another one (A) of uncertain value.
2) This is more likely to happen the higher $A_0, 1 - \lambda$ and s, and it is always welfare reducing.

These results relate to some important social and economic issues.

1) Resistance to structural change. Trade and technical change alter the relative payoffs to working in the modern, international sector and in traditional activities. 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, to which one might need to return one day.

2) Resistance to assimilation. Immigrants and their descendents experience strong tensions between integrating into Western societies and preserving their specific culture. This is particularly acute for the young, who are locally born and have citizenship but often do not feel British, German or French. Yet neither do they feel Pakistani, Turkish or Algerian, having little

knowledge of the "old country" or its language. As seen earlier, it is in situations of uncertainty over one's own values that identity threats and investments become most relevant.[47] Laws and proposals such as the French ban on the veil or the Home Secretary's [2001] urging that newcomers adopt British "norms of acceptability", take an oath of allegiance and embrace "our laws, our values, our institutions" then elicit significant opposition from those whose who feel that complying would represent a betrayal of their own culture or religion.[48] Conversely, native populations feel that the values and traditions they "believe in" (religious, secular, political, etc.) are undermined by visible displays of adherence to other cultures among newcomers, and especially their locally-born descendants.

In a related vein, it has been suggested that low educational achievement among African-Americans students may partly reflect a desire to maintain an oppositional ethnic identity. Austen-Smith and Fryer [2005] offer a model of "acting White" in which some minority students forsake educational investment in order to signal loyalty to their peers. The idea there is a different one, namely that having demonstrably low labor market prospects makes one less likely to leave when called upon to "give back to the community" in the future.[49]

2) 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. This corresponds in the model to the case where $c_B < 0$, so that the costly action is now one that reduces B or prevents it from growing ($b_0 = 0$). In the events that shook the suburbs of French cities in 2005, for instance, the young rioters attacked and destroyed a number of schools, nursery schools and cars in their own communities.

It is also interesting to note *two factors that can "tip" the equilibrium* from one in which people optimally invest in B to one in which they self-defeatingly destroy those assets (i.e., affecting (23) while leaving (21) and (22) unchanged). The first is a lower perceived chance of success in those investments (π) or the associated payoff (r_B). Thus, if minority youth become more pessimistic about their chances of mobility through education, or perceive that even with diplomas the jobs to which they can aspire will be low-paying ones, they will switch to the destructive-identity scenario, even when π and r_B remain high enough that investing in B (education, integration) *would still make them better off* in the long run. A second potentially

[47] The results in Proposition 6 on the effects of A_0 and ρ are also consistent with the findings by Constant et al. [2006] that, among immigrants to Germany, the probability of assimilation decreases with age at arrival and with having had primary or secondary schooling in the country of origin.

[48] See Hoge [2002]. Here again, self-perceived intentions matter: infiltrated members of an extremist organization feel much less conflict in submitting to such requirements, pledges, dress codes, etc., because they know that their doing it really signals commitment to, rather than abandonment of, their chosen values.

[49] In our case, the (stochastic) returns to education are common knowledge and there is no incentive to deceive others. Instead, the individual wants to sincerely believe, and thus tries to convince himself, that his community is very valuable to him –instead of his being valuable to them. Moreover, since this mechanism does not involve any community enforcement of membership "payments" through the expulsion of defectors, the relevant community or identity capital can be far away, uncoordinated, or even virtual (e.g., native country, culture, religious faith).

important factor is the salience s of the "alternative" A identity and the benefits anticipated from it –as with advertising in the single-identity case. This is where ideological or religious indoctrination may come into play, as well as the amplification mechanism of media coverage.

While we have focused here on the anticipatory-utility or self-image case, which is somewhat simpler and seems more appropriate to the applications just discussed, similar insights apply when the demand for identity stems from a commitment problem. If the individual expects sufficient temptation to underinvest in A relative to B at $t = 1$, he will not invest in B at $t = 0$ even if it has a high return, and may even destroy B capital. Such a strategy serves *not* as a physical commitment (investment costs and returns are independent of the stocks) but as a *cognitive* one, aimed at *defining oneself* as an A-person rather than a B-person. From Proposition 4 we know that welfare may go up in this case, but need not.

VI Conclusion

We examined in this paper how moral identity shapes individual and collective behavior. More generally, we developed a simple, flexible framework for analyzing a broad class of economically important beliefs which people value and invest in. The model also offers a unified account of many seemingly disparate or contradictory findings by psychologists and experimental economists. Others, such as endowment effects, could easily be obtained (see Gottlieb [2008]).

Rather than restate here the paper's results, we will single out two interesting avenues for further research which they point to. The first one is that of *sacred values and taboos,* where our framework offers a way of bringing the debate over markets and morals into the realm of formal analysis. The second one concerns the role, in bargaining and other distributive conflicts, of self-serving beliefs linked to pride and dignity concerns. In Bénabou and Tirole [2009], these are shown to reduce the range of sustainable sharing agreements and, beyond a point, inevitably cause a bargaining impasse in spite of fully symmetric information. Many interesting questions remain to explore along this line, such as the optimal design of contracts and organizations or the political economy of reforms when agents have motivated beliefs.

Appendix A: Main Proofs

Proof of Proposition 1. The difference between the two types' incentives to invest in (12) is

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

If $V_{12} = 0$ (as with anticipatory utility) then $\Delta > 0$, so *any* equilibrium must have $x_L\left(1 - x_H\right) = 0$. 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:

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

$$\mathbf{V}\left(v_H, \overline{v}, A_0\right) \geq \mathbf{V}\left(v_H, v_H, A_0 + r_0\right) - c_0^H. \tag{A.2}$$

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

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

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

$$\mathbf{V}\left(v_H, v_L, A_0\right) \leq \mathbf{V}\left(v_H, v_H, A_0 + r_0\right) - c_0^H, \tag{A.3}$$

$$\mathbf{V}\left(v_L, v_L, A_0\right) \geq \mathbf{V}\left(v_L, v_H, A_0 + r_0\right) - c_0^L. \tag{A.4}$$

(d) Mixing by v_L: $1 = x_H > x_L > 0$, hence $\hat{v}(0) = v_L$ and $\overline{v} < \hat{v}\left(1\right) < v_H$, with

$$\mathbf{V}\left(v_L, v_L, A_0\right) = \mathbf{V}\left(v_L, \hat{v}\left(1\right), A_0 + r_0\right) - c_0^L. \tag{A.5}$$

(e) Full investment $x_H = x_L = 1$, hence $\hat{v}(0) = v_L$ and $\hat{v}\left(1\right) = \overline{v}$, with

$$\mathbf{V}\left(v_L, v_L, A_0\right) \leq \mathbf{V}\left(v_L, \overline{v}, A_0 + r_0\right) - c_0^L. \tag{A.6}$$

We can first rule out equilibria of type (b), in which type H randomizes: since $\mathbf{V}_2 > 0$, the no-investment equilibrium also exists if an equilibrium of type (b) exists. Furthermore, since $V\left(v, \overline{v}, A_0\right) > V\left(v, \hat{v}\left(0\right), A_0\right)$ for all v, both types are better off in the no-investment equilibrium, so we can apply the Pareto criterion in order to select the policy equilibrium. For the same reason, we can rule out the separating equilibrium (type (c)) whenever it coexists with the no-investment equilibrium (type (a)).

We now show that there exists a unique equilibrium, which involves no investment when (A.2) holds and, when this condition fails, separation, randomization by v_L or full investment, depending respectively on whether (A.3)-(A.4), (A.5) or (A.6) holds.

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

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

$$\mathbf{V}(v_H, \bar{\rho}v_H + (1-\bar{\rho})v_L, A_0) \equiv \mathbf{V}(v_H, v_H, A_0 + r_0) - c_0^H \tag{A.7}$$

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 (A.4) holds, the equilibrium is separating: $x_H = 1$ and $x_L = 0$. By contrast, if $\mathbf{V}(v_L, v_L, A_0) < \mathbf{V}(v_L, v_H, A_0 + r_0) - c_0^L$, the L type must invest with positive probability. If (A.6) holds there can be no solution to (A.5) with $x_L < 1$, so the only equilibrium is full investment on $[0, \bar{\rho})$. If (A.6) is reversed, on the other hand, it involves mixing: by (10),

$$\hat{v}(1) = \frac{\rho}{\rho + (1-\rho)x_L} v_H + \frac{(1-\rho)x_L}{\rho + (1-\rho)x_L} v_L, \tag{A.8}$$

and by (A.5) this expression must be independent of ρ. Thus, $x_L = (\gamma - 1)/(1/\rho - 1)$, where $\gamma = 1/\hat{\rho}(1) > 1$ 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,

$$\mathbf{V}(v_L, v_L, A_0) = \mathbf{V}(v_L, \tilde{\rho}v_H + (1-\tilde{\rho})v_L), A_0 + r_0) - c_0^L, \tag{A.9}$$

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

Proof of Proposition 2. (1)(i) When λ decreases, each type's incentive to invest, $\mathbf{V}(v, \hat{v}(1), A_0 + r_0) - \mathbf{V}(v, \hat{v}(0), A_0)$ increases: by (11), its derivative with respect to $1 - \lambda$ is

$$V(v, \hat{v}(1), A_0 + r_0) - V(v, v, A_0 + r_0) + V(v, v, A_0) - V(v, \hat{v}(0), A_0)$$

which exceeds $\int_{\hat{v}(0)}^{\hat{v}(1)} V_2(v, x, A_0)dx > 0$ if either $v = v_L$ (by the assumption $V_{23} \geq 0$), or if $v = v_H$ and $\hat{v}(1) = v_H$. Since the full-investment region (e) and the mixing region (d) are both governed by type L's incentives, the first case implies that as $1 - \lambda$ increases region (e) expands and x_L decreases in region (d). The second case implies that the no-investment region (a) shrinks.

(ii) It is easily verified from (A.7), (A.8) and (A.9) that a decrease in c_0^H increases $\bar{\rho}$ while a decrease in c_0^L decreases $\tilde{\rho}$ and reduces $\hat{v}(1)$ in the mixing region, thus increasing x_L. Thus, again investment unambiguously increases.

(iii) and (iv). In the AU case,

$$\mathbf{V}(v, \hat{v}(1), A_0 + r_0) - \mathbf{V}(v, \hat{v}(0), A_0) = s\left[\lambda v r_0 + (1-\lambda)[\hat{v}(1)(A_0 + r_0) - \hat{v}(0)A_0] + \delta v r_0\right.$$

rises with s and A_0. The rest of the proof follows the steps of part (i).

(2) The result is obvious when $x_L(\tilde{\rho}) = 0$ (separating equilibrium), since $x_L(\rho) \equiv 0$ in that case. When $x_L(\tilde{\rho}) > 0$ (equilibrium with randomization), it follows from the fact that $\widehat{v}\,(1)$ and therefore $\hat{\rho}(1) = \rho/\left[\rho + (1-\rho)x_L(\rho)\right]$ must remain constant over $[0, \tilde{\rho}]$. ■

Proof of Proposition 3 Consider (13). If $(s+\delta)\, v_L r_0 \geq c_0^L$, it is a dominant strategy for both types to invest, so $x_H = x_L = 1$ and changes in λ do not affect behavior, nor W. If $(s+\delta)\, v_L r_0 < c_0^L$ and $(s+\delta)\, v_H r_0 < c_0^H$, then $c_0^H > c_0^L > 0$ so W decreases with both x_H and x_L; a decrease in λ can therefore only (weakly) lower welfare. Finally, when $(s+\delta)\, v_H r_0 - c_0^H \geq 0 > (s+\delta)\, v_L r_0 - c_0^L$, type H always invests ($x_H = 1$); hence λ can only affect x_L, and any increase in x_L reduces welfare since $c_0^L > 0$. The proof for small changes in A_0 around the no-investment threshold (given by (14)) is similar, since the direct effect on the last term in (13) is infinitesimal, whereas the jump in x_H (and possibly x_L) is discrete. ■

Proof of Proposition 4 We construct an appropriate mixed equilibrium. Choose $c_1^* \in (0,1)$ such that $\beta\delta r_1 \bar{v} < c_1^* < \beta\delta r_1 v_H$. Next, define $v^* \in (\bar{v}, v_H)$ as $v^* \equiv c_1^*/\beta\delta r_1$ and $x_L \in (0,1)$ by

$$\hat{\rho}(1) \equiv \frac{\rho}{\rho + (1-\rho)(1-x_L)} = \frac{v^* - v_L}{v_H - v_L}. \tag{A.10}$$

Suppose now that $F(c_1)$ puts mass 1 on c_1^*; by continuity, the arguments below will continue to hold when the mass is close enough to 1. By construction, the agent invests at $t=1$ when $\hat{v} \geq v^*$. As to (A.10), it means that if the L type mixes at $t=0$ with probability x_L, the posterior following $a_0 = 1$ is exactly v^*, inducing $a_1 = 1$ for both types. Next, choose c_0^H and c_0^L such that mixing with probability x_L defined by (A.10) is indeed the equilibrium:

$$c_0^H < \delta r_0 v_H + (1-\lambda)\,(\delta r_1 v_H - c_1^*) = \mathbf{V}(v_H, v_H, A_0 + r_0) - \mathbf{V}(v_H, \bar{v}, A_0), \tag{A.11}$$

$$c_0^L \equiv \delta r_0 v_L + (1-\lambda)\,(\delta r_1 v_L - c_1^*) = \mathbf{V}(v_L, v^*, A_0 + r_0) - \mathbf{V}(v_L, v_L, A_0). \tag{A.12}$$

Compared to the equilibrium that prevails when $\lambda = 1$, in which $\hat{v} = v$ always, this yields a gain in $E[V]$ given by (16) and a loss term equal to zero; hence a positive contribution to welfare.

Turning now to period 0, in order for the equilibrium with $\lambda = 1$ to be one where neither type invests in spite of the fact that choosing $a_0 = 1$ would be be ex ante efficient for both (making the first two terms in (15) positive), if suffices that

$$\beta c_0^L < \delta v_L r_0 < \delta v_H r_0 < c_0^H. \tag{A.13}$$

Compatibility with (A.11)-(A.12) requires that $(1-\lambda)\,(\delta r_1 v_H - c_1^*) > 0$ and $(1-\lambda)\,(\delta r_1 v_L - c_1^*) < (1/\beta - 1)\,\delta v_L r_0$, neither of which contradicts any other condition. ■

Proof of Proposition 5 Since this proof is fairly long, we provide it in online Appendix C. ■

Appendix B: Extensions and Variants of the Model

• *Disappointment aversion.* Let $S(v,\hat{v},A_1) \equiv D((v-\hat{v})A_1)$ be part of the agent's date-2 payoff, where $D' > 0 \geq D$ and $-xD''(x)/D'(x) < 1$ for all x. Concavity, which means that negative surprises weigh more than positive ones, implies $S_{12} > 0$, while the elasticity condition ensures that $S_{13} > 0$ nonetheless. Thus, adding this term into the continuation value V only reinforces the sorting condition in Assumption 2, while generating a demand for "defensive pessimism". For V_2 to remain positive, this last effect must not be too strong relative to that generated by s_1. Alternatively, it could be so strong as to make V_2 negative everywhere; all that is needed is that $s_1 v + \delta D((\hat{v}-v)A_1)$ be monotonic in v over all feasible values of $v, \hat{v}$ and A_1.

• *Anticipatory utility and procrastination* Let $z_1(v) \equiv vr_1(v)$ for all v. Consider now an agent with self-view defined by $\hat{v} \in [v_L, v_H]$, or equivalently $\hat{\rho} \equiv (\hat{v}-v_L)/(v_H-v_L) \in [0,1]$. Denoting $\hat{z}_1 \equiv \hat{\rho} z_1(v_H) + (1-\hat{\rho}) z_1(v_L)$, the agent invests at $t=1$ if $\beta(s+\delta)\hat{z}_1 \geq c_1$, leading to

$$V(v,\hat{v},A_1) \equiv (s\hat{v}+\delta v)A_1 + \int_0^{\beta(s+\delta)\hat{z}_1} \left[(s+\delta)z_1(v) - c_1\right] \, dF(c_1). \tag{B.1}$$

Since $\partial \hat{z}_1/\partial \hat{v} = \left[z_1(v_H) - z_1(v_L)\right]/(v_H - v_L)$, this function satisfies $V_{13} > 0$, $V_{23} > 0$ if $s > 0$ and $V_{12} > 0$ as long as $z_1(v)$ is strictly monotonic in v, in either direction. In the case where $z_1(v)$ is decreasing, one just needs to impose conditions such that V_2 remains positive (over the relevant range).

• *Resource rivalry.* Taking for simplicity an extreme case of the investment rivalry described in Section II.B, suppose that: (a) the agent can invest in either A or B ($a_t = 1 - b_t \in \{0,1\}$), with respective returns r_{At}, r_{Bt}, salience s_A, s_B, and similar notation for other parameters; (b) his long-term values are subject to a relative preference shock: $v_A = \bar{v}_A + v/2$ and $v_B = \bar{v}_B - v/2$, where $v = \varepsilon > 0$ with probability ρ and $v = -\varepsilon$ with probability $1-\rho$. The model is then essentially isomorphic to the basic one, with all variables redefined as differentials. The relevant asset is now the row vector $A' \equiv (A \ -B)$, so that "a higher stock" means a higher A, a lower B or both (with enough parameter symmetry, only the scalar $A-B$ matters, but that need not generally be the case) and similarly for $r' \equiv (r_{At} \ -r_{Bt})$, $s' \equiv (s_{At} \ -s_{Bt})$, etc.

• *Social signaling.* The expected value function playing the role of (11) 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 $(v,\hat{v},A_1) \longmapsto V(v,\hat{v},A_1,\hat{v})$ 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.

REFERENCES

Adams, H. Wright, L., and B. Lohr (1996) "Is Homophobia Associated With Homosexual Arousal?" *Journal of Abnormal Psychology,* 105(3),440-445.

Akerlof, G., and W. Dickens, (1982) "The Economic Consequences of Cognitive Dissonance." *American Economic Review,* 72(3), 307–319.

Akerlof, G., and R. Kranton (2000) "Economics and Identity," *Quarterly Journal of Economics*, 115, 716-753.

—— (2002) "Identity and Schooling: Some Lessons for the Economics of Education," *Journal of Economic Literature*, 40(4), 1167-1201.

—— (2005) "Identity and the Economics of Organizations," *Journal of Economic Perspectives*, 19, 9–32.

Austen-Smith, D., and R. Fryer (2005) "An Economic Analysis of "Acting White"," *Quarterly Journal of Economics*, 120(2): 551–583.

Basu, K. (2006) "Identity, Trust and Altruism: Sociological Clues to Economic Development," Cornell University mimeo, April.

Battaglini, M., Bénabou, R., and J. Tirole (2005) "Self-Control in Peer Groups," *Journal of Economic Theory*, 123: 105–134.

Baumeister, R. (1986) *Identity: Cultural Change and the Struggle for Self.* Oxford: Oxford University Press.

Bem, D. J. (1972). "Self-Perception Theory," in L. Berkowitz, ed., *Advances in Experimental Social Psychology*, Vol. 6, 1-62. New York: Academic Press.

Bénabou, R. (2008) "Groupthink: Collective Delusions in Organizations and Markets," Princeton University mimeo.

Bénabou, R. and J. Tirole (2002) "Self Confidence and Personal Motivation," *Quarterly Journal of Economics*, 117(3): 871–915.

—— (2004) "Willpower and Personal Rules," *Journal of Political Economy*, 112(4): 848–886.

—— (2006a) "Belief in a Just World and Redistributive Politics," *Quarterly Journal of Economics*, 121(2), 699-746.

—— (2006b) "Incentives and Prosocial Behavior," *American Economic Review*, 96(5), 1652-1678.

—— (2009) "Over My Dead Body: Bargaining and the Price of Dignity," *American Economic Review,* Papers and Proceedings, 99(2), 459-46.

Bernheim, D. (1994) "A Theory of Conformity," *Journal of Political Economy*, 102(5), 842–877.

Bernheim, D. and R. Thomadsen (2005) "Memory and Anticipation," *The Economic Journal*, 115, 271–304.

Benjamin, D., Choi, J. and J. Strickland (2010) "Social Identity and Preferences," *American Economic Review,* forthcoming.

Bodner, R. and D. Prelec (2003) "Self-signaling and Diagnostic Utility in Everyday Decision Making," in I. Brocas and J. Carrillo eds. *The Psychology of Economic Decisions. Vol. 1: Rationality and Well-being*, Oxford University Press, 105-126.

Branscombe, N., Wann, D., Noel, Jeffrey, G. and J. Coleman (1993) "In-Group or Out-Group Extemity: Importance of the Threatened Social Identity," *Personality And Social Psychology Bulletin,* 19(4), 381-388.

Brunnermeier, M. and J. Parker (2005) "Optimal Expectations," *American Economic Review*, 95, 1092–1118.

Caplin, A., and J. Leahy (2001) "Psychological Expected Utility Theory and Anticipatory Feelings," *Quarterly Journal of Economics*, 116, 55–80.

Carlsmith, J., and A. Gross (1969) "Some Effects of Guilt on Compliance," *Journal of Personality and Social Psychology*, 11, 232–239

Carrillo, J. (2005) "To Be Consumed with Moderation," *European Economic Review*, 49, 99–111.

Carrillo, J., and T. Mariotti (2000) "Strategic Ignorance as a Self Disciplining Device," *Review of Economic Studies*, 67(3), 529–544.

Cassone, A. and C. Marchese (1999) "The Economics of Religious Indulgences," *Journal of Institutional and Theoretical Economics,* 155, 429-442.

Constant, A., Gataullina, L., and K. Zimmermann (2006) "Ethnosizing Immigrants," IZA Discussion Paper No. 2040, March.

Dal Bo, E. and M. Terviö (2008) "Self-Esteem, Moral Capital, and Wrongdoing," NBER W.P. 14508.

Dana, J., Cain, D.M., and Dawes, R. (2006). "What You Don't Know Won't Hurt Me: Costly (but Quiet) Exit in a Dictator Game," *Organizational Behavior and Human Decision Processes,* 100(2), 193-201.

Dana, J., Kuang, J., and R. Weber (2007) "Exploiting Moral Wriggle Room: Experiments Demonstrating an Illusory Preference for Fairness," *Economic Theory* 33(1), 67-80.

De Jong, H.W. (1979) "An Examination of Self-Perception Mediation of the Foot-in-the-Door Effect," *Journal of Personality and Social Psychology*, 37, 2221–2239.

Dessi, R. (2008) "Collective Memory, Cultural Transmission and Investments." *American Economic Review,* 98, 534-560.

Durkheim, E. (1976) *The Elementary Forms of the Religious Life*. 2nd edition. London: Allen and Unwin (original work: 1925).

Fang, H. and Loury, G. (2005) ""Dysfunctional Identities" Can Be Rational," *American Economic Review,* 95(2), 104-111.

Fehr, E. and S. Gächter (2000) "Cooperation and Punishment in Public Goods Experiments," *American Economic Review,* 90(4), 980-994.

Fehr, E. and K. Schmidt (1999) "A Theory Of Fairness, Competition, And Cooperation," *Quarterly Journal of Economics,* 114 (3), 817-868

Fershtman, C., Gneezy, U. and M. Hoffman (2008) "Taboos: Considering the Unthinkable," CEPR Discussion Paper No. 6854, June.

Festinger, L. and J. Carlsmith (1959) "Cognitive Consequences of Forced Compliance." *Journal of Abnormal and Social Psychology,* 58, 203–210.

Fiske, A., and P. Tetlock (1997) "Taboo Trade-offs: Reaction to Transactions that Transgress the Spheres of Justice," *Political Psychology*, 18, 255–297.

Freud, Sigmund (1919) *Totem and Taboo: Resemblances Between the Mental Lives of Savages and Neurotics.* P.86. London, UK: Routledge and Sons.

Fryer, R. and M. Jackson (2003) "Categorical Cognition: A Psychological Model of Categories and Identification in Decision Making," NBER Working Paper 9579, March.

Geanakoplos, J., Pearce, D. and E. Stacchetti (1989) "Psychological Games and Sequential Rationality," *Games and Economic Behavior* 1, 60–79.

Gottlieb, D. (2008) "Imperfect Memory and Choice under Risk," MIT mimeo.

Greif, A. (2009) "Morality and Institutions: Moral Choices Under Moral Network Externalities," Stanford University mimeo, November.

Hamman, J., Weber, R. and G. Loewenstein (2010) "Self-Interest Through Delegation: An Additional Rationale for the Principal-Agent Relationship." *American Economic Review,* forthcoming.

Herrmann, B., Thöni, C. and S. Gächter (2008) "Antisocial Punishment Across Societies," *Science,* 319, 1362-1367.

Hoge, W. (2002) "Britain's Nonwhites Feel Un-British, Report Says," *New York Times,* April 4.

Kahneman, D., and P. (1997) "Back to Bentham? Explorations of Experienced Utility," *Quarterly Journal of Economics*, 112, 75–407.

Kanbur, R. (2004) "On Obnoxious Markets", in S. Cullenberg and P. Pattanaik (eds.), *Globalization, Culture and the Limits of the Market: Essays in Economics and Philosophy.* Oxford, England: Oxford University Press.

Kay, A., Jimenez, M. and J. Jost (2002) "Sour Grapes, Sweet Lemons, and the Anticipatory Rationalization of the Status Quo," *Personality and Social Psychology Bulletin,* 9, 1300-1312.

Konow, J. (2000) "Fair Shares: Accountability and Cognitive Dissonance in Allocation Decisions," *American Economic Review,* 90(4), 1072-1091.

Köszegi, B. (2009) "Utility from Anticipation and Personal Equilibrium," forthcoming, *Economic Theory.*

Kunda Z. (1987) "Motivated Inference: Self-Serving Generation and Evaluation of Causal Theories," *Journal of Personality and Social Psychology,* 53(4), 636-647.

Kuran, T. (1996) "The Discontents of Islamic Economic Morality." *American Economic Review,* Papers and Proceedings 86(2), 438-442.

Landier, A. (2000) "Wishful Thinking and Belief Dynamics," MIT mimeo.

Lazear, E., Malmendier, U. and R. Weber (2009) "Sorting and Social Preferences," Stanford University mimeo, June.

LeBoeuf, R., Shafir, E. and J. Bayuk (2010) "The Conflicting Choices of Alternating Selves," *Organizational Behavior and Human Decision Processes,* 111(1), 48-61

Loewenstein, G. (1987) "Anticipation and the Valuation of Delayed Consumption," *Economic Journal*, 97, 666-84.

Loewenstein, G. and Schkade D. (1999) "Wouldn't It Be Nice? Predicting Future Feelings" in D. Kahneman, E. Diener and N. Schwartz, eds. *Well-Being: Foundations of Hedonic Psychology.* New York, NY: Russel Sage Foundation.

Maas, A. Cadinu, M., Guarnieri, G. and Grasselli, A. (2003) "Sexual Harassment Under Social Identity Threats: The Computer Harassment Paradigm," *Journal of Personality and Social Psychology,* 85(5), 853-870.

Mazar, N., Amir, O. and D. Ariely (2008) "The Dishonesty of Honest People: A Theory of Self-Concept Maintenance." *Journal of Marketing Research,* 45, 633-634.

Mazar, N. and Zhong, C.-B. (2010) "Do Green Products Make us Better People?", *Psychological Science,* XX(X), 1-5.

Jordan, A. and Monin, B. (2008) "From Sucker to Saint: Moralization in Response to Self-Threat." *Psychological Science,* 19, 683-689.

Monin, B. (2007). "Holier Than Me? Threatening Social Comparison in the Moral Domain." *International Review of Social Psychology,* 20(1): 53-68.

Monin, B. and Miller, D. (2001). "Moral Credentials and the Expression of Prejudice." *Journal of Personality and Social Psychology,* 81(1), 33-43.

Monin, B., Sawyer, P., and M. Marquez (2008). *"The Rejection of Moral Rebels: Resenting Those Who Do the Right Thing,"* in press, *Journal of Personality and Social Psychology,* 95, 76-93.

Oxoby, R. (2003) "Attitudes and Allocations: Status, Cognitive Dissonance and the Manipulation of Preferences," *Journal of Economic Behavior and Organization,*, 52(3), 365–385.

—— (2004) "Status, Cognitive Dissonance, and the Growth of the Underclass," *The Economic Journal,* 114(498), 727–749.

Piccione, M. and A. Rubinstein (1997) "On the Interpretation of Decision Problems with Imperfect Recall," *Games and Economic Behavior,* 20, 3-24.

Pyszczynski, T. (1993) "Cognitive Strategies for Coping with Uncertain Outcomes," *Journal of Research in Psychology,* 16, 386-399.

Quattrone, G., and Tversky, A. (1984) "Causal Versus Diagnostic Contingencies: On Self-Deception and the Voter's Illusion," *Journal of Personality and Social Psychology*, 46(2), 237-248.

Rabin, M. (1994) "Cognitive Dissonance and Social Change," *Journal of Economic Behavior and Organization*, 23, 177-194.

—— (1995) "Moral Preferences, Moral Constraints, and Self-Serving Biases," Berkeley Department of Economics Working Paper No. 95-241, August.

Ross, L. and Nisbett, R. (1991) *The Person and The Situation: Perspectives of Social Psychology.* McGraw Hill.

Rotemberg J. (1994) "Human Relations in the Workplace," *Journal of Political Economy,* 102, 684-718.

Roth, A. (2007) "Repugnance as a Constraint on Markets," *Journal of Economic Perspectives,* 21(3), 37-58.

Shayo, M. (2009) "A Model of Social Identity with an Application to Political Economy: Nation, Class and Redistribution." *American Political Science Review,* 103(2), 147-174.

Schelling, T. (1985) "The Mind as a Consuming Organ." In J. Elster (Ed.), *The Multiple Self.* New York: Cambridge University Press, 177-195.

Sen, A. (1985) "Goals, Commitment, and Identity," *Journal of Law, Economics and Organization,* 1(2), 341-355.

Smith, A. (1759) *The Theory of Moral Sentiments.* Reedited (1997), Washington, D.C.: Regnery Publishing, Conservative Leadership Series.

Staw, B. (1976) "Knee-Deep in the Big Muddy: A Study of Escalating Commitment to a Chosen Course of Action," *Organizational Behavior and Human Performance,* 16(4),27-44.

Steele, C. and J. Aronson (1995) "Stereotype Vulnerability and the Intellectual Test Performance of African Americans," *Journal of Personality and Social Psychology,* 69, 797-811.

Stutzer, A. and B. Frey (2007) "What Happiness Research Can Tell Us About Self-Control Problems And Utility Misprediction," in Bruno S. Frey and Alois Stutzer (eds). *Economics and Psychology: A Promising New Cross-Disciplinary Field.* Cambridge: MIT Press, 169-195.

Tetlock, P., Kristel, O., Elson, B, Green, M. and J. Lerner (2000) "The Psychology of the Unthinkable: Taboo Trade-Offs. Forbidden base Rates and Heretical Counterfactuals," *Journal of Personality and Social Psychology,* 78(5), 853-870.

Thompson, L. and G. Loewenstein (1992) "Egocentric Interpretations of Fairness in Negotiation," *Organization Behavior and Human Decision Processes,* 51, 176-197.

Young, P. (2006) "Self Knowledge and Self-Deception," John Hopkins University mimeo, November.

Zelizer V. (1997) *Morals and Markets: The Development of Life Insurance in the United States.* New York: Columbia University Press.

Zhong, C. B., Ku, G., Lount, R. B., Jr. and Murnighan, J. K. (2010), "Compensatory Ethics," *Journal of Business Ethics,* in press.

Online Appendix C

Proof of Proposition 5 (1) Since we focus on equilibria in which there is no investment when $\xi \leq 0$, while the H type always invests when $\xi = 1$, observing a peer who invests is necessarily good news about one's v (benchmarking on the person) and / or about the usefulness of the task, ξ. In particular, if $a^j = a^j = 1$, then necessarily $\xi = 1$. Excluding one's peer would then entail not only the direct cost of ostracism but also, if individual values are correlated, bad news about one's own v. Thus, ruling out weakly dominated strategies, as we assumed, no one chooses to ostracize. Similarly, if $a^i = a^j = 0$, staying with one's peer economizes on the cost of ostracism but also serves as a signal that $\xi \leq 0$, providing reassuring news about one's own valuation. In conclusion, ostracism can occur only when actions differ.

(2) From there on, we focus on subgames in which the two agents' actions differ, with $a^j = 1$ and $a^i = 0$.

(a) Benchmarking on the person. Note first that ostracism is a weakly dominated strategy for the agent who fails to invest: since own actions are always recalled, having chosen $a_0^i = 0$ unambiguously identifies him as an L type. For his partner, however, having chosen $a_0^j = 1$ is an imperfect signal (due to partial pooling by L's), so the deviation by i represents damaging news, given the (perfect) correlation between v^i and v^j.

The first equilibrium condition is that agent j chooses to ostracize agent i. Viewed from the point of view of j, a separation is exogenous with probability ν and endogenous with probability $(1-\nu)(1-\rho)x(1-x)$ (the probability that there is no exogenous split, that $v^1 = v^2 = v_L$ and that $a_0^j = 1$ and $a_0^i = 0$). Agent j's expected type conditional on the separation being exogenous is therefore

$$\widehat{v}(x) \equiv \frac{\rho}{\rho + (1-\rho)x} v_H + \frac{(1-\rho)x}{\rho + (1-\rho)x} v_L, \tag{C.1}$$

while it is v_L conditional on the separation being endogenous (due to his excluding i for $a^i = 0$). Let

$$\widehat{\widehat{v}}(x) \equiv \frac{\nu}{\nu + (1-\nu)(1-\rho)x(1-x)} \widehat{v}(x) + \frac{(1-\nu)(1-\rho)x(1-x)}{\nu + (1-\nu)(1-\rho)x(1-x)} v_L. \tag{C.2}$$

be the resulting expected value of v for a separation of *unknown* origin**.** Because $V_{12} \geq 0$, the condition that types H and L ostracize a noninvesting partner need be imposed only for the latter:

$$(1-\lambda)\left[V(v_L\ ,\ \widehat{\widehat{v}}(x),\ A_0 + r_0) - V(v_L\ ,\ v_L\ ,\ A_0 + r_0)\right] \geq b. \tag{C.3}$$

The second equilibrium condition relates to investment. Since values are perfectly correlated, an individual with type L knows that his partner will be of the same type, and thus invest with probability x. The individual thus invests if:

$$0 \leq -c_0^L + \lambda[V(v_L, v_L, A_0 + r_0) - V(v_L, v_L,\ A_0) + (1-\nu)xb]$$

$$
\begin{aligned}
&+ (1-\lambda)\Big\{\nu[V(v_L\ ,\ \widehat{\widehat{v}}(x),\ A_0+r_0) - V(v_L\ ,\ v_L\ ,\ A_0)] \\
&+ (1-\nu)(1-x)[V(v_L\ ,\ \widehat{\widehat{v}}(x),\ A_0+r_0) - V(v_L\ ,\ v_L\ ,\ A_0) - b] \\
&+ (1-\nu)x[V(v_L\ ,\ \overset{o}{v}(x),\ A_0+r_0) - V(v_L\ ,\ v_L\ ,\ A_0) + b]\Big\}
\end{aligned} \tag{C.4}
$$

where $\overset{o}{v}(x)$ denotes the expected value when both have invested and are still together:

$$
\overset{o}{v}(x) \ \equiv \frac{\rho}{\rho+(1-\rho)x^2}v_H + \frac{(1-\rho)x^2}{\rho+(1-\rho)x^2}v_L \geq \widehat{\widehat{v}}(x), \tag{C.5}
$$

with strict inequality for $x > 0$.

We look for conditions under which $x = 0$ (with no ostracism, $y^i = y^j = 0$) and $x = 1$ (with both types excluding any partner who fails to invest when they did, $y^j = 1$ if $(a_0^j, a_0^i) = (1, 0)$) are both equilibrium strategies for type L.

It will be useful to denote, for all λ,

$$
K_\lambda \equiv c_0^L - \lambda[V(v_L, v_L, A_0+r_0) - V(v_L, v_L, A_0)], \tag{C.6}
$$

and note that $K_\lambda \geq K_1 > 0$ by Assumption 4.

Substituting $\overset{o}{v}(0) = \ \widehat{\widehat{v}}(0) = v_H$ into conditions (C.3)-(C.4), $x = 0$ is an equilibrium if and only if

$$
b \leq \Delta_H \equiv (1-\lambda)\left[V(v_L,v_H,A_0+r_0) - V(v_L, v_L,\ A_0)\right] \leq K_\lambda + (1-\lambda)(1-\nu)b. \tag{C.7}
$$

Similarly, using the fact that $\overset{o}{v}(1) > \widehat{\widehat{v}}(1) = \overline{v}$ in conditions (C.3)-(C.4), sufficient conditions for $x = 1$ to be an equilibrium are given by

$$
\bar{\Delta} \equiv (1-\lambda)\left[V(v_L, \bar{v}, A_0+r_0) - V(v_L, v_L,\ A_0)\right] \geq \max\left\{b,\ K_\lambda - (1-\nu)b\right\}. \tag{C.8}
$$

Since $\Delta_H > \bar{\Delta}$, these conditions hold simultaneously if and only if

$$
\min\left\{(\Delta_H - K_\lambda)/(1-\lambda),\ K_\lambda - \bar{\Delta}\right\} \leq (1-\nu)b \leq (1-\nu)\bar{\Delta}. \tag{C.9}
$$

This defines a nonempty range for b if

$$
\Delta_H - (1-\lambda)(1-\nu)\bar{\Delta} < K_\lambda < (2-\nu)\bar{\Delta}, \tag{C.10}
$$

For ρ close to 1, the prior $\bar{v}$ is close to v_H, so $\bar{\Delta}$ is close to Δ_H and (C.10) therefore defines a nonempty range in R_+ for K_λ, or equivalently for c_0^L. Consequently, there exists a positive-measure range of parameters for which equations (C.7)-(C.8) hold jointly. Finally, to ensure that $a_0 = 1$ is a dominant strategy for the H type, it suffices that c_0^H be small enough (relative to

v_H and b). This concludes the proof that both equilibria coexist over a positive-measure range of parameters.

(b) Benchmarking on the situation. Consider first the event when investment is socially harmful or useless ($\xi \leq 0$). The following assumption ensures that it is an equilibrium for no one to invest, ($a_0^i = a_0^j = 0$).

Assumption 5 *For $k \in \{H, L\}$,*

$$\begin{aligned}\tilde{c}_0^k \geq -b + \lambda \left[V(v_k\ ,\ v_k\ ,\ A_0 + \xi r_0) - V(v_k\ ,\ v_k\ ,\ A_0)\right] \\ + (1-\lambda)\left[V(v_k\ ,\ v_k\ ,\ A_0 + \xi r_0) - V(v_k\ ,\ v_L\ ,\ A_0)\right]\end{aligned} \tag{C.11}$$

There are two simple conditions under which this will hold. First, provided V is unbounded (as in our *AU* and *SC* specifications), one can find a $\xi^* < 0$ that makes the right-hand side small enough, for all $\xi \leq \xi^*$. This is not necessarily very plausible, however, as it may require that investment create great harm to society and the individual's relational capital. Alternatively, we can observe that the right hand-side of (C.11) is increasing in ξ and the term multiplying $1-\lambda$ increasing in v_k ; since the costs now are $\tilde{c}_0^H \geq \tilde{c}_0^L$, a sufficient condition is that

$$\tilde{c}_0^L \geq -b + (1-\lambda)\left[V(v_H\ ,\ v_H\ ,\ A_0) - V(v_H\ ,\ v_L\ ,\ A_0)\right]. \tag{C.12}$$

We now turn to agents' strategies in the more interesting event where $\xi = 1$. We first take as given that the v_H always invests ($x_H = 1$) in that case, and focus on the behavior of the L type. We then check that $x_H = 1$ is indeed a best response by the H type.

The key idea is that since agents do not recall the realization of ξ, the possibility of the event $\xi \leq 0$ supplies a potential excuse for not investing. This excuse can only work, however, if the other agent is also not investing. Indeed, if $a_0^i = 0$ while $a_0^j = 1$, the presence of the virtuous agent j deprives agent i of an excuse. Agent i then ostracizes agent j if and only if

$$(1-\lambda)\left[V(v_L\ ,\ \widehat{\widehat{v}}(x),\ A_0) - V(v_L\ ,\ v_L\ ,\ A_0)\right] \geq b \tag{C.13}$$

where $\widehat{\widehat{v}}(x)$ now denotes i's self image following $a_0^i = 0$ and a subsequent separation. To compute this posterior, consider the following two events. With probability $1-\theta$, $\xi = 0$ so no one invests and separation occurs only exogenously, with probability ν. In this case nothing is learned about agents' types, so the posterior remains at the prior, $\bar{v}$. With probability θ, $\xi = 1$, in which case $a_0^j = 0$ occurs with probability $(1-\rho)(1-x)$ and separation occurs if there is an exogenous breakdown or if the other agent, having chosen $a_0^j = 1$, excludes agent i. When $\xi = 1$, the probability that $a^i = 0$ and a separation occurs is therefore

$$\sigma(x) \equiv (1-\rho)(1-x)\left[\nu + (1-\nu)(\rho + (1-\rho)x)\right]. \tag{C.14}$$

Consequently, agent i's self-image following $a_0^i = 0$ and a separation is given by

$$\widehat{\widehat{v}}(x) = \frac{(1-\theta)\nu}{(1-\theta)\nu + \theta\sigma(x)}\overline{v} + \frac{\theta\sigma(x)}{(1-\theta)\nu + \theta\sigma(x)}v_L. \tag{C.15}$$

Consider next the investment decision of type L when $\xi = 1$. The decision hinges on

$$\begin{aligned}
&- c_0^L + \lambda V(v_L, v_L, A_0 + r_0) + (1-\lambda)V(v_L, \widehat{v}(x),\, A_0 + r_0) + (1-\nu)[\rho + (1-\rho)x]b \\
&\gtrless \lambda[V(v_L, v_L, A_0) + (1-\nu)(1-\rho)(1-x)b] + (1-\lambda)\{[\rho + (1-\rho)x]V(v_L\ ,\ \widehat{\widehat{v}}(x),\, A_0) \\
&+ (1-\rho)(1-x)[\nu V(v_L\ ,\ \widehat{\widehat{v}}(x),\, A_0) + (1-\nu)[V(v_L\ ,\ \overset{o}{v}(x),\, A_0) + b]]\},
\end{aligned} \tag{C.16}$$

where $\widehat{v}(x)$ is still given by (C.1) while the beliefs following $a_0^i = a_0^j = 0$ are

$$\overset{o}{v}(x) \equiv \frac{1-\theta}{1-\theta+\theta(1-\rho)^2(1-x)^2}\overline{v} + \frac{\theta(1-\rho)^2(1-x)^2}{1-\theta+\theta(1-\rho)^2(1-x)^2}v_L. \tag{C.17}$$

We now look for conditions under which $x = 1$ and $x = 0$ are both equilibria, sustained respectively by L types ostracizing any partner who invests, and by no ostracism.

For $x = 1$, $\widehat{\widehat{v}}(1) = \overset{o}{v}(1) = \widehat{v}(1) = \overline{v}$, so by (C.13) and (C.16) $x = 1$ is an equilibrium if and only if

$$b \leq (1-\lambda)\left[V(v_L\ ,\ \overline{v},\, A_0) - V(v_L\ ,\ v_L\ ,\ A_0)\right], \tag{C.18}$$

$$\begin{aligned}
0 \leq\ & -c_0^L + \lambda[V(v_L, v_L, A_0 + r_0) - V(v_L, v_L, A_0)] \\
&+ (1-\lambda)[V(v_L, \overline{v}, A_0 + r_0) - V(v_L, \overline{v}, A_0)] + (1-\nu)b.
\end{aligned} \tag{C.19}$$

For $x = 0$, (C.13) becomes

$$(1-\lambda)\left[V(v_L\ ,\ \widehat{\widehat{v}}(0),\, A_0) - V(v_L\ ,\ v_L\ ,\ A_0)\right] \geq b, \tag{C.20}$$

which implies (C.18), since $\widehat{\widehat{v}}(0) < \overline{v}$ by (C.15). As to (C.16), since $\hat{v}(0) = v_H$ it takes the form

$$\begin{aligned}
&(1-\lambda)\Big\{V(v_L, v_H, A_0 + r_0) - \rho V(v_L, \widehat{\widehat{v}}(0), A_0) \\
&- (1-\rho)[\nu V(v_L\ ,\ \widehat{\widehat{v}}(0),\, A_0) + (1-\nu)V(v_L\ ,\ \overset{o}{v}(0),\, A_0)]\Big\} \\
&\leq c_0^L - \lambda[V(v_L, v_L, A_0 + r_0) - V(v_L, v_L, A_0)] - (2\rho - 1)(1-\nu)b \\
&= K_\lambda - (2\rho - 1)(1-\nu)b.
\end{aligned} \tag{C.21}$$

To summarize, both equilibria will coexist if (C.19) to (C.21) hold, together with (C.12), which ensures that no one invests when $\xi = 0$, plus a condition stating that when $\xi = 1$ the H type always finds it optimal to invest. This last requirement is automatically satisfied in the candidate

equilibrium where $x = 1$, since even the L type wants to invest. For $x = 0$, the H type must want to invest even though this will get him ostracized if his partner is an L type; by not investing, on the other hand, his action would identify him as such a type. The last condition is thus

$$c_0^H + (1-\nu)(1-\rho)b < \lambda\left[V(v_H, v_H, A_0 + r_0) - V(v_H, v_H, A_0)\right] + (1-\lambda)\left[V(v_H, v_H, A_0 + r_0) - V(v_H, v_L, A_0)\right]. \tag{C.22}$$

We now show that these five conditions are compatible over a nonempty range of parameters. For θ close to 0, $\bar{v}$, $\widehat{\widehat{v}}(x)$ and $\overset{o}{v}(x)$ are all close to $\bar{v}$ for all x, so (C.12), (C.20), (C.19) and (C.21) will respectively hold if

$$\tilde{c}_0^L + b > (1-\lambda)\left[V(v_H\ ,\ v_H\ ,\ A_0) - V(v_H\ ,\ v_L\ ,\ A_0)\right], \tag{C.23}$$

$$b < (1-\lambda)\left[V(v_L\ ,\ \bar{v},\ A_0) - V(v_L\ ,\ v_L\ ,\ A_0)\right], \tag{C.24}$$

$$K_\lambda - (1-\nu)b < (1-\lambda)[V(v_L, \overline{v}, A_0 + r_0) - V(v_L, \overline{v}, A_0)], \tag{C.25}$$

$$K_\lambda - (2\rho - 1)(1-\nu)b > (1-\lambda)[V(v_L, v_H, A_0 + r_0) - V(v_L, \bar{v}, A_0)] \tag{C.26}$$

Next, taking limits as r_0 and ν tend to 0, these four inequalities will hold for r_0 and ν small enough if

$$\tilde{c}_0^L + b > (1-\lambda)\left[V(v_H\ ,\ v_H\ ,\ A_0) - V(v_H\ ,\ v_L\ ,\ A_0)\right] \equiv (1-\lambda)\,P \tag{C.27}$$

$$b < (1-\lambda)\left[V(v_L\ ,\ \bar{v},\ A_0) - V(v_L\ ,\ v_L\ ,\ A_0)\right] \equiv (1-\lambda)Q(\rho), \tag{C.28}$$

$$c_0^L < b, \tag{C.29}$$

$$c_0^L - (2\rho - 1)b > (1-\lambda)[V(v_L, v_H, A_0) - V(v_L, \bar{v}, A_0)] \equiv (1-\lambda)R(\rho) \tag{C.30}$$

Let ε and η be small positive numbers, and define

$$b \equiv (1-\varepsilon)(1-\lambda)Q(1), \tag{C.31}$$

$$c_0^L \equiv \tilde{c}_0^L = [1 - \eta(1-\rho)]b. \tag{C.32}$$

Condition (C.28) will hold provided ε is small enough, while (C.29) holds for all $\eta > 0$, as long as $\rho < 1$. As to (C.27) and (C.30), they become:

$$(1-\varepsilon)[2 - \eta + \eta\rho)]Q(1) > P, \tag{C.33}$$

$$(1-\varepsilon)(2-\eta)Q(1) > \frac{R(\rho)}{1-\rho}. \tag{C.34}$$

One can find ε and η small enough such that (C.33) holds for all ρ close enough to 1, provided $2Q(1) > P$, that is

$$2\left[V(v_L\ ,\ v_H,\ A_0) - V(v_L\ ,\ v_L\ ,\ A_0)\right] > V(v_H\ ,\ v_H\ ,\ A_0) - V(v_H\ ,\ v_L\ ,\ A_0). \tag{C.35}$$

In the AU/SE case, $V(v\ ,\ \hat{v},\ A) \equiv (s\hat{v} + \delta v)A$, so the condition clearly holds. In the SC case, $V(v\ ,\ \hat{v},\ A) \equiv \delta vA + \int_0^{\beta\delta\hat{v}r_1}\ (\delta v r_1 - c_1)\ dF(c_1)$, so it holds if and only if

$$\int_{\beta\delta v_L r_1}^{\beta\delta v_H r_1}\ \left[\delta r_1\left(2v_L - v_H\right) - c_1\right]\ dF(c_1) > 0, \tag{C.36}$$

for which it suffices that $\delta r_1\left(2v_L - v_H\right) > \beta\delta v_H r_1$, or $v_L/v_H > (1+\beta)/2$. This condition can be imposed as long as $\beta < 1$, and it then automatically implies (6).

Turning next to (C.34) and noting that $R(1) = 0$, one can find ε and η small enough such that (C.34) holds for all ρ close enough to 1, provided $2Q(1) > R'(1)$, or

$$2\left[V(v_L\ ,\ v_H,\ A_0) - V(v_L\ ,\ v_L\ ,\ A_0)\right] > V_2(v_L, v_H, A_0)(v_H - v_L). \tag{C.37}$$

A sufficient condition (much stronger than necessary) is that $V(v\ ,\ \hat{v},\ A)$ be weakly concave in $\hat{v}$. In the AU/SE case, V is linear. In the SC case, we have

$$V_2(v, \hat{v}, A) =\ (v - \beta\hat{v})\ f(\beta\delta\hat{v}r_1)\ \beta(\delta r_1)^2\ . \tag{C.38}$$

Since $v_L > \beta v_H$ by (6), the first term is always positive, so a sufficient condition for concavity is that f be nonincreasing on its support.

The last condition to check is (C.22). Given (C.31), it will hold for r_0 and ν close to 0 if

$$c_0^H + (1-\rho)(1-\varepsilon)(1-\lambda)Q(1) < (1-\lambda)\,P. \tag{C.39}$$

Letting $0 < c_0^H < (1-\lambda)\,P$ ensures that this inequality is satisfied for all ρ close enough to 1.

This concludes the proof that conditions (C.12) and (C.19) to (C.22) are mutually consistent over a positive-measure set of parameters, ensuring the claimed multiplicity of equilibria. ■

Proof of Proposition 6 Given (21) and (22), the intertemporal utility of an agent with value $v_A \in \{v_H, v_L\}$ who starts with stocks (A_0, B_0) and chooses $b_0 \in \{0, 1\}$ is:

$$\begin{aligned}\tilde{W}(v_A, A_0, B_0, b_0) &\equiv b_0\pi\left(\delta + s\right)v_B\left(B_0 + r_B\right)\\ &+ (1-b_0)\left[\delta v_A + s\left(\lambda v + (1-\lambda)\hat{v}_A(1-b_0)\right)\right]A_0\\ &+ b_0(1-\pi)[\delta v_A + s(\lambda v + (1-\lambda)\hat{v}_A(b_0))]A_0 - b_0 c_B.\end{aligned} \tag{C.40}$$

Let us now define $a_0 \equiv 1 - b_0$, $R_0 \equiv \pi A_0$, $c_0 \equiv c_B - \pi\left(\delta + s\right)v_B\left(B_0 + r_B\right)$, common to both types, and the functions:

$$V(v, \hat{v}, A_1) \equiv c_0 + \left(\delta v_A + s\hat{v}_A\right)A_1, \tag{C.41}$$

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

It is then easy to see that (C.40) can be rewritten as

$$W(v_A, A_0, B_0, a_0) = -c_0 a_0 + \mathbf{V}(v_A, \hat{v}_A(a_0), A_0(1 - \pi) + a_0 R_0) \tag{C.42}$$

and that V is the same as in (4) apart from a constant, so it satisfies Assumption 3. Thus, although c_0 and V no longer individually correspond to the date-zero costs and date-1 expected value function (e.g., c_0 includes payoffs received at dates 1 and 2), their sum still defines the agent's objective function, with the only change with respect to the one dimensional problem being a minor one in the "fictitious" law of motion for A_t, which is now $A_1 = A_0(1-\pi) + a_0 R_0$. The depreciation" term in $1-\pi$ will not change anything (qualitatively), while the fact that the return $R_0 = \pi A_0$ now increases with the initial stock will only reinforce the fact that investment increases with A_0. The agent therefore invests at $t = 0$ if and only if

$$\mathbf{V}(v_A, \hat{v}(1), A_0(1 - \pi) + R_0) - \mathbf{V}(v_A, \hat{v}(0), A_0(1 - \pi)) \geq c_0 \tag{C.43}$$

and all results in Propositions 1 and 2 applicable to the anticipatory-utility case remain unchanged. In particular, equilibrium generally results in excessive "investment" in A, which mean suboptimally low investments in B. ■