A Cognitive Theory of Identity, Dignity and Taboos [1]

Roland Bénabou[2] Jean Tirole[3]

This version: October 2006

[1]We are grateful for helpful comments and suggestions on this project to Andrew Caplin, Glenn Weyl and participants at the Scribner lectures (Princeton 2002), the AEA Meetings (Philadelphia 2005), the Laffont memorial conference (Toulouse 2005) and seminars at Université de Paris I and the Institute for Advanced Study (2006). Bénabou gratefully acknowledges support from the National Science Foundation (SES 0424015) and the Canadian Institute for Advanced Research.

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

[3]IDEI and GREMAQ (UMR 5604 CNRS), Toulouse, and MIT.

Introduction

This paper develops a theory of why and how people "invest themselves" in a personal, professional, social or cultural identity, then uses it to analyze some of the economic and political implications of this phenomenon. The theory is cognitive in that it explicitly treats identity, dignity and related concepts as beliefs about one's deep preferences or "values" and emphasizes the self-inference process –defining oneself by one's actions– through which it operates. The affective and functional dimensions of identity are also an important part of the framework, however, as the motivations for holding particular self-views are grounded in more basic aspects of preferences documented by psychology.

People's beliefs concerning their long-term preferences and prospects have important economic effects. First, they directly impact welfare when self image (as, e.g., a caring, honorable, smart, or hard-working person) has consumption value, or when future experiences to be derived from one's economic and social assets give rise to anticipatory feelings such as savoring or dread. Second, a desired identity can be held at the cost of generating behavioral distortions, such as overconfidence; alternatively, a strong sense of self may provide clear priorities and directions that help mobilize energy and make better decisions. Third, and precisely because certain self-views are more pleasant or functional to have than others, people invest substantial resources in trying to achieve, maintain and defend these beliefs. Thinking of oneself as a moral person requires spending time and money to help others, refraining from cheating or polluting, or consuming certain products. Upholding religious beliefs requires conforming to rituals, regularly rehearsing sacred texts and abstaining from proscribed behaviors, or even thoughts. A strong occupational or political identity precludes deviating from the chosen path even when evidence indicates that the return may be quite low. Protecting one's dignity demands turning down "insulting" offers that could profitably be accepted, refusing "charity", fighting to defend one's honor or that of the clan, ostracizing and harassing norm transgressors, etc.

The importance of identity for economics has been emphasized by Akerlof and Kranton in an influential series of papers (2000, 2002, 2006). Our work naturally relates to theirs, but the two perspectives (one drawing more on sociology, the other on psychology) and modeling approaches (one emphasizing preferences, the other beliefs) are different and complementary. Our framework also provides an account of personal "commitments" (Sen (1985)) that is consistent with standard, consequentialist, economic rationality.

The starting point of our model is the recognition that people often do not know their own preferences and motives very well, and tend instead to infer them from their own actions.[1] On the

[1] See, e.g., Festinger and Carlsmith (1959) on cognitive dissonance, Kiesler et al. (1969) and especially Bem (1972) on self-perception, and Quattrone and Tversky (1984) on the self-manipulation of "diagnostic" actions. On imperfect retrospective and prospective access to one's feelings and desires, see e.g., Kahneman et al. (1997) and Loewenstein and Schkade (1999). Discussions of these phenomena and experimental evidence can be found in Bodner and Prelec (2003), Bénabou and Tirole (2004) and Battaglini et al. (2005).

demand side of belief formation, people have a desire for favorable self-views that reflects either purely affective benefits from self-image or anticipatory utility, or instrumental benefits, as when a strong identity improves motivation and helps overcome deficient willpower. These mechanisms can also be combined, allowing for interactions between wishful thinking and imperfect self-control. On the *supply side* of motivated beliefs, we model identity investments as self-signals, made possible by imperfect memory: because people have better and more objective access to their past actions than to the exact mix of motivations that led to them, they are led to judge themselves by what they do. This, in turn, implies that when contemplating choices they take into account "what kind of a person" each alternative would "make them" and the desirability of those self-views –a form of cognitive dissonance reduction. In the terminology of Hill (2006), our model thus incorporates both "identity payoffs" and the role of "identity as a perceptual lens," relating each to its psychological underpinnings.

The first part of the paper develops the basic framework and derives general results, both positive and normative, for a single dimension of identity. Three main positive results emerge. First, identity investments are higher in situations where objective information (whether confirming or disconfirming) is generally scarce, and conversely they are easily affected by even minor manipulations of salience and attention. Second, the model can account for *"escalating commitments",* in which an individual who has built up a lot of identity-specific capital (vita or bank account, close group of friends, children, family farm, familiarity with a given culture) will continue to invest in it even when the marginal return is very low. The reason is that he will be better off if he believes that these assets will yield high future utility than if he has doubts about how much he truly enjoys working, being wealthy, spending time with family, friends or relatives, etc. Continued investment is the way to "demonstrate" such preferences, but it also augments the stock of identity-specific capital, thus further raising the stakes of continued identification. This mechanism, which provides a formal account of the "self-justification" phenomenon emphasized by psychologists, tends to result in excessive specialization (e.g., work versus family) and persistence in unproductive activities.

The third result is that the intensity of investment is nonmonotonic (hump-shaped) with respect to the strength of prior beliefs. Identity-related behavior is therefore most important in situations of uncertainty over one's true values, such as those faced by adolescents, immigrants, new converts, workers in traditional sectors facing the economic and social changes brought about by globalization, etc. Most importantly, the non-monotonicity result leads to a distinctive pattern of responses to identity threats that allows us to capture and reconcile a number of previously dissonant findings from psychology. Whereas challenges to a weakly held identity (low prior) elicit *conformity* effects, effective challenges to a strongly held one (high prior) elicit possibly very strong *counter*-reactions aimed at restoring the threatened beliefs. The latter is common with religious and sexual identity (e.g., Mass et al. (2003)). It also corresponds, in the realm of "moral" identity, to the *"transgression-compliance"* effect, whereby subjects who

are led to believe that they have harmed someone show later on an increased willingness to accept requests to perform a good deed (Carlsmith and Gross (1969)). Confirmatory responses following a moderate manipulation of an identity that is relatively "fragile", on the other hand, correspond to the *"foot in the door"* effect (DeJong (1979)), in which being presented with and freely accepting an initial request for a small favor raises the probability of accepting a much more costly request in the future. The same mechanism can also help account for the *"stereotype threat"* findings of Steele and Aaronson (1995) in the academic context.

While these positive results are quite general, depending primarily on the "supply" side of the motivated-beliefs mechanism (self-signaling), the welfare consequences of identity, dignity and similar concerns depend critically on whether the "demand" side originates in anticipatory utility or self-control. In the first case, we show that identity investments always reduce ex-ante welfare, being *in fine* a form of wasteful signaling. An individual is thus always worse off with malleable beliefs or memory than with non-manipulable ones. Most strikingly, he can even be made worse off by a higher capital stock, as the escalating-commitment mechanism leads to a *treadmill effect,* in which higher levels of wealth, social status, professional achievement, etc., induce a self-defeating pursuit of the belief that happiness lies in the accumulation of those assets. By contrast, when the demand for identity stems from a self-discipline motive, more malleable beliefs and the resulting ability to invest in them can (under conditions which we characterize) raise ex ante welfare, by improving the individual's capacity to make consistent choices and persevere in the face of adversity.

These positive and normative results apply in particular to one very important set of identity-like beliefs which we examine in more detail, namely *taboos.* In contrast to economists, most societies, religions and cultures proclaim certain goods to be "priceless" or sacred: life, justice, liberty, love, friendship, one's children, democratic citizenship, religious faith, etc. For such goods, not only are markets often banned as "contrary to human dignity", but even the mere thought of up of placing a monetary value on them is considered to be appalling or sacrilegious. We show how such *"taboo tradeoffs"* (Fiske and Tetlock (1997)) can be understood as a special case of our model, in which upholding certain valued beliefs (or illusions) concerning things one "would never do" and the "incommensurable" value of certain goods requires shunning even mental comparisons which might reveal the terms of trade that could actually be obtained.

The second part of the paper examines interactions between multiple dimensions of identity, which may occur through three channels: consumption rivalry, resource rivalry, and correlation. Consumption rivalry occurs when two identities are likely to compete for resources in the future, for instance because they are associated to different lifestyles or locations. In such cases, investing in one (B) "depreciates" the other (A), as it suggests that the individual may not value it that highly. If he has substantial capital vested in A but the ultimate value of this identity is more uncertain than that of B, he may then refrain from even highly desirable investments in B, and end up worse off as a result. This *dysfunctional identity* mechanism can explain

such phenomena as resistance to structural change brought about by international trade or technological innovation, and resistance to assimilation into the local culture by immigrants and their descendents. Not investing in B in order to safeguard A can also mean actually destroying productive B capital. Such occurrences (e.g., rioting youths attacking their own schools) become more likely when individuals turn more pessimist about their chances of success for investing in the B identity –even though it remains the optimal thing to do– or when the salience of the alternative identity A is amplified by media attention or ideological manipulations.

The third part of the paper centers on some of the social dimensions of identity and dignity. We first consider peer effects, and in particular how people deal with transgressors. Peers' behavior matters here not because of payoff interdependence but because of what it may reveal, as the deep preferences and past signals of similar individuals are likely to correlated. "Deviant" behavior (violating norms and taboos, fraternizing with outsiders, reading forbidden books, cross-dressing, etc.) sends a negative signal about the value of the existing capital stock (anticipated utility version) or that of motivation-sensitive future investments (imperfect willpower version). In line with the above comparative-statics results, we show that when such transgressions effectively threaten a strongly held identity, they elicit a forceful investment response, designed to "repair" the damaged belief; this may involve excluding non-conformers to suppress the undesirable reminders created by their presence, or even harassing them, which can serve as a form of self-signaling. When the initial identity was relatively weak, on the other hand, transgressions will further "sap morale" and depress investment. In both cases, the norm violator's behavior is more significant, the more similar he is to oneself, that is, the more correlated the values. Consequently, the harshest condemnations and punishments are reserved for "insiders" who, by their words or their acts, threaten a group's valued identity (e.g., apostates and heretics).

The second main application we consider is how identity concerns such as pride, "dignity" or anticipatory feelings about one's future options lead to a *failure of Coasian agreements* in matching or bargaining situations with symmetric information. We consider partners (spouses, capital and labor, majority and minority populations) engaged in joint production and faced with a situation in which output is low (disappointing marriage, firm or economy, lost war) and they must decide whether to continue together or split the team –possibly with a resource-consuming fight. Continuation still leads to a positive surplus, but the fact that output is low means that at least one party (possibly both) has low productivity, and hence low outside prospects. Moreover, while joint output is "hard" data that is easy to remember and verify, individual contributions to it –that is, "who is to blame", or "who is getting a raw deal"– are soft signals, symmetrically observed initially but later on only imperfectly recalled by each side. In such situations, individuals with anticipatory-utility preferences or self-image concerns will suffer an identity loss if they continue to operate under equal (or a fortiori, inferior) terms in a low-productivity team. Conversely, the way to convince themselves of their better worth or of

the justness of their cause is to refuse an "insultingly low" share of the joint product and destroy the match when the other side is not willing to make enough of a concession. By doing so, each seeks to *shift the blame* on the other party and escape bleak realities to take refuge in feelings of self-righteousness or wishful hopes of "a better tomorrow", be they individual or in the form of political utopias. In equilibrium, the range of sustainable sharing rules is shown to shrink with the importance of identity concerns; beyond a point, a bargaining impasse is reached, in spite of gains from trade and symmetric information. The model is consistent with observers' accounts of the importance of self-delusion and costly "standing for one's principles" in trials, divorces, strikes, etc. (e.g., Bewley (1999)). It also matches very well, and provides a formal account of, the experimental findings of Babcock et al. (1995) and Thompson and Loewenstein (1992) on how self-serving judgments of fairness spontaneously arise and lead to costly delays and failures of bargaining.

The paper relates to two main bodies of economic literature. The first one concerns the general phenomenon of motivated beliefs, self-deception, wishful thinking and the like. We unify in the "demand" side of our model those mechanisms that are based on the consumption value of beliefs, whether due to anticipatory feelings (Akerlof and Dickens (1982), Loewenstein (1987), Caplin and Leahy (2001), Landier (2000), Brunnermeier and Parker (2005)) or a pure concern for self-image (Köszegi (2004)) and those that center more (though not exclusively) on functional motives (Carrillo and Mariotti (2000), Bénabou and Tirole (2002), (2004), (2006a), Battaglini et al. (2005), Dessi (2005)). On the "supply side" of cognitive distortion, the emphasis on the role of imperfect memory as the channel through which belief management operates builds on our earlier work in this area. The combination of anticipatory utility with imperfect recall is also the focus of Bernheim and Thomadsen (2005), while the idea of self-signaling is shared with Bodner and Prelec (2003).[2]

The second body of literature is that on identity (see Davies (2004) and Hill (2006) for recent surveys). A first approach, starting with Akerlof and Kranton (2000), emphasizes how the endogeneity and social interdependence of preferences can be structured by the choice of an identity label. Identity is thus represented as an argument in the utility function that depends on the individual's assigned or chosen social category, on the match between (exogenous) "prescriptions" for that category and the individual's given characteristics and behavior, and on his and others' actions. One of our paper's methodological contributions is to endogenize the identity payoffs, categorical prescriptions and interpersonal spillovers in Akerlof-Kranton and related frameworks (e.g., Oxoby (2003), Shayo, (2005), Smith (2005), Basu (2006)); in a different context, see also Becker and Murphy (2000)). A second line takes an evolutionary approach (e.g., Bisin and Verdier (2000), Horst et al. (2005), Wichardt (2005)). More cognitive

[2] On imperfect recall, see also Piccione and Rubinstein (1993). The parallel to social signalling games also relates our model to, e.g., Bernheim (1994) and Austen-Smith and Fryer (2005).

aspects of identity appear in only a couple of papers such as Freyer and Jackson (2003), who show optimal categorization can lead to ethnic stereotypes, and Fang and Loury (2005), who model group identity as a shared convention (akin to a language) for the transmission of information.

The paper is organized as follows. Section I develops the basic framework and general results for a single dimension of identity. Section II extends the model to multiple, interacting identities, highlighting in particular the phenomenon of dysfunctional identities. Section III turns to social aspects, focussing first on peer effects and responses to transgressions, then on how self-perception concerns can lead to bargaining impasses. Proofs are gathered in the Appendix.

I Single identity

"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."
Roy Baumeister, "Identity: Cultural Change and the Struggle for Self" (1986).

A Preferences and beliefs

There are three periods, $t = 0, 1, 2$, as illustrated on Figure I. An individual starts at date 0 with an endowment A_0 of some physical or intangible asset which we shall refer to as identity-specific capital. This could be accumulated recognition, wealth, status, good deeds (possibly religion-specific), knowledge of a language or culture, number of friends or children, stock of experiences and memories shared with them, etc. In the case of negatively valued endowments, such as social stigma, disease-causing genes or unhealthy past behaviors, our convention will be that the individual is endowed with a low stock of a positively valued attribute (say, immunity from heart disease or lung cancer).

At dates $t = 0, 1$, the individual can "invest" ($a_t = 1$), with return r_t, or "not invest" ($a_t = 0$). The new capital stock is thus

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

in which A_t is to be understood in a "net of depreciation" sense. For instance, shifts in trade or technology render specific human capital obsolete, friends and spouses may move away, scientific contributions fall out of fashion, wealth may be wiped out by market crashes or wars, etc. To economize on notation we do not represent such depreciation as $(1 - d_t)A_t$, but simply as a reduction in the level of A_t.

The "investment" action has a dual role in this model. The first one is standard accumulation: $r_t > 0$ when the stock can be increased (vita, wealth, friends), whereas $r_t = 0$ for an immutable trait (gender, race). The second and more important role is informational: even when $r_t \equiv 0$,

an individual's choice will constitute a signal about how much he values the benefits that flow from the asset A.[3]

Indeed, the central ingredient in the model is that the individual is, at times, unsure of his own deep preferences and "values" –moral standards, personal priorities, strength of faith, commitment to culture or career: the stock A_2 that he will eventually have built up may prove to be very valuable to his long-term welfare, or not that meaningful.

• *Date* 0. At the start of period 0, the individual receives a signal (intuitive feeling, conscious self-assessment, external feedback) about his type or "objective identity":

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

with $v_H > v_L$ and $\bar{v} \equiv \rho v_H + (1-\rho) v_L$ denoting the prior expectation. Conditional on v, the expected long-run utility to be derived from A_2 is vA_2. Following the signal, the individual makes his investment decision, $a_0 \in \{0,1\}$, resulting in a flow payoff $U(v, A_0, a_0)$.

Assumption 1 *The instantaneous utility $U(v, A_0, a_0)$ received by the agent at date 0 satisfies $U_{13} \geq 0$ and $U_{23} \geq 0$.*

The condition $U_{13} \geq 0$ allows us to represent the date-0 impact of investment as a type-dependent cost (or benefit if negative), $c_0^i \equiv U(v_i, A_0, 0) - U(v_i, A_0, 1)$ for $i = H, L$, such that

$$c_0^H \leq c_0^L. \tag{3}$$

When $U_{23} = 0$, as will be the case in most of our applications, the costs c_0^i are independent of the initial stock A_0.

• *Date 1.* The individual's perception of his type at $t = 1$ will often differ from what it was at $t = 0$. The usual assumption is that he receives additional information, leading to a more accurate view of his long-run preferences. For a person's past actions to define his sense of identity, on the contrary, it must be that he *no longer has perfect access* to his true initial feelings, deep motives, and introspective insights –an information *loss.* Otherwise, past behavior conveys no useful information, so there is no sense in which the individual can make (or claim to make) choices intended to "be true to myself," "stand for my principles, "keep my dignity", "maintain my integrity," "not betray my values", "be able to live with myself," and the like. As suggested by these expressions, such behaviors are quite common and in fact represent the essence of what it means to care about one's identity. Psychologists also provide extensive evidence that people's recall of their past feelings and true motivations is very imperfect and often self-serving, that

[3] This informational effect disappears when behavior is externally forced rather than chosen; our model in one which *intentions* matter critically.

they judge themselves by their actions, and that many decisions are shaped by a concern to achieve or maintain certain desirable self-views.[4]

Assumption 2 *(Self-inference). At date* 1, *the individual is aware (or reminded) of his past motivational state* 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 $\hat{v} \equiv \hat{\rho} v_H + (1-\hat{\rho}) v_L$ 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 actual signal v, and with probability $1-\lambda$ it is equal –with a slight abuse of notation– to the conditional expectation $\hat{v}(a_0) \in [v_L, v_H]$ formed on the basis of previous behavior.

This cognitive mechanism of *self-inference* can be thought of as representing the "supply side" of motivated beliefs in the model.[5] We next turn to the "demand side," which encompasses a number of mechanisms that make certain beliefs more desirable to hold than others. These include pure self-image concerns, anticipatory utility and imperfect self-control, all of which can be cast as alternative specifications of the 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* $V_{13} > 0$.[6]

The positive first derivative is mainly a "good identity" convention.[7] The conditions on the cross-partial derivatives, together with those in Assumption 1, will generate a sorting condition leading the high-valuation type to always invest at least as much as the low-valuation one, so that actions indeed have informational content.[8] Before analyzing the equilibrium, however, we show how different preferences known to generate a demand for self-serving beliefs map into the value function V. The two main examples are summarized in Figure I.

• *Example 1*: *anticipatory utility (AU).*

In period 2, the agent obtains from the stock A_2 a utility vA_2 ("primary experience"). During period 1 he derives from the prospect of that future consumption a pleasure or pain ("secondary experience") $s_1 \hat{v} A_2$, where $\hat{v}$ is his date-1 expectation of v and s_1 a "savoring" parameter. An

[4] See the references in footnote 1.

[5] It could easily be combined with other ones, such as standard learning or motivated cognition. Thus, the recall or awareness probability could be different for good and bad signals, $\lambda_H \geq \lambda_L$, whether exogenously or endogenously (see Bénabou and Tirole (2002, 2004)). We focus here on the case in which $\lambda_H = \lambda_L$ for simplicity and to highlight the role of self-inference, which seems most relevant to "identity".

[6] When $r_0 = 0$ (immutable characteristic), no condition on V_{13} is necessary.

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

[8] Since a_t and A_t can, like v and $\hat{v}$, always be redefined as their opposites, all one really needs is that there exist $(\varepsilon, \eta) \in \{-1, 1\}^2$ such that the functions $U(\varepsilon v, \eta A_0, \eta a_0)$ and $V(\varepsilon v, \varepsilon \widehat{v}, \eta A_1)$ satisfy Assumptions 1 and 3. We shall make use of such transformations in some of the examples presented below.

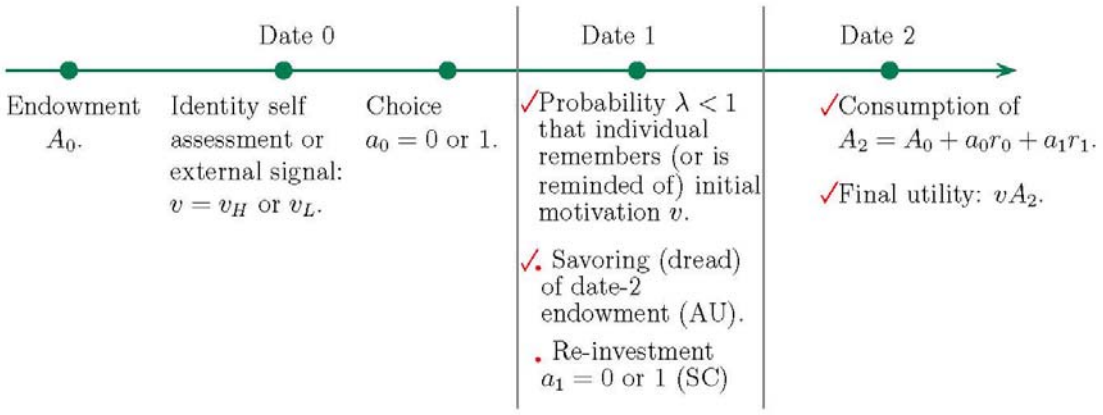

Figure I: Timing of moves and actions

important determinant of s_1 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.

We focus here on the "pure" anticipatory-utility case, in which there is no further decision to be made at date 1 (Example 3 will add a re-investment). Thus $a_1 = 0$, $A_2 = A_1$ and the continuation value of entering period 1 with subjective identity $\hat{v}$ is

$$V(v, \hat{v}, A_1) \equiv (\delta_1 s_1 \hat{v} + \delta_2 v) A_1, \tag{4}$$

where δ_1 and δ_2 reflect standard time discounting (back to $t = 0$), with possibly different lengths of periods 1 and 2. It is clear from (4) that Assumption 2 is satisfied, with $V_{13} > 0$, $V_{23} > 0$ and $V_{12} = 0$.

Turning now to date-0 payoffs, let

$$U(v, a_0, A_0) = -ca_0 + \tau v a_0 + s_0 v(A_0 + a_0 r_0), \tag{5}$$

The first term is a time, effort or monetary cost, independent of type. The second one represents the consumption benefits derived or "sampled" in the process of investment: going out with friends to make more friends, spending time with family to increase the stock of shared experiences, attending church to strengthen one's religious beliefs, working the farm to improve the land, etc. The third term arises when the agent derives anticipatory utility at $t = 0$, as he will at $t = 1$, from his long-term ($t = 2$) consumption prospects. These last two terms capture intuitive effects that make identity investments less costly, or more pleasant, for the high-valuation type. Thus Assumption 1 is satisfied, with $U_{13} = U_{23} = 0$ and investment costs $c_0^i = c - [s_0 r_0 + \tau] v^i$ such that $c_0^H < c_0^L$.

Finally, when performing welfare analysis, our criterion will be total intertemporal utility

$$W \equiv E[U + V], \tag{6}$$

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

This simple benchmark model easily accommodates a number of extensions.

a) Disappointment aversion. Whereas savoring provides a motive to be optimistic about how satisfying the future will be, the fear of being disappointed when consumption from A_2 $(= A_1)$ actually occurs generates an opposing incentive to maintain low expectations. Let $S(v, \hat{v}, A_1) = \delta_2 \varphi((v-\hat{v})A_1)$ represent the corresponding period-2 payoff, where φ is an increasing and concave function such that $-x\varphi''(x)/\varphi'(x) < 1$ for all x. Concavity, which means that negative surprises are more unpleasant than positive ones (see Gul (1991)), implies that $S_{12} > 0$, 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".[9]

b) Utility from memories or pure self-image. So far, the utility derived from beliefs was anchored on some final consumption through which the true of value of v will be directly experienced. Pure "mental consumptions" (Schelling (1985)) are just a special case in which this moment of truth never comes. Thus $\delta_2 = 0$, and hence $V(v, \hat{v}, A_1) = \delta_1 \hat{v} A_1$ corresponds to an individual who simply cherishes memories of how useful, productive, or generous he has been. If the stock corresponds to a fixed personal trait $(A_1 = A_0)$, moreover, this specification (or any nonlinear variant) captures a pure demand for-self-esteem with respect to intelligence, attractiveness, and the like.

c) Wishful thinking impairs later decisions. The savoring motive will, as we shall see, lead individuals to distort their initial $(t = 0)$ decisions in the pursuit of more pleasant beliefs. Once beliefs have been altered, moreover, any subsequent decision-making will also be impaired as a result. Example 3 below will incorporate this effect by allowing investment to also take place at $t = 1$, with the main difference being that the value function becomes nonlinear in beliefs.[10] It will also integrate the effects on period-1 decisions of anticipatory utility with those of imperfect self-control, to which we now turn.

[9] For V_2 to remain positive, this 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; see footnote 8. What is needed is that $\delta_1 s_1 v + \delta_2 \varphi((\hat{v}-v)A_1)$ be monotonic in v over all feasible values of $v, \hat{v}$ and A_1.

[10] To capture phenomena such as anxiety or focused savoring, Caplin and Leahy (2001) and Köszegi (2003) posit hedonic benefits from anticipation that are non-linear in probabilities. Our *positive* results (Propositions 1 and 2) apply as well to such date-1 payoffs $\pi(\hat{v}, A_2)$, as long as $\pi_1 > 0$ and $\pi_{12} > 0$. As Propositions 3 and 4 make clear, however, *normative* conclusions depend importantly on linearity or the specific form of nonlinearity.

• *Example 2: imperfect self-control (SC)*

Whereas individuals with anticipatory or self-esteem preferences want to hold certain beliefs for purely *affective* reasons, having a strong, stable sense of identity is also very valuable for making consistent choices and persevering in long-term projects. This *functional* motive, equally stressed by psychologists, leads to our second main benchmark.

As before, the stock A_2 generates consumption benefits vA_2 at date 2, but now accumulation can take place both at $t = 0$ and at $t = 1$. The latter involves a cost c_1 (which, for simplicity, we take to be type-independent),[11] with

$$\delta_2 v_L r_1 > \delta_1 c_1, \tag{7}$$

where δ_1 and δ_2 have the same interpretation as above. Ex-ante, it is therefore always efficient to invest at $t = 1$, even for someone with relatively low valuation for the identity-related good. Come date 1, however, weakness of will can make the immediate disutility of effort much more salient than the distant benefit, giving rise to a self-control problem. Let the individual's "Self 1" thus perceive the current cost as equal to c/β_1, where the willpower (time-consistency) parameter β_1 is drawn at $t = 1$ from a continuous distribution F on $[0, 1]$.[12] Given a posterior belief $\hat{v}$, the individual invests at $t = 1$ only if

$$\beta_1 \delta_2 \hat{v} r_1 \geq \delta_1 c_1, \tag{8}$$

which defines a cutoff level of β_1 that decreases with $\hat{v}$. The continuation value is thus

$$V(v, \hat{v}, A_1) \equiv \delta_2 v A_1 + (\delta_2 v r_1 - \delta_1 c_1) \left[1 - F\left(\frac{\delta_1 c_1}{\delta_2 \hat{v} r_1}\right)\right], \tag{9}$$

which again satisfies all the conditions of Assumption 2, with $V_3 = 0$. In period 0, finally, let the corresponding payoff U, as perceived contemporaneously (i.e., by "Self 0"), be defined as in (5) but with $s_0 = 0$, resulting in net investment costs $c_0^H \leq c_0^L$ that again satisfy Assumption 1.

With regard to welfare analysis, it may no longer be appropriate to just add up (the expectations of) U and V, as the individual may suffer from present–biased preferences at date 0, as he does at date 1. Suppose that his perceptions of contemporaneous payoffs are magnified by $1/\beta_0$, where $\beta_0 \leq 1$ measures willpower at $t = 0$ (one could easily make it stochastic, as with β_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 $\beta_0 c_0$. Recalling that V is a value function and therefore (unlike

[11] More generally, it suffices that c_1 either be only imperfectly informative about v, or that the agent need to make the $t = 1$ investment decision before having experienced the full cost.

[12] Alternatively, it could be the date-1 cost c_1 that is unknown at date 0. The role of uncertainty here is only to smooth over $t = 1$ decisions so as to make V differentiable (which we use only to simplify the exposition).

U) not subject at date-0 to salience of the present, our welfare criterion will be:

$$W = E[\beta_0 U + V], \tag{10}$$

where, as before, the expectation is taken with respect to the prior distribution of types and the posterior distribution of beliefs.

• *Example 3: wishful thinking and procrastination*

When does the desire to indulge in pleasant beliefs and avoid unpleasant ones aggravate the self-control problem, and when does it alleviate it? The AU and SC models are easily integrated together within our framework, allowing us to show in particular how the answer to this question depends on whether effort and "identity" are *substitutes* or *complements.*

Let us simply combine the two previous preference specifications and generalize the investment technology to allow for type-dependent returns. We denote those as $r_t(v)$ and the resulting contributions to final utility vA_2 as $z_t(v) \equiv vr_t(v)$, $t = 1, 2$. For an agent with self-view $\hat{v} \in [v_L, v_H]$, or equivalently $\hat{\rho} \equiv (\hat{v} - v_L)/(v_H - v_L) \in [0, 1]$, the corresponding marginal expected utility is then

$$z_t(\hat{v}) \equiv \hat{\rho} z_t(v_H) + (1 - \hat{\rho}) z_t(v_L). \tag{11}$$

He invests at $t = 1$ if $\beta_1 (\delta_1 s_1 + \delta_2) z_1(\hat{v}) \geq \delta_1 c_1$, leading to

$$V(v, \hat{v}, A_1) \equiv (\delta_1 s_1 \hat{v} + \delta_2 v) A_1 + [\delta_1 s_1 z_1(\hat{v}) + \delta_2 z_1(v) - \delta_1 c_1] \left[1 - F\left(\frac{\delta_1 c_1}{(\delta_1 s_1 + \delta_2) z_1(\hat{v})}\right)\right], \tag{12}$$

which again satisfies $V_{12} > 0$ as long as z_1 is either increasing or decreasing; moreover, $V_{13} > 0$, and $V_{23} > 0$ if $s_1 > 0$.[13] Note that when $z_1(v)$ is increasing, it is again the more optimistic agents who invest; when $z_1(v)$ is decreasing, on the contrary, they are the ones who think they can afford to "coast".[14] This dichotomy maps into two main types of situations, and related forms of identity, in which anticipatory utility and imperfect willpower combine very differently to affect behavior and welfare.

a) Wealth accumulation, status-seeking, and other entrepreneurial behaviors (complementarity). When r_1 is type-independent (as with financial assets) or when v corresponds to an ability variable that increases the expected value of effort $z_1(v)$ (e.g., raising the probability of winning in a competitive situation, or the market value of an invention), wishful thinking can help alleviate the self-motivation problem (if there is one; otherwise, it only results in excessive activism). Thus, dreams of riches and glory –and of how enjoyable those will be– propel entrepreneurs,

[13] In the limiting case in which there is anticipatory utility but no present bias, the term in $1 - F$ becomes $\mathbf{1}_{\{\delta_1 s_1 + \delta_2 z(\hat{v}) \geq \delta_1 c_1\}}$, which is non-differentiable but retains the key increasing-differences properties. It is then easily seen from (12) that distorted beliefs, $v \neq \hat{v}$, *always* lead to (weakly) suboptimal decisions at $t = 1$, namely overinvestment or underinvestment, depending on $z_1(\hat{v}) \gtrless z_1(v)$.

[14] In the case $z_1' < 0$, one needs to impose conditions such that V_2 remains positive (over the relevant range).

explorers, athletes and scientists to significant sacrifices and persistence in the pursuit of risky, long-term endeavors.

b) Health investments, safe driving and other risk-prevention behaviors (substitutability). In such cases $z_1(v)$ is decreasing in v, which, depending on the context, can be a favorable genetic endowment that protects from disease and makes taking care of one's health less of a necessity, or good driving skills and reflexes that permit faster speeds.[15] Wishful thinking then leads to a "care-free" complacency that further impairs decision-making, aggravating the self-control problem. Thus, understating the likelihood of illness, accident or death makes the present more enjoyable but further encourages negligent behaviors –unhealthy lifestyle, addictions, careless driving, failing to save for old age– that are precisely those to which weakness of will already makes one tootempted to succumb.[16]

Finally, we again specify date-0 payoffs as in (5), leading to effective costs of investment $c_0^i = c - \left[s_0 r_0(v^i) + \tau\right] v^i$ for type $i = H, L$, with: $\tau \geq 0$ in case (a), capturing as before the idea that in the process of accumulating social status or good deeds, mastering a culture, etc., some consumption benefits inherent to the activity are derived; $\tau \leq 0$ in case (b), meaning that investing in (say) health confers more immediate benefits (relief from pain, weight loss, preventing the onset of a latent crisis) to the low-immunity type than to the high-immunity one.[17]

B Equilibrium

At date 0, each type chooses his action optimally, taking into account the impact that may result for his sense of identity at date 1 and the affective and/or functional payoffs that derive from it. Thus a_0 is a solution to

$$\max_{a_0 \in \{0,1\}} \left\{U\left(v, A_0, a_0\right) + \lambda V\left(v, v, A_0 + a_0 r_0\right) + (1-\lambda) V\left(v, \hat{v}\left(a_0\right), A_0 + a_0 r_0\right)\right\}, \tag{13}$$

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 probabilities that types v_H and v_L respectively invest at $t = 0$, this means

[15] In the health case, for instance, the individual's long-term health is $vA_2 = vA_0 + z_0(v)a_0 + z_1(v)a_1$, where v is his endowment of "good" genes and A_0 a constant that can be normalized to 1.

[16] On how denial of death impairs decision-making, see Becker (1973) and Kopczuk and Slemrod (2005).

[17] Thus $s_0 z_0(v) + \tau v$ is increasing in v in case (a), satisfying Assumption 1, and decreasing in case (b), contributing to a "reverse" sorting condition that will make the v_L type more likely to invest at $t = 0$, as he is at $t = 1$. In the latter case one can just redefine "identity investment" as $b_t = 1 - a_t$ (see footnote 8).

[18] The problem we study thus has the structure of a dynamic "psychological game" (Geanakoplos et al. (1989)) between the individual's time 0 and time 1 "selves". By modeling agents as Bayesian, and thus conscious at date 1 that they sometimes make decisions so as to maintain or enhance a valued identity, we are treating them as relatively sophisticated. One can easily relax, in part or totally, this "metacognition" assumption (see, e.g., Bénabou and Tirole (2002, 2004)). This would not change any of the model's positive results (it would actually make them stronger), but in certain cases it would lead to different welfare implications (see footnote 25).

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

$$\hat{\rho}(1) = \frac{\rho x_H}{\rho x_H + (1-\rho) x_L} \quad \text{and} \quad \hat{\rho}(0) = \frac{\rho(1-x_H)}{\rho(1-x_H) + (1-\rho)(1-x_L)} \tag{14}$$

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

$$\mathbf{V}(v, \hat{v}, A_1) \equiv \lambda V(v, v, A_1) + (1-\lambda) V(v, \hat{v}, A_1), \tag{15}$$

which brings together the *demand* (preferences) and *supply* (cognition) sides of the model and inherits from V all the properties in Assumption 3. Investing at $t = 0$ is thus an optimal strategy for type $v_i \in \{v_H, v_L\}$ if

$$\mathbf{V}(v_i, \hat{v}(1), A_0 + r_0) - \mathbf{V}(v_i, \hat{v}(0), A_0) - c_0^i \geq 0. \tag{16}$$

There are three reasons why this net return is typically greater for the v_H type than the v_L type (a sorting condition), implying that if $x_L > 0$ then $x_H = 1$ (hence $\hat{v}(1) \geq \hat{v}(0)$ on the equilibrium path). First, the high-valuation type has a lower effective cost, $c_0^H \leq c_0^L$. Second, when $V_{13} > 0$, he attaches greater value to any addition to the capital stock. Finally, when $V_{12} > 0$ he also cares more about having a "strong" identity at date 1, which investing helps achieve if $\hat{v}(1) > \hat{v}(0)$.[19]

From now on, we shall restrict attention to *monotonic equilibria,* defined as those in which: i) the high-value type always invests more: $x_H \geq x_L$, which given (16) again means that $x_H = 1$ whenever $x_L > 0$; ii) 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.[20]

Finally, over a certain range of parameters there may be multiple (three) monotonic equilibria, among which one is Pareto-dominant and will be the one selected.[21]

[19] Formally, the difference between the two types' net incentives to invest in (16) equals

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

If $V_{12} = 0$ (as is the case for anticipatory utility), then Δ is always strictly positive, so any equilibrium must have the monotonicity property, $x_L(1 - x_H) = 0$. When $V_{12} \geq 0$ the same holds provided $\hat{v}(1) \geq \hat{v}(0)$, but since those beliefs are endogenous, less intuitive forms of equilibria may also exist.

[20] It is implied for instance by Cho and Kreps' (1987) Never a Weak Best Response (NWBR) criterion if $V_{12} = 0$ (as is the case for *AU*). The proof of this result is similar to that of Lemma 1.

[21] An equilibrium Pareto dominates another equilibrium if it yields a weakly higher payoff to both types and a strictly higher payoff to at least one of them. While the possibility of inferior "self-traps" is not uninteresting, we have studied it in some other context (Bénabou and Tirole (2002)) and will abstract from it here.

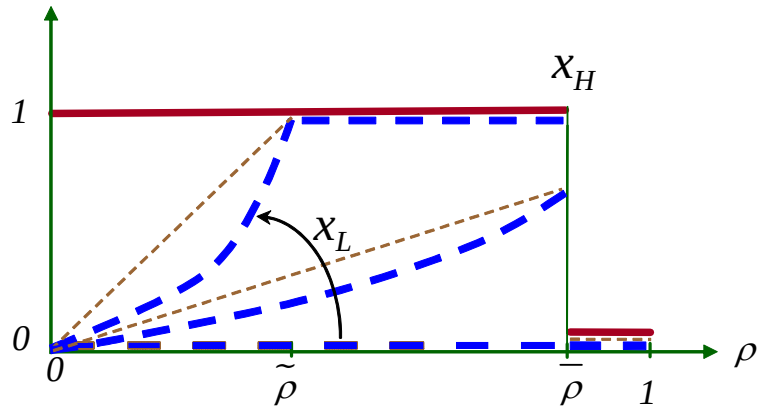

Figure II: Equilibrium as a function of ρ, for decreasing values of c_0^L. Solid line = $x_H(\rho)$, thick dashed line = $x_L(\rho)$, thin dashed line = average investment $\overline{x}(\rho)$.

Proposition 1 *There exists a unique (monotonic, undominated) equilibrium, characterized by thresholds* $\tilde{\rho}$ *and* $\bar{\rho}$ *with* $0 \leq \tilde{\rho} \leq \bar{\rho} \leq 1$ *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* $\rho \geq \bar{\rho}$;
(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 the case where $0 < \bar{\rho} < 1$ and for decreasing values of c_0^L, so as to illustrate all the cases of interest:

(i) no investment: when ρ is high enough ($\rho > \bar{\rho}$), the v_H type can afford not to invest, knowing that since the other type also abstains, the posterior will equal the prior, which is already close to 1 and thus could not be increased much anyway.

When initial self-confidence is below the threshold $\bar{\rho}$,on the other hand, the v_H type needs to invest in order to "affirm his values" and separate from the more common v_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 never finds it worthwhile to invest ($\tilde{x} = 0$), whereas the high-valuation does, for $\rho < \bar{\rho}$;

(iii) randomization by v_L : for lower values of c_0^L, it becomes desirable for the v_L type to imitate the v_H type, but his ability to do so profitably is limited by the initial prior ($0 < \tilde{x} < 1, \tilde{\rho} = \bar{\rho}$). The lower is ρ, the more truthful (low x_L) his strategy must be in order for investment to signal a high type with sufficient credibility (see (14)).

(iv) full investment: for c_0^L still lower, even a small signaling gain is profitable, so the low-valuation type can afford to completely pool with the other one ($\tilde{x} = 1$), provided ρ is above some threshold $\tilde{\rho}$ (which decreases with c_0^L).

Having fully characterized equilibrium behavior, we now turn to comparative-statics predictions, then draw their implications. We shall say that the individual is "more likely to invest in

identity" if both x_H and x_L (hence also the total probability of investment $\rho x_H + (1-\rho)\, x_L$) (weakly) increase.[22]

Proposition 2 *(1) The individual is more likely to invest in identity:*
(i) the more malleable his beliefs (the lower λ),
(ii) the lower the investment cost (the lower c_0^L or c_0^H),
(iii) the more salient the identity in the AU case (higher s_1).
(iv) in the presence of AU, and more generally when $V_{23} \geq 0$, the higher the identity-specific capital A_0.
(2) Initial beliefs have a nonmonotonic, hump-shaped, effect on the overall probability of investment: it increases linearly in ρ on $[0, \tilde{\rho})$, equals 1 on $[\tilde{\rho}, \bar{\rho})$, then falls to 0 beyond.

C Applications

The comparative statics derived in Proposition 2 have a number of interesting implications.

*1) Malleability of beliefs. A*n increase in the probability λ that the individual remains aware, or is reminded of, his true motives and values, reduces the level of investment. Identity-management is thus more likely to occur in environments in which objective (and especially, dissonant) information is scare, such as closed, tight-knit communities where most members share the same endowments and objectives.

2*) Salience of identity.* Messages or cues that "remind" individuals of specific components of their identity will elicit investments along the same dimensions. LeBoeuf and Shafir (2004) thus find that even minor manipulations emphasizing alternative aspects of subjects' identity, such as scholar versus socialite, or ethnic Chinese versus American citizen, trigger identity-consistent consumption choices.

3) Escalating commitment. The more identity-relevant capital they have, the more identity-affirming investment people will make, thereby raising the stock even further. This result is not due to any increasing returns in the investment technology: in our benchmark model, $U_{23} = 0$. The reason is instead that someone with more A_0 has a greater vested interest in viewing this asset as valuable rather than worthless, and further investment is the way to demonstrate such beliefs –as in the psychology literature on self-justification. Thus, a farmer faced with adverse market or personal signals may obstinately refuse to quit rather than admit that his efforts and sacrifices (or those of his parents) have been in vain. A manager may keep throwing good money after bad on a doomed project (as in the original "escalating commitment" experiments of Staw (1976)). 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 value

[22] Given the results of Proposition 1, illustrated in Figure II, the fact that (for all ρ) x_H increases means that $\bar{\rho}$ decreases, and the fact that (for all ρ) x_L increases means that either $\tilde{x} < 1$ increases or $\tilde{x} = 1$ and $\widetilde{\rho}$ decreases.

of earlier ones –that is, to safeguard or strengthen the belief (true or false) that these assets will bring happiness over the course of their lifetime, or a favorable fate in some hereafter.

The escalating commitment result relies on V_{23} being positive, meaning that individuals have a higher demand for optimistic beliefs when they have more at stake. This assumption has substantial empirical support. For instance, 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 large changes in tuition (or either sign). Kunda (1987) had subjects read a (bogus) medical article linking cumulated caffeine consumption to risks of fibrocystic disease and breast cancer. Among the women, heavy coffee drinkers judged the information to be significantly less credible than light drinkers, whereas the men (who were not "at risk" of either disease and thus served as a control group) showed no such difference.[23] In the field, finally, there is the well-known Stockholm syndrome, whereby people taken hostage often come to see their captors in a favorable light, most plausibly so as to maintain hope that they will not harm them.

4) Uncertain values. The high-valuation type's incentive to "prove himself", which exists only when he is sufficiently insecure ($\rho < \bar{\rho}$), and the low-valuation type's incentive to "pass", which can be effective only when he has sufficient self-confidence ($\rho > \tilde{\rho}$), combine to make the overall (ex ante) probability of investment hump-shaped with respect to ρ. This means, first of all, that identity-affirming investment are characteristic of people with uncertain or unsettled preferences and values: hence the zeal of the new convert (religious or political), the fierce nationalism or culturalism of the recent immigrant (whether in favor of his new country or the old one), and the shifting but always uncompromising attitudes and ritual codes of the adolescent. People who are confident of "who they are", on the contrary, have no use for purely identity-affirming investments (they invest only if r_0 is large enough to justify the cost). Second, the model's predictions can help understand and unify a number of disparate or seemingly contradictory findings from the psychology literature on people's responses to manipulations of their self-image.

a) Substantial *identity threats* trigger large opposing responses aimed at restoring the damaged self-image –as occurs in the model when ρ is caused to fall below $\bar{\rho}$. In Maas et al. (2003), for instance, males 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

[23] A similar pattern is that HIV-seropositive individuals tend to underestimate the incidence of full-blown AIDS in the seropositive population, relative to HIV-seronegative individuals. (In this case, however, reverse correlation is also likely to be at work).

rather than a meek, family oriented traditionalist; it was also more pronounced, the more the subjects had initially self-rated themselves as masculine. Turning now from gender identity to "good person" identity, many experiments (e.g., Carlsmith and Gross (1969)) have documented the *"transgression-compliance"* effect, whereby subjects who are led to believe that they have harmed some other person (e.g., by administering painful electric shocks, or by carelessly ruining some of her work) show an increased willingness to later on accept requests to perform a "good" action, even when the requester is not their "victim" and does not even know about their (supposedly) harmful deed.

b) More subtle challenges to, or affirmations of, an identity that is desirable but relatively unfamiliar or "fragile", on the other hand, are likely to lead to confirmatory rather than fighting responses –as occurs in the model when ρ changes marginally, starting from a value below $\tilde{\rho}$. Such is the case with the *"foot in the door"* effect (see DeJong (1979) for a survey), whereby being presented with and freely accepting an initial request for a small favor raises the probability of accepting a much more costly request in the future.[24] Finally, the model can also account for the *"stereotype threat"* findings of Steele and Aaronson (1995). A social stereotype of female or African-American students as having a lower distribution of (say) comparative mathematical abilities than their White or Asian counterparts means precisely that "society" places a lower subjective prior on their being a high type (with v now representing ability rather than taste, or a combination of both). Making gender or race subtly more salient before a test reminds the subjects of this statistical perception and thus (consciously or unconsciously) lowers their own 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.

D Identity and welfare: treadmill effect or empowerment?

While equilibrium behavior and most comparative results are very general, relying only on Assumptions 1-3, the welfare implications of identity management depend critically on whether the demand originates in anticipatory utility or self-control.

- *Anticipatory utility and the treadmill effect.*

  Equations (4)-(6) lead to

$$W = \rho x_H \left[(\delta_1 s_1 + \delta_2) v_H r_0 - c_0^H \right] + (1-\rho) x_L \left[(\delta_1 s_1 + \delta_2) v_L r_0 - c_0^L \right] \quad (17)$$

$$+ \left[s_0 + \delta_1 s_1 + \delta_2 \right] \overline{v} A_0. \quad (18)$$

The last term is constant: although agents actively manage their beliefs, the average self-view

[24] Conversely, an initial costly request, which most people turn down, decreases the average probability of accepting a later smaller one. In neither case are the results due to self-selection, since the probabilities being compared are the average compliance rates between all members of the experimental group (who get two requests) and a control group (who get only the second request).

remains fixed, by the law of iterated expectations.[25] 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 for an immutable characteristic like gender, race, or nationality: with $r_0 = 0$, there remains only a loss of $-\rho x_H c_0^H - (1-\rho) x_L c_0^H$. The result (wasteful signaling) applies equally with an accumulable asset, however.[26]

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

$$\mathbf{V}(v_H, v_H, A_0 + r_0) - \mathbf{V}(v_H, \overline{v}, A_0) = (\delta_1 s_1 + \delta_2) v_H r_0 + (1-\lambda)\,\delta_1 s_1 (v_H - \bar{v})\, A_0 \leq c_0^H,$$

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. More generally, the model yields a type of *treadmill effect:* higher levels of wealth, social status, professional achievements, etc., do no generate much of an increase in life satisfaction, or may even reduce it –and this precisely due to a self-defeating pursuit of the belief that these assets will really bring happiness.[27]

Proposition 3 *In the anticipatory utility case,*
(1) An increase in the malleability of beliefs $(1-\lambda)$ always reduces welfare.
(2) An increase in the identity-specific capital A_0 can also make the individual worse off.

Note that the same result shown for A_0 holds for the salience of anticipatory feelings and other self-image concerns, s_1. Therefore identity-targeted *advertising,* which raises s_1, can be quite effective in inducing consumers to purchase ($a_0 = 1$) but at the same time significantly lower their average welfare –and even social welfare, if the costs borne by consumers include not only the price of the goods purchased, but also time or other opportunity costs.

• *Willpower and the commitment value of identity*

In the self-control version of the model, A_0 has no behavioral impact (unless some complementarity with a_0 is assumed), as (9) shows that $V_{23} = 0$. The malleability of self-image, on the other

[25] Conversely, for investments in anticipatory feelings to yield welfare gains, it must be that either: (i) agents' updating is at least partially naïve: when $a_0 = 1$, they do not properly correct for pooling by the v_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., along the lines in Bénabou and Tirole (2002)); or (ii) the savoring value of beliefs is nonlinear (and thus not purely anticipatory in the standard sense), as in some cases considered by Caplin and Leahy (2001) and Köszegi (2003).

[26] If $(\delta_1 s_1 + \delta_2) v_L r_0 \geq c_0^L$, 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 $(\delta_1 s_1 + \delta_2) v_H r_0 < c_0^H$, then W decreases with both x_H and x_L, so a decrease in λ can only (weakly) lower welfare. Finally, when $(\delta_1 s_1 + \delta_2) v_H r_0 - c_0^H \geq 0 > (\delta_1 s_1 + \delta_2) v_L r_0 - c_0^L$, type v_H always invest ($x_H = 1$); hence λ can only affect x_L, and any increase in x_L reduces welfare.

[27] Our is thus a different mechanism and explanation for treadmill effects from the traditional one, which is based on preferences or "aspirations" adapting to changes in consumption levels. Note also that, in this model, agents in the treadmill zone rationally know that they are trapped ("money does not buy happiness"), but cannot refrain from overinvesting. Ex-ante, they would then like to reduce A_0, which may or may not be feasible (one can give away wealth, but it is harder to dispose of past achievements or an immutable identity).

hand, now affects behavior both at $t = 0$ and (through posterior beliefs) at $t = 1$. To demonstrate the key results most simply, we shall compare here intertemporal welfare, $W = E[\beta_0 U + V]$, when $\lambda = 1$ and when $\lambda < 1$.

Date-1 behavior. If the equilibrium involves separation ($x_H = 1, x_L = 0$) at $t = 0$, the individual's self-view at $t = 1$ is always the same as with $\lambda = 1$, hence so are behavior and welfare. More interestingly, when some pooling occurs there will be states of the world at $t = 1$ in which the agent has imperfect knowledge of his type. This uncertainty *boosts* the self-confidence and propensity to invest of the v_L type, but simultaneously *weakens* those of the v_H type. Under (7), $a_1 = 1$ is optimal for both, so the first effect leads to a welfare gain, the second to a loss.

Suppose, for instance, that the equilibrium with $\lambda < 1$ involves mixing by the low-valuation type: $0 < x_L < x_H = 1$, so $\hat{v}(0) = v_L < \overline{v} < \hat{v}(1) < v_H$.[28] The difference, between the malleable and nonmalleable beliefs cases, in the contributions to intertemporal welfare of date-1 investments is then $1 - \lambda$ times

$$\begin{aligned} \Delta V &= (1-\rho)x_L \left[F\left(\frac{\delta_1 c_1}{\delta_2 r_1 v_L}\right) - F\left(\frac{\delta_1 c_1}{\delta_2 r_1 \hat{v}(1)}\right) \right] (\delta_2 v_L r_1 - \delta_1 c_1) \\ &\quad -\rho \left[F\left(\frac{\delta_1 c_1}{\delta_2 \hat{v}(1) r_1}\right) - F\left(\frac{\delta_1 c_1}{\delta_2 v_H r_1}\right) \right] (\delta_2 v_H r_1 - \delta_1 c_1). \end{aligned} \tag{19}$$

Clearly, if $F(\beta_1)$ is such that the support of $(\delta_1 c_1/\delta_2 r_1)/\beta_1$ is concentrated in the interval $(v_L, \hat{v}(1))$ there is only a gain from malleability, whereas if it is concentrated in $(\hat{v}(1), v_H)$ there is only a loss.[29] When the two scenarios have equal probability, the net welfare effect is negative, since investment is more valuable when the true valuation is high. Such is the case, for instance, if $1/\beta_1$ is uniformly distributed on any subinterval of $[1, +\infty)$.[30]

In summary, the net impact of malleability on the (ex ante) efficiency of date-1 decisions depends on whether the distribution of willpower makes underinvestment by the low type or the high type more of a problem. The first case occurs when the self-control problem and the difficulty of the typical task are relatively moderate, the latter when they are really severe.

Date-0 behavior. Suppose that with $\lambda = 1$, the only equilibrium is non-investment: $\delta_2 v_H r_0 < c_0^H$. With $\lambda < 1$, investing yields not only the return r_0 but also raises the individual's sense of

[28] 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 $\overline{v}$. Of course, the nature of the equilibrium, including the value of $\hat{v}(1)$, is endogenous and depends on the distribution $F(\beta_1)$. The proof of Proposition 4 takes this fixed-point aspect into account.

[29] In the first case, if the support lies in $(v_L, \overline{v})$, the mere coarsening of information is beneficial (as in Carrillo and Mariotti (2000)), so $\lambda < 1$ leads to a gain even when the individual has no identity-relevant choice to make at $t = 0$ (i.e., he is constrained to $a_0 = 0$ or to $a_1 = 1$). Morenovel is the case in which the support is concentrated on or extends to $(\overline{v}, \hat{v}(1))$: it is then *only when combined with the ability to act* and take advantage of the malleability of beliefs that imperfect (self) knowledge can be beneficial.

[30] The probabilities of a $\delta_2 v_L r_1 - \delta_1 c_1$ gain and a $\delta_2 v_H r_1 - \delta_1 c_1$ loss in (19) are then respectively proportional to $(1-\rho)x_L[\hat{v}(1) - v_L]$ and $\rho[v_H - \hat{v}(1)]$, and thus exactly equal.

identity from $\hat{v}(0)$ to $\hat{v}(1)$. Consistent with the above discussion, let this additional benefit be sufficient to induce an equilibrium with mixing by the v_L type: $\mathbf{V}(v_L, \hat{v}(1), A_0+r_0)-\mathbf{V}(v_L, v_L, A_0) = c_0^L$. This occurs when

$$F\left(\frac{\delta_1 c_1}{\delta_2 v_L r_1}\right) - F\left(\frac{\delta_1 c_1}{\delta_2 \overline{v} r_1}\right) < \frac{c_0^L - \delta_2 v_L r_0}{1-\lambda} < F\left(\frac{\delta_1 c_1}{\delta_2 v_L r_1}\right) - F\left(\frac{\delta_1 c_1}{\delta_2 v_H r_1}\right). \tag{20}$$

Intertemporal welfare under malleability then differs from that achieved when $\lambda = 1$ by

$$\Delta W = \rho\left(\delta_2 v_H r_0 - \beta_0 c_0^H\right) + (1-\rho) x_L\left(\delta_2 v_L r_0 - \beta_0 c_0^L\right) + (1-\lambda)\,\Delta V, \tag{21}$$

where ΔV is given by (19) and β_0 is date-0 willpower. When β_0 is low enough that (say) both of the first two terms are positive even though $\delta_2 v_H r_0 < c_0^H$, ex-ante efficient investments fail to occur in period 0 under $\lambda = 1$. Because the ability to manage one's identity ($\lambda < 1$) provides additional motivation for such investments, it tends to raise welfare by improving decision-making at $t = 0$. When β_0 is equal or close to 1, on the other hand, increased period-0 investment due to $\lambda < 1$ is a net cost, which only pays off in terms of improved decision-making at $t = 1$ if ΔV is positive and sufficiently large.

Proposition 4 *In the self control case, a greater malleability of beliefs (a lower λ) can raise welfare by improving choices at $t = 0$ and / or at $t = 1$.*

E Taboos

While economists tend (at least, in their professional "identities") to view all activities as fungible or "secular," that is, subject to trade-offs, most societies, religions and cultures hold, or at least declare, certain goods to be "priceless" or "sacred": life, justice, liberty, honor, love, friendship,, one's children, democratic citizenship, adherence to religious commandments, etc. (see, e.g., Durkheim (1925), Fiske and Tetlock (1997)).

It is thus considered highly immoral to attribute a monetary equivalent to the value of marriage, friendship or loyalty to a cause. Sexuality, death, body organs and military duty are not to be "commodified", nor are childbearing permits an acceptable policy for population control. Admittedly such rules are often observed in the breach and the boundaries between the secular and the sacred are evolving ones, as demonstrated by the 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. They also testify to widespread views that the mere existence of certain markets would be "contrary to human dignity" and harmful even to people who do not transact in them, because they would allow or "invite" comparisons and that, to use Fiske and Tetlock (1997)'s memorable phrase, "to compare is to destroy". Yet what exactly is

being destroyed by placing a monetary value on certain goods or activities, and how this damage occurs, is never really explained.

Taboos and sacred values are closely related to the preservation of identity, in the sense of upholding 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. Accordingly, they can be accounted for, and normatively evaluated, as a special case of our general framework.

Consider the same model as before, with $v \in \{v_H, v_L\}$ representing (a signal of) the long-term value to an individual of an important activity or state of being: freedom, bodily integrity, non-addiction, relationship to a person (child, spouse, friend) or to a more abstract entity (country, religion), with associated capital A_0. For the usual anticipatory-utility (including prospects for an afterlife) or self-control motives, the individual may want to be optimistic about v, resulting in a value function $V(v, \hat{v}, A_1)$ of the type studied earlier.

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

$$p = \begin{cases} p_H & \text{with probability} \quad z \\ p_L & \text{with probability} \quad 1-z \end{cases} . \tag{22}$$

The actual value may be learned, depending on the context, by checking what is being offered on a formal or informal market (for switching loyalties, selling one's vote, organ or children; for prostitution, fraud, crime, etc.) or by simply engaging in deliberate, "coldhearted" calculations about the costs and benefits of different courses of action.

To simplify the problem, let p_H be high enough and p_L low enough such that, if the 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 $p = p_L$.[31] In either case, he will later recall that he entertained the possibility of a transaction and evaluated whether maintaining his identity, dignity, etc., was "worth it" or not. Thus, at $t = 1$, the agent will remember whether or not he had looked into the price and, with probability $1 - \lambda$, he will draw from it the appropriate inferences about where his "true values" lie.

Investing in identity ($a_0 = 1$) consists here in upholding a rule never to not place a price on certain goods –staying away from markets where such transactions occur, not entertaining offers one may receive, and avoiding even "forbidden thoughts" of commensurability. The cost of upholding the taboo is the option value $c_0 = zp_H$ of the potential transaction thus foregone,[32]

[31] Formally, this is a dominant strategy for both types $i = H, L$, 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".

[32] Transacting without first finding out the price is either infeasible, or else unprofitable (due to the average

and an individual of type $i = H, L$ will do so if

$$\mathbf{V}(v,\hat{v}(1), A_0) - \mathbf{V}(v,\hat{v}(0), A_0 - z) \geq z p_H, \tag{23}$$

with the same notation as usual.[33] This is clearly a special case of our general model, with $r_0 = z$ and initial stock $A_0' \equiv A_0 - z$; therefore, all the previous results apply directly. 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 and how taboo-breaking by others can lead to reaffirmation or collapse.[34] On the normative side, Propositions 3 and 4 show how the welfare effect (at the individual level) of taboos depends critically on whether they reflect "mental consumption" or self-control motives. In the first case, taboos generally lower ex ante welfare (unless agents are sufficiently non-Bayesian or the consumption value of beliefs is appropriately nonlinear). In the latter, they can increase it, but only under specific conditions involving priors and the severity of the self-control problem.

II Multiple identities

We now extend the analysis to multiple identities. For simplicity, let there be two, defined by specific assets A_t and B_t with potentially uncertain valuations v_A and v_B in $\{v_L, v_H\}$. There are three main types of interactions among identities, which we shall analyze in turn:

a) Consumption rivalry. At some future date ($t = 2$), the agent may have to choose between identities due to time, geographical, legal or other exclusivity constraints. Important examples include national, regional or religious identities. In the extreme case of full rivalry, date-2 utility is $\max\{v_A A_2, v_B B_2\}$. More generally, the individual faces a tradeoff between reaping the benefits of A_2 and B_2, resulting in a payoff of the form $\max_{t\in[0,1]} \{t^\alpha v_A A_2 + (1-t)^\alpha v_B B_2\}, 0 < \alpha \leq 1$.

b) Resource rivalry. At date 0 (and possibly date 1), the individual has limited time or resources to devote to competing identities. The leading example here is that of professional versus family identity, particularly for working women. Formally, the investments a_0 and b_0 are subject to

"auction" price $z p_H + (1-z) p_L$ being too low).

[33] In writing the second term in (23) we take advantage of the linearity of $\mathbf{V}$ in A_1 under both the anticipatory-utility and the self-control 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.

[34] See Section III.A for more details on peer effects. Because they involve the avoidance of normally valuable information, taboos are related to the strategic-ignorance argument of Carrillo and Mariotti (2000) and Carrillo (2005) and especially the rule-based behavior in Bénabou and Tirole (2004). There are, however, two important differences. First, on the demand side, imperfect willpower is here only one of the potential sources from which motivated beliefs may arise. Second, 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.

congestion or a budget constraint, $a_0 + b_0 \leq 1$; given our (simplifying) assumption of discrete choices, the latter means that at most one of the two can be non-zero.

c) Trait affiliation. Identities that compete for neither consumption nor investment resources will still interact if their values v_A and v_B are perceived to be correlated, as with certain clusters of attributes thought to define personality types ("a good Christian", a "doer", a moral person, etc.).

A Consumption rivalry

When two identities are likely to compete in the future, investing in one (say, B) depreciates the other (A), as it suggests that the individual may not value it that highly. If the individual has substantial capital vested in A but the ultimate value of this identity is more uncertain than that of B, he main then refrain from even highly desirable investments in B, and end up worse off as a result.

We demonstrate here this mechanism of *dysfunctional identity* using the anticipatory or pure self-image case, 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 new or "modern" one –for instance, in the context of farmers and workers faced with sectoral shifts brought about by globalization and technical change, or that of immigrants and their children confronting the issue of assimilation into a Western country.

(a) Modern identity. At $t = 0$, the agent decides whether to invest in identity B ($b_0 = 1$), at a cost c_B: acquiring a new type of human capital, mastering a new language and culture, socializing with an unfamiliar group, etc. The investment succeeds with probability $z \in (0, 1)$, in which case the initial stock B_0 rises to $B_1 = B_0 + b_0 r_B$; it fails with probability $1 - z$ ($B_1 = B_0$). This uncertainty captures the idea that the agent may end up consuming A even after investing in B, for instance because such investment is a new activity to which he may or may not be suited; A thus serves as a "fallback" option. The (per unit) value of B capital, on the other hand, is a known v_B. For instance, the material benefits of 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 by 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 individual 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 ($\hat{v}_A = \hat{v}(a_0)$). 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 A or B, thus achieving $\max\{v_A A_2, v_B B_2\}$.

To focus on the interesting case, suppose that, *ex post,* the agent will end up consuming B only if he had successfully invested in it,

$$v_B B_0 < v_L A_0 < v_H A_0 < v_B (B_0 + r_B), \tag{24}$$

but that, *ex ante,* the expected return is sufficiently high that, when beliefs are not malleable ($\lambda = 1$), even agents with a high value for A will optimally invest:

$$z(\delta_1 s_1 + \delta_2)[v_B (B_0 + r_B) - v_H A_0] > c_B. \tag{25}$$

When identity concerns are operative, however, both types will fail to make this efficient investment, as long as

$$z(\delta_1 s_1 + \delta_2)[v_B (B_0 + r_B) - v_L A_0] - (1-z)\delta_1 s_1 (1-\lambda)(\bar{v} - v_L) A_0 < c_B. \tag{26}$$

The first term is the "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, however, such savoring- or affect-motivated identity management always ends up lowering welfare, as in the single-identity case. Indeed, while the nonlinear value function makes the model 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 5 *Assume the AU specification, with (24)-(25). The individual invests less in a known identity (B) when it will compete in the future with another one (A) of uncertain value. This is more likely to happen the higher $A_0, 1-\lambda$ and s_1, and it is always (weakly) welfare reducing.*

It is interesting to relate these results to recent trends and controversies.

1) Resistance to structural change. Identity preservation can account for inertia (or even hostility) in the adjustment to a new world. International 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 devaluing past investments in the old (rural, extended-family, blue-collar, etc.) identity.

2) Resistance to assimilation. In many Western countries, immigrants and their descendents experience strong tensions between integration and the preservation of their specific culture. This is particularly acute among the youth, who are locally born and have citizenship yet, for many, do not "feel" British, German or French. But neither do they feel Pakistani, Turkish or Algerian, having seldom been to the "old country" or learned its language. As seen earlier, it is in situations of uncertainty over one's own values that identity threats and investments become most relevant.[35] Relatedly, laws and proposals requiring "demonstrations" of assimilation, such as the Home Secretary's (2001) urging that minorities adopt British "norms of acceptability" and newcomers be required to take an oath of allegiance, study British history and culture and embrace "our laws, our values, our institutions," elicit significant anger. The concerned groups feel that complying would represent a betrayal or denial of their own identity, culture or religion.[36] In the same vein, Fordham and Ogbu (1986) suggest that some academic failure among African-Americans may partly reflect a desire to maintain racial identity.[37]

3) Destructive identity, discrimination and communitarianism. "Not investing in B" in order to safeguard A can also mean actually destroying productive B capital. This simply corresponds in the model to the case where $c_B < 0$, meaning that the costly action is now one ($b_0 = 0$) that reduces B or prevents it from growing. In the events that shook the suburbs of French cities in 2005, the young rioters attacked and destroyed a number of schools and nursery schools, pharmacies and many cars, mostly in their in *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 (26) while leaving (24) and (25) unchanged). The first is a lower perceived chance of success in those investments (z) or associated payoff (r_B). Thus, if minority youth become more pessimistic about their chances of mobility through education, or perceive, rightly or wrongly, that even with diplomas the jobs to which most of them will be able to aspire will be low paying ones (whether due to discrimination of economic trends), they will switch to the destructive-

[35] One can also relate to the results in Proposition 2 on the effects of A_0 and ρ 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. (Those with tertiary education, on the other hand, were best able to combine their original and new identities; higher education is typically more internationalized and "portable", and higher income probably relaxes some of the constraints underlying consumption rivalry).

[36] See Hoge (2002). Similar controversies characterized the French ban on the veil in public schools, although in this case there was a sharp divide in the North-African community –particularly along gender lines. Note also that, here again, self-perceived intentions matter: infiltrated members of an extremist organization presumably feel much less conflict in submitting to such requirements, pledges, dress codes, etc., because they clearly know that their doing it signals commitment to, rather than possible abandonment of, their chosen "values".

[37] Austen-Smith and Fryer (2005) develop a formal model for this idea, with which this section shares some common elements. In their model, a youth's schooling behavior sends a signal to multiple audiences: the (mainly White) labor market, and the minority in-group, which wants reliable members who can be depended upon in difficult times. Our account emphasizes instead the valuation of competing identities, and the target audience is primarily oneself.

identity scenario, even when z 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 important factor is the salience s_1 of the "alternative" A identity and the benefits anticipated from it. This is where ideological manipulation and religious indoctrination may come into play (as with advertising in the single-identity case), as well as the amplification mechanism of media coverage.

Finally, 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, it is clear that similar insights apply when the demand for identity-consistent beliefs stems from a self-motivation 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, or may even destroy B capital. Such a strategy serves not as a physical commitment (investment costs and returns at $t = 1$ 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.[38]

B Resource rivalry

Suppose now that the two identities are independent and without consumption rivalry. The continuation value function is then simply

$$V\left(v_A, \hat{v}_A, A_1\right) + V\left(v_B, \hat{v}_B, B_1\right), \tag{27}$$

with each component arising from any of the mechanisms seen earlier to generate a demand for motivated beliefs and satisfying Assumption 3. As to date 0, resource rivalry mean that investing in A raises the cost (or shadow cost) of investing in B, and vice-versa, so one would expect it to result in less total investment than when there is no such congestion. This is always the case when $\lambda = 1$ and it is also easy to obtain when $\lambda < 1$. The more surprising result is that when identity concerns are operative, resource competition can actually *increase* total investment.

To show this most simply, let us assume that the two identities are symmetric in all respects, including the initial stocks ($A_0 = B_0$) and that resource rivalry is so severe that at most one investment can be undertaken: $b_0 = 0$ if $a_0 = 1$ and $a_0 = 0$ if $b_0 = 1$. As before, c_0^i will denote the corresponding cost (equal for both identities) for type $i = H, L$. Consider now the condition under which a full-investment equilibrium exists, that is, under which even type (v_L, v_L) invests in one of the two identities. Absent resource rivalry, the corresponding condition for A is

$$\mathbf{V}\left(v_L, \overline{v}, A_0 + r_0\right) - \mathbf{V}\left(v_L, v_L, A_0\right) \geq c_0^L, \tag{28}$$

[38] Indeed, one may observe that the earlier analysis of taboos is also a case of consumption rivalry, with the agent choosing between a "sacred" and a "secular" identity, by evaluating or not evaluating payoffs.

and by symmetry the same one will ensure that full investment also prevails for identity B.[39] Under resource rivalry, let $\hat{v}^+$ and $\hat{v}^-$ denote the posterior expected valuations for the identities in which there has been and not been investment, respectively, with $\hat{v}^+ > \overline{v} > v_L > \hat{v}^-$. When no investment has occurred, the perception of the agent's type is (v_L, v_L).[40] The condition for full investment in either A or B is therefore:

$$\mathbf{V}\left(v_L, \hat{v}^+, A_0 + r_0\right) - \mathbf{V}\left(v_L, v_L, A_0\right) + \mathbf{V}\left(v_L, \hat{v}^-, B_0\right) - \mathbf{V}\left(v_L, v_L, B_0\right) \geq c_0^L. \tag{29}$$

Comparing (28) with (29) shows that full investment is more likely in the latter case (e.g., holds for a wider range of c_0^L), through the conjunction of two related effects. First, investment in identity A, say, is a stronger signal of v_A being high ($\hat{v}^+ > \bar{v}$), as the alternative would have been investment in B, which under (28) is preferred to no investment at all. Second, not investing in A when one does not invest in B generates some collateral damage on B since it reveals the valuation for B to be v_L for sure, whereas with investment in A it might sill have been v_H with some probability ($\hat{v}^- < v_L$). A social signaling analogy may help grasp these two competition effects, which could be labelled *winner's distinction* and *loser's comfort*: when receiving two invitations for dinner for the same evening and accepting one, one knows that the selected host will be particularly pleased and that the one who is turned down will find some comfort in the fact that one was facing a tough choice.

Proposition 6 *Let the two identities be symmetric.*
(1) Under resource rivalry, investing in A, say, is a better signal about v_A and not investing in B carries less stigma about v_B than when the two identities do not compete for resources.
(2) As a result of these winner's distinction and loser's comfort effects:
(i) there may be more investment under resource rivalry than with independent identities;
(ii) there is always more investment than if there were a single identity with marginal distribution equal to that of $\max\{v_i\}$.

C Value affiliation and value conflicts

We now assume away consumption and resource rivalry. The continuation value is thus the same as in (27) and the date-0 utility simply reflects the sum of the two (effective) costs:

$$U = -c_0(v_A)a_0 - c_0(v_B)b_0. \tag{30}$$

We retain the assumption of symmetry for simplicity and assume that the marginal distributions remain unchanged: $\Pr\left(v_A = v_H\right) = \rho = \Pr\left(v_B = v_H\right)$. The only interaction between the two

[39] Suppose for example that types (v_H,v_H) and (v_L,v_L) choose to invest in A or B at random (the analysis is more general). Then, following $a_1 = 1$, $\widehat{\rho}_A = \rho(1-\rho/2)$ and $\widehat{\rho}_B = (1-\rho)^2/2$.

[40] We impose here the same type of restriction on out-of-equilibrium beliefs as in the one-dimensional case.

identities comes from the affiliation of values: $\Pr(v_B = v_H | v_A = v_H) = \Pr(v_B = v_L | v_A = v_L) \equiv \sigma$,with $\sigma > 1/2$ for positive correlation and $\sigma < 1/2$ for a negative one.

Suppose, for simplicity, that there is a single investment opportunity at date $0 : a_0 \in \{0, 1\}$ and $b_0 \equiv 0$. It can then be shown that there is more investment in identity A than would take place if values were independent. For example, the full-investment equilibrium requires that

$$\mathbf{V}(v_L, \overline{v}, A_0 + r_0) + \mathbf{V}(v_B, \overline{v}, B_0) - c_0^L \geq \mathbf{V}(v_L, v_L, A_0) + \mathbf{V}(v_B, E(v_B | v_A = v_L), B_0).$$

The more positively correlated v_A and v_B are, the less stringent this condition becomes, due to the increased collateral damage of a non-investment in A on identity B. Similarly, for negative correlation, a greater conflict between the two values ($E(v_B | v_A = v_H)$ decreases, $E(v_B | v_A = v_L)$ increases) makes the condition harder to satisfy.

Proposition 7 *Let the two identities be symmetric, but with an investment opportunity in A only, and let V satisfy $V_{12} = 0$ and $V_{22} \leq 0$ (e.g., V_{22} in the AU specification). Keeping marginal distributions constant, an increase in affiliation raises investments in identity. Conversely, value conflicts discourage identity investments.*

Affiliation may be related to the emergence of clusters of identity components. In her study of working men's values, Lamont (2002) identifies two such clusters. Some men stress the *"disciplined self",* emphasizing responsibility, work ethic, willpower, etc., and often associated with conservative political ideas. Others, by contrast, stress the *"caring self,"* built around the values of generosity, solidarity, and the appreciation of what one has (family, friends, community) rather than the constant pursuit of wealth and status. Lamont also observes that Black workers tend to stress relatively more their caring self over their disciplined self than their White counterparts, while the latter see the (essentially White) middle and upper classes above them as lacking in the caring-self dimension. Indeed we saw how, under consumption rivalry, a group that experiences discrimination in "self-reliance" activities such as education and employment (B), and conversely a greater need for solidarity (A), will develop an identity (perceptions of v_A and v_B) that "fits" its situation and can persist even when discrimination abates, creating a form of history dependence.

A second illustration of clustering is the standard cross-cultural comparison of how people define themselves (e.g., Kunda (2002), Nisbett (2003)). Westerners tend to emphasize their "independent self', namely unique attributes, abilities, thoughts and feelings. Eastern cultures, on the other hand, stress the "interdependent self", i.e., the relationship of the individual to parents, family, coworkers or firm.

Another application of this section's results concerns *military identity.* As Akerlof and Kranton (2005) note, an important step in building a military identity is the eradication of the soldier's "old self" (through haircut, uniform, hazing, drilling in of hierarchical values, etc.).

Indeed, there is typically a tension (negative correlation) between certain values essential in war and others that most civilians hold. As Proposition 7 shows, this will make conscripts or new recruits reluctant to invest in military identity. Moreover, the anticipation that most of them have of eventually returning to civilian life leads to "consumption rivalry," which by Proposition 5 creates a further incentive not to identify too much with military values. The strategies employed by the military to overcome this resistance involve important cognitive aspects that aim at (partially) "erasing" memories of one's civilian self. This corresponds in the model to lowering the recall or accessibility rate λ and depreciating the stock A_0 –both of which, by the analysis of the last two sections, encourage investments in the military identity.

III Identity in social interactions

A Peer effects

People's social environment shapes their identity through two main channels. First, interactions with family, peers or coworkers directly affects the payoffs (cost c_t, productivity r_t or value v) of investing in different identity-relevant assets. Second, they affect the two cognitive mechanism involved in identity formation, by exposing the individual to signals and cues relevant to "who he is" (supply side) and by altering his incentives to perceive himself, and be perceived by others, as a certain type of person (demand side). Our focus here is on the cognitive channel.

(a) Signaling to the in-group. A person's behavior conveys information to others with whom he interacts, or whose opinion he cares about. For example, people are more likely to invest in a relationship with someone whom they believe shares common long-term interests with them.[41] When the individual's future payoffs (whether material or affective) are thus affected by others' perceptions $\hat{v}'$ of his type, as well as by his own sense of identity $\hat{v}$, the continuation value is of the form $V(v, \hat{v}, A_1, \hat{v}')$. Since others form their beliefs primarily by observing behavior, $\hat{v}' = \hat{v}(a_0^i)$ and the expected value function playing the role of (15) is now

$$\mathbf{V}(v, \hat{v}, A_1) \equiv \lambda V(v, v, A_1, \hat{v}) + (1 - \lambda)\, V(v, \hat{v}, A_1, \hat{v}). \tag{31}$$

Thus, as long as $V(v, \hat{v}, A_1, \hat{v})$, as a function of $(v, \hat{v}, A_1)$, satisfies Assumption 3, the whole analysis carries over.[42] Adding a social signaling concern is then akin to increasing the intensity of the self-signaling motive (from V_2 to $V_2 + V_4$) in the original model. This leads, as one would

[41] In Rotemberg (1994), such complementarities lead agents to "invest in altruism" (even unilaterally), thus altering their own preferences rather than their beliefs and external image, as is the case here.

[42] Such is for instance the case if, in addition to Assumption 3: a) the individual cares about others' opinion per se (i.e., not for instrumental purposes): $V = \tilde{V}\,(v, \widehat{v},\, A_1) + \mu\ \hat{v}'$, with $\mu > 0$; or b) in the self-control case, others are more likely to invest at date 1 in an action that is complementary to a_1 if they think it more likely that the individual himself will choose $a_1 = 1$.

expect, to increased identity investment (as can be shown from a minor adaptation of the proof of Proposition 1).

An interesting implication is that a person's social environment can induce him to make investments that will have lasting effects, through the capital accumulated or the self-image achieved, even after he has left that environment. As to the welfare impact of such social interactions, it depends again on whether identity is functional (e.g., a demanding but cooperative work environment can have a long-lasting impact on an individual otherwise prone to procrastination) or dysfunctional (in the anticipatory utility model augmented by attention to others' opinions, the individual is trapped in a "double treadmill").

(b) Responding to transgressions. We now examine how observing others' behavior, rather than being observed by them, influences one's identity. We consider here a very simple, two-agent version of the basic model, with sequential moves.[43] At date 0, j moves first, choosing a_0^j; then, after observing j's action, i makes his own choice a_0^i. The two agents need not be symmetrical. The actions a_0^i and a_0^j are just investments in specific assets and do not directly enter into the other agent's payoff (so, at this stage, $a_0^i = 1$ is not to be interpreted as aggressiveness, ostracism, etc.).

The only link between the two individuals is that they are "similar", that is, their values (v^i, v^j) are affiliated. Let ρ_0 be the prior on i's type and ρ^+ and $\hat{\rho}^-$ the corresponding posteriors after observing agent j invest ($a_0^j = 1$) or not invest ($a_0^j = 0$). Under a monotonic strategy for j, and with positive affiliation between v^i and v^j,

$$\rho^- < \rho_0 < \rho^+, \tag{32}$$

since j's investing (say) is "good news" about her v^j and therefore also about v^i. Furthermore, if one fixes x_L^j and x_H^j, which is legitimate if j does not observe i's action, or else perfectly remembers his own motivation, then as the correlation of types increases, ρ^- decreases and ρ^+ increases (weakly). To derive how agent i's behavior is influenced by j's, we can then directly apply Propositions 1 and 2, with the initial belief ρ set to ρ^- or ρ^+. In particular, Proposition 2(2) yields the comparative statics of average investment with respect to the degree of correlation, which acts like a mean-preserving spread in ρ.[44]

These results then help understand the nature of identity threats coming from other persons or groups and how people deal with transgressors. "Deviant" behavior by peers (non investment) sends a negative signal about the value of the existing capital stock (anticipated utility version)

[43] The case of simultaneous moves is more complicated, as it involves mutual informational spillovers between agents' actions, plus coordination of their expectations. Battaglini, Bénabou and Tirole (2005) provide a detailed analysis of peer interactions among agents with a self-control problem related to that in Example 2.

[44] Given that $\rho x_H + (1-\rho)x_L$ is concave up to $\bar{\rho}$ then falls to zero, a mean-preserving spread reduces average investment when starting from a prior $\rho_0 < \bar{\rho}$; conversely, it increases it when starting from $\rho_0 > \bar{\rho}$, provided ρ^- falls below $\bar{\rho}$.

or that of motivation-sensitive future investments (imperfect willpower version). For example, members of an ethnic, religious or national community who mingle with "outsiders", or are not fully supportive of the group's positions, undermine others' own sense of commitment to the common value. Or, as discussed by Akerlof and Kranton (2000), a woman in a construction job or a man wearing a dress threaten masculine identity –more specifically, in our model, men's beliefs about abilities "only they" have, or attractions they "could never" have. When such transgressions represent sufficiently bad news to an initially strongly held identity ($\rho^- < \overline{\rho} < \rho_0$), they will elicit a strong positive investment response, designed to *"repair"* the damaged self-view; the different forms (constructive or destructive) that it may take are discussed below. When the initial identity was relatively weak, on the other hand ($\rho^- < \min\{\tilde{\rho}, \rho_0\} < \overline{\rho}$), transgressions will further *"sap morale"* and depress investment.[45]

Focussing on the first case, strong reactions to deviant behaviors can then be partly explained by cognitive strategies.[46] First, the exclusion of mavericks from the in-group suppresses the undesirable reminders created by their presence: "out of sight, out of mind". That is, exclusion lowers λ.[47] Second, actively excluding or even harassing deviators can be a form of identity damage control by self-signaling: one must incur the cost of losing beneficial interactions with the excluded and expend resources or take risks to support norm enforcement (including punishing others those who themselves fail to enforce it). If those most truly committed to the group identity (the v_H types) face effectively lower costs in such activities, the sorting condition will hold here as well and these exclusionary or harassing behaviors can themselves serve as identity investments –an alternative, less benign interpretation of a_0^i.

The bad news conveyed by the divergence of behaviors (or opinions) is more identity-threatening, the more similar the norm violator is to oneself, that is, the more correlated their values are *a priori.* The harshest moral condemnations and punishments are therefore reserved for people with similar attributes and endowments who, by their words or their acts, threaten the group's valued identity. The canonical example (so to speak) is apostasy. The Catholic Church long imposed excommunication without hope of pardon on apostates, tortured and executed heretics, and Islamic Sharia law still prescribes that apostates should be put to death, lose their children and their property.

Because "renegades" are more of a threat to one's identity than members of differentiated outgroups, sharp conflicts are less likely to arise if people entertain different reference groups. If group i believes (accurately or not) group j to be sufficiently different along some attributes that

[45] Symmetric results apply following the observation of a peer's identity-affirming behavior, which raises the prior to ρ^+.

[46] In addition to "instinctive" anger and contempt, hardwired or learned early on in life, which serve broad functional purposes as well. Such emotions also provide a suitably noisy "rationalization" for having excluded previous members other than just censoring their messages.

[47] More precisely, it lowers the recall or visibility rate λ_L of identity-threatening signals ($a_0^j = 0$), while leaving unaffected that for identity-affirming ones (λ_H, for $a_0^j = 1$).

either correlate with v, or change what can be inferred about it from investment or non investment, it will feel less threatened when group j behaves in a dissonant way. It will also feel less "validated" when group j behaves consonantly, however, but in the case of a sufficiently strongly held initial beliefs the fear of identity losses dominates the desire for identity gains, as shown earlier. Similarly, in such circumstances the perception of a certain amount of heterogeneity in the in-group can also reduce benchmarking.

Paradoxically, a peer's deviancy can sometimes strengthen one's identity: the sight of a "bad Christian", corrupt executive or other norm violator can make other members of the group feel better (e.g., morally superior) rather than worse. Such a form of *Schadenfreude* arises when correlation affects costs more than values. If the effort costs of "behaving well" are imperfectly accessible in retrospect and are positively correlated across individuals, a peer's deviant behavior suggests a high investment cost and will therefore make one's own investment a better signal of having "the right values".

B Dignity and scapegoating in bargaining and group conflict

"If you cut the pay of all but the superperformers, you have a big morale problem. Everyone thinks they are a superperformer."
(Head of human resources of a nonunion manufacturing company with 200 employees)

"A pay cut also represents a lack of recognition. This is true of anybody. People never understand and don't want to understand. They don't want to believe that the company is in that much trouble. They live in their own world and make very subjective judgments.
(Owner of a small business with 30 employees)

Interviews from Truman Bewley, *Why Wages Don't Fall During a Recession* (1999).

We consider here another set of economic and political applications of the model: how identity concerns such as pride, dignity or wishful thinking about one's options ("keeping hope") lead individuals or groups to walk away from "reasonable" offers, try to shift blame for failure onto others, or take refuge in political utopias –resulting in costly delays, impasses and conflicts. Examples include trials, divorces, strikes, the scapegoating of minorities and certain wars. The importance of belief distortion in those phenomena is attested by field observers (e.g., Bewley (1999) in the context of labor relations, Woods et al. (2006) in that of war), as well as by recent experiments. In particular, Thompson and Loewenstein (1992) and Babcock et al. (1995) demonstrate how subjects in bargaining situations devoid of any informational asymmetry (common knowledge) spontaneously generate, through self-serving processing and recall of the evidence, divergent beliefs about the fairness of their cause and wishful predictions of outcomes that, in turn, result in costly delays and failures to agree.

To capture these phenomena we consider a "partnership" between two individuals or groups –husband and wife, labor and management, majority and minority populations, etc. Each

may be of high or low type, $i = H, L$, corresponding to ability, motivation, honesty, outside opportunities, etc. At date 0, the joint output or productivity of the partnership is revealed: it is either good or bad, $y \in \{y_B, y_G\}$, with $y_G > y_B$. The technology exhibits complementarity, in that $y = y_G$ if and only if $i = j = H$. The interesting case will then be when $y = y_L$, since this means that at least one of the parties is "to blame" for the low output –disappointing marriage, firm or economy, lost war, etc.

At the end of period 0, the two partners must decide whether to: (i) remain together, in which case they will continue to produce the same (expected) output in period 2 (the long run) and must bargain over how it will be shared; or (ii) split, in which case each type i will get a reservation value v_i, with $v_H > v_L$, corresponding for instance to producing in autarky (or, as in a dynamic version of the model, searching for a new match). In a more conflictual setting, these outside options can also capture the outcome of a fight between the two sides involving expropriation and net resource dissipation. In all that follows, we abstract from discounting ($\delta_1 = \delta_2 = 1$).

Let parameters be such that staying together is efficient for all teams, both balanced (HH or LL) and unbalanced (HL), but that in the latter case a compensating transfer (or share of y_B exceeding 1/2) is needed to induce the more productive partner to stay:

$$y_G > 2v_H > \ y_B \ > v_H + v_L > 2v_L. \tag{33}$$

When bargaining and making their stay or quit decisions at the end of period 0, the two parties will be assumed to know not only the joint output y, but also each one's type. Such *symmetric-information* will make inefficient-breakdown results all the more interesting, and allow us to provide the first formal model of the Babcock et al. (1995) findings described above. In keeping with the rest of our self-inference based theory, we further assume that, at date 1 :

(i) Whereas the level of joint output y is "hard" data that is easy to remember and verify, individuals's separate contributions to it –their types v– are soft, unverifiable information, which later on is only imperfectly recalled. Indeed, it would always be more pleasant, *ceteris paribus,* to "recall" that one was the competent and honest partner and the other was entirely to blame for the team's poor performance. For notational simplicity we shall take here the recall probability of the v's to be $\lambda = 0$, but this is inessential.

(ii) Individuals experience anticipatory feelings from their long-run (date-2) consumption prospects, with savoring coefficient s_1, common to both for simplicity. Alternatively, they may derive utility from their self-view about their talent or usefulness to society; this slight variant leads to similar results.

We formalize the bargaining process over future output as a standard Nash demand game.[48]

[48] We treat the period-0 allocation of output as sunk (e.g., allocated *ex ante* on a 50-50 basis, before types are

At $t = 0$, players 1 and 2 simultaneously make demands for shares θ_1 and θ_2 of y_B; if $\theta_1 + \theta_2 \leq 1$ each gets what they asked for, whereas if $\theta_1 + \theta_2 > 1$ the negotiation breaks down and the pair dissolves. Demanding a larger share (or offering a smaller one) may correspond to requesting either a monetary transfer, the allocation of a control right (e.g., regional autonomy, child custody), the attribution of a prestigious position or other ego-gratifiying benefit, or a reform of the performance-measurement system that will alter the sensitivity of income shares to individual contributions. We assume that offers are later remembered (having been formally recorded, submitted to an arbitrator, etc.), although the key results are similar when they are not.

We look for a symmetric, pure-strategy equilibrium, with shares $\theta_H^* > 1/2 > \theta_L^*$ for the high and low valuation types respectively in an unbalanced partnership, and a common share $1/2$ in a balanced one. Finally, we restrict out-of equilibrium beliefs as follows. Let $\Theta \equiv \{\theta_L^*, 1/2, \theta_H^*\}$ the set of equilibrium individual offers. For $\theta_i \in \Theta$ and $\theta_j \notin \Theta$, player i is presumed to have played on the equilibrium path, which is sufficient to tie down both players' types. If both θ_i and θ_j belong to Θ but are jointly inconsistent with equilibrium, on the other hand, then: (i) if $\theta_i = \theta_j$ both payers are considered equally likely to have deviated, resulting in the same posterior $\hat{v}_i = \hat{v}_j = \bar{v}$; (i) if $\theta_i > \theta_j$, then $\hat{v}_i = v_H$ and $\hat{v}_j = v_L$; this is in the spirit of standard equilibrium refinements, since it always the strong type who has less to lose from breaking up the match.

Let us first observe that in any equilibrium with agreement it must be that the shares demanded by both sides sum to 1; otherwise, either party can ask for ε percent more and gain $(1 + s_1)\varepsilon y$, since the team will still stay together. For the same reason, downward deviations by either type (asking for less than the equilibrium share) are never profitable. he binding constraints will thus correspond to upward deviations.

Given $(1 + s_1)y_G/2 > (1 + s_1)v_H > (v_H + s_1\hat{v})$ for any feasible value of $\hat{v}$, matched strong partners (HH) always stay together, sharing output equally. The interesting case is that of low-productivity pairs, $y = y_B$. Consider first bargaining in an unbalanced (HL) team. For the H type to be satisfied with his share, it must be that:

$$\theta_H^* y_B \geq v_H. \tag{34}$$

Otherwise he could ask for more, which would break up the team while maintaining his posterior belief $\hat{v} = v_H$ and achieving $(1 + s_1)v_H > (1 + s_1)\theta_H^* y_B$. Next, for the weak partner (L type) to accept the bargain, it must be that:

$$(1 + s_1)\,\theta_L^* y_B \geq v_L + s_1\bar{v}, \tag{35}$$

revealed). Since (expected) output is equal in both periods, however, allowing period 0 resources to be part of the bargaining would simply amount to a renormalization (doubling) of the size of the pie.

otherwise he could deviate and demand θ_H^* (mimicking the strong partner), thus achieving (and savoring at $t = 1$) the posterior identity $\hat{v} = \bar{v}$, even though his true outside option is only v_L. Other deviations to $\theta' > \theta_L$ with $\theta' \neq \theta_H$ would still identify him as the weak type, $\hat{v} = v_L$, and be *a fortiori* unprofitable under (35).

The set of agreeable sharing rules $(1 - \theta_L^*, \theta_L^*)$ is thus defined by

$$\frac{v_L + s_1 \bar{v}}{1 + s} \leq \theta_L^* y_B \leq y_B - v_H \tag{36}$$

It clearly shrinks as identity concerns increase, up to the critical threshold

$$s^* \equiv \frac{v_L + v_H - y_B}{y_B - v_H - \bar{v}}, \tag{37}$$

beyond which a bargaining impasse arises, in spite of gains from trade and symmetric information. Intuitively, a higher s makes the loss of self-image involved in "admitting blame" more costly for the L type, who then requires a higher share θ_L^* to be compensated. At some point this becomes more than the H type is willing to grant given his outside option (including identity concerns) and no agreement can be reached. The two parties then split (or fight), with each side getting $v_i + s_1 \bar{v}$; thus, once again, there is actually no net gain in self-esteem or anticipatory utility, only a destruction of surplus.

We turn finally to bargaining in an LL team. By asking for any share $\theta' > 1/2$, either side can break up the match and achieve self image v_H (by either of our refinement assumptions). Therefore, the partnership remains sustainable only if $(1 + s_1)\, y_B/2 \geq v_L + s_1 \bar{v}$, or $s_1 \leq s^{**}$, where

$$s^{**} \equiv \frac{y_B - 2v_L}{2v_H - y_B} > s^*. \tag{38}$$

Otherwise the match is dissolved, as each side seeks to convince himself that he is better than the other, even though in reality both are equally bad.

Proposition 8 *(1) For $s_1 \leq s^*$, both balanced (LL) and unbalanced (HL) low-output partnerships successfully negotiate, splitting resources equally in the first case and according to any sharing rule θ_L^* satisfying (36) in the second; this agreement range shrinks with s_1.*
(2) For $s^* < s_1 \leq s^{**}$, the two sides can still agree if they share equal blame but not if one must shoulder it all: LL matches survive but HL ones are destroyed. For $s_1 > s^{**}$, not even balanced (LL) partnerships can find a sustainable agreement.
(3) All dissolutions are inefficient.

In summary, our model of bargaining with malleable beliefs identifies a new and potentially important limit to the achievement of Coasian deals, namely the preservation of dignity, pride, or "hope" about the future.

IV Conclusion

We developed in this paper a simple but flexible framework for analyzing a broad class of beliefs –such as identity, dignity, or religion– which people value and invest in, with important economic implications. The model also offers a unified account of many seemingly disparate phenomena documented by psychologists; yet others, such as endowment effects, could be fairly easily obtained. Rather than restate the paper's main themes and results, we will single out here the two that are most novel and, having been treated here only in their simplest form, suggest interesting avenues for further research. The first one is that of sacred values and taboos, which demonstrates how the debate over the interplay of markets and morals can be brought into the realm of formal analysis. The second concerns the role, in bargaining and other distributive conflicts, of endogenously arising self-serving beliefs linked to pride, dignity or wishful thinking. Many applications, ranging from the design of contracts and organizations to political economy, can be envisioned.

Appendix

Proof of Proposition 1. We first list the potential equilibrium configurations, given the monotonicity property:

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

$$\mathbf{V}(v_H, \overline{v}, A_0) \geq \mathbf{V}(v_H, v_H, A_0 + r_0) - c_0^H. \tag{A.1}$$

(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

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

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

$$\mathbf{V}(v_H, v_L, A_0) \leq \mathbf{V}(v_H, v_H, A_0 + r_0) - c_0^H, \tag{A.2}$$

$$\mathbf{V}(v_L, v_L, A_0) \geq \mathbf{V}(v_L, v_H, A_0 + r_0) - c_0^L. \tag{A.3}$$

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

$$\mathbf{V}(v_L, v_L, A_0) = \mathbf{V}(v_L, \hat{v}(1), A_0 + r_0) - c_0^L. \tag{A.4}$$

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

$$\mathbf{V}(v_L, v_L, A_0) \leq \mathbf{V}(v_L, \overline{v}, A_0 + r_0) - c_0^L. \tag{A.5}$$

We can first rule out equilibria of type (b), in which type v_H randomizes: since $\mathbf{V}_2 > 0$, the no-investment equilibrium also exists if an equilibrium of type (b) exists. Furthermore, since $V(v, \bar{v}, A_0) > V(v, \hat{v}(0), A_0)$ 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.1) holds and, when this condition fails, separation, randomization by v_L or full investment, depending respectively on whether (A.2)-(A.3), (A.4) or (A.5) 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

$$\mathbf{V}(v_H, \bar{\rho}v_H + (1-\bar{\rho})v_L, A_0) \equiv \mathbf{V}(v_H, v_H, A_0 + r_0) - c_0^H \tag{A.6}$$

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.3) 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 v_L type must invest with positive probability. If (A.5) holds there can be no solution to (A.4) with $x_L < 1$, so the only equilibrium is full investment on $[0, \bar{\rho})$. If (A.5) is reversed, on the other hand, it involves mixing: by (14),

$$\hat{v}(1) = \frac{\rho}{\rho + (1-\rho)x_L} v_H + \frac{(1-\rho)x_L}{\rho + (1-\rho)x_L} v_L, \tag{A.7}$$

and by (A.4) 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 v_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,

$$\mathbf{V}\left(v_L, v_L, A_0\right) = \mathbf{V}\left(v_L, \tilde{\rho}v_H + (1-\tilde{\rho})v_L\right), A_0 + r_0) - c_0^L. \tag{A.8}$$

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 v's incentive to invest, $\mathbf{V}(v, \hat{v}(1), A_0 + r_0) - \mathbf{V}(v, \hat{v}(0), A_0)$, increases: indeed, by (15), its derivative with respect to $1 - \lambda$ is

$$V(v, \hat{v}(1), A_0 + r_0) - V(v, v, A_0 + r_0) + V(v, v, A_0) - V(v, \hat{v}(0), A_0) \geq \int_{\hat{v}(0)}^{\hat{v}(1)} V_2(v, x, A_0) dx > 0,$$

where the first inequality follows from the assumption $V_{23} \geq 0$. Consequently, the no-investment region shrinks, $\bar{\rho}$ increases, $\tilde{\rho}$ rises and $\hat{v}(1)$ decreases in the mixing equilibrium: investment increases (weakly) for each type, at any value of ρ.

(ii) It is easily verified from (A.6), (A.7) and (A.8) 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,

$$\mathbf{V}(v, \hat{v}(1), A_0 + r_0) - \mathbf{V}(v, \hat{v}(0), A_0) = \delta_1 s_1 \left[\lambda v r_0 + (1-\lambda)[\hat{v}(1)(A_0 + r_0) - \hat{v}(0)A_0] + \delta_2 v r_0\right.$$

rises with $s_1$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 form 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 Propositions 3 and 4 The first proposition was shown in the text, following (17). The proof of the second one is by construction. Let us choose $\beta^* \in (0,1)$ such that

$$\beta^* \delta_2 r_1 \bar{v} < \delta_1 c_1 < \beta^* \delta_2 r_1 v_H. \tag{A.9}$$

Next, define $v^* \in (\bar{v}, v_H)$as $v^* \equiv (1/\beta^*)(\delta_1 c_1/\delta_2 r_1)$ 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 $F(\beta)$ puts mass 1 on β^*; by continuity, the arguments below will continue to hold when the mass is close enough to 1. By (8), the agent invests at $t=1$ when $\hat{v} \geq v^*$. As to (A.10), it means that if the v_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:

$$c_0^H < \delta_2 r_0 v_H + (1-\lambda)(\delta_2 r_1 v_H - \delta_1 c_1) = \mathbf{V}(v_H, v_H, A_0 + r_0) - \mathbf{V}(v_H, \bar{v}, A_0), \tag{A.11}$$

$$c_0^L \equiv \delta_2 r_0 v_L + (1-\lambda)(\delta_2 r_1 v_L - \delta_1 c_1) = \mathbf{V}(v_L, v^*, A_0 + r_0) - \mathbf{V}(v_L, v_L, A_0). \tag{A.12}$$

Compared to the equilibrium that prevails when $\lambda = 1$, in which $\hat{v} = v$ always, this yields a gain in V given by (19) but with the 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 (21) positive), if suffices that

$$\beta_0 c_0^L < \delta_2 v_L r_0 < \delta_2 v_H r_0 < c_0^H. \tag{A.13}$$

Compatibility with (A.11)-(A.12) requires that

$$\begin{aligned} (1-\lambda)(\delta_2 r_1 v_H - \delta_1 c_1) &> c_0^H - \delta_2 r_0 v_H > 0, \\ (1-\lambda)(\delta_2 r_1 v_L - \delta_1 c_1) &< (1/\beta_0 - 1)\,\delta_2 v_L r_0, \end{aligned}$$

neither of which contradicts any other condition. ■

Proof of Proposition 5 We shall take advantage of the fact that the value function is piecewise linear to rewrite the agent's optimization problem (over some relevant range) as a linear one that can be mapped isomorphically into the single-identity anticipatory utility case.

Given (24) and (25), the intertemporal utility of an agent of type $v_A \in \{v_H, v_L\}$ who starts with stocks (A_0, B_0) and chooses $b_0 \in \{0,1\}$ is:

$$\tilde{W}(v_A, A_0, B_0, b_0) \equiv b_0 z (\delta_2 + \delta_1 s_1) v_B (B_0 + r_B)$$

$$+(1-b_0)\left[\delta_2 v_A+\delta_1 s_1\left(\lambda v+(1-\lambda)\hat{v}_A(1-b_0)\right)\right]A_0$$
$$+b_0(1-z)[\delta_2 v_A+\delta_1 s_1(\lambda v+(1-\lambda)\hat{v}_A(b_0))]A_0-b_0 c_B. \tag{A.14}$$

Let us now define the variables $a_0 \equiv 1-b_0$, $R_0 \equiv zA_0$ and the functions:

$$\begin{aligned} U(v_a,a_0,A_0;B_0) &\equiv (1-a_0)\left[z\left(\delta_2+\delta_1 s_1\right)v_B\left(B_0+r_B\right)-c_B\right], \\ V(v,\hat{v},A_1) &\equiv \left(\delta_2 v_A+\delta_1 s_1\hat{v}_A\right)A_1, \\ \mathbf{V}(v,\hat{v},A_1) &\equiv \lambda V(v,v,A_1)+(1-\lambda)V(v,\hat{v},A_1), \end{aligned}$$

It is then easy to see that (A.14) can be rewritten as

$$W(v_A,A_0,B_0,a_0)=U(v_a,a_0,A_0)+\mathbf{V}(v_A,\hat{v}_A(a_0),A_0(1-z)+a_0R_0) \tag{A.15}$$

and that the function U satisfies $U_3=0$, hence Assumption 1, while V is exactly the same as in (4) and therefore satisfies Assumption 3. Thus, while $U(v_a,a_0,A_0;B_0)$ and $\mathbf{V}(v,\hat{v},A_1)$ no longer individually correspond to the date-zero flow payoffs and date-1 expected value function (note that U 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 the state variable A_t : instead of $A_1=A_0+a_0r_0$, we now have $A_1=A_0(1-z)+a_0R_0$; the "depreciation" term in $1-z$ will not change anything (qualitatively), while the fact that the return $R_0=zA_0$ now increases with the initial stock will only reinforce the fact that investment increases with A_0. Thus, the agent will invest at $t=0$ if and only if

$$\mathbf{V}(v_A,\hat{v}(1),A_0(1-z)+R_0)-\mathbf{V}(v_A,\hat{v}(0),A_0(1-z))\geq c_0, \tag{A.16}$$

where $c_0 \equiv c_B-z\left(\delta_2+\delta_1 s_1\right)v_B\left(B_0+r_B\right)$ is now the same for both types. All results in Proposition 1 and all those in Proposition 2 pertaining to the anticipatory utility case thus remain unchanged. In particular, equilibrium generally results in excessive "investment" in A, which mean suboptimally low investments in B. ■

Proof of Proposition 6 In the text it is proved that the full investment equilibrium (investment by type (v_L,v_L)) exists for a wider range of parameters than the range defined by (A.5). More generally, with two identities and complete resource rivalry, the "low type" corresponds to type (v_L,v_L) and the "high type" to $\max\{v_i\}=v_H$, i.e., to the composite of types (v_H,v_H), (v_H,v_L) and (v_L,v_H). The probability of type (v_L,v_L) must be $1-\rho$ for comparison purposes, so the probability of type v_L for any given identity must be $1-\chi\equiv\sqrt{1-\rho}$.

If $\mathbf{V}(v_H,v_L,A_0)\geq\mathbf{V}(v_H,v_H,A_0+r_0)-c_0^H$, we know that there is no investment with a single identity, so there cannot be less with two. Let us therefore assume that $\mathbf{V}(\mathbf{v}_H,v_L,A_0)<\mathbf{V}(v_H,v_H,A_0+r_0)-c_0^H$ and look for the separation, mixing and no-investment regions.

(a) No investment. As in the single-identity case, we assume that a deviation toward investment is interpreted as coming from the highest type (v_H, v_H). The condition for no-investment,

$$2\mathbf{V}(v_H, \bar{v}, A_0) \geq \mathbf{V}(v_H, v_H, A_0 + r_0) + \mathbf{V}(v_H, v_H, A_0) - c_0^H, \tag{A.17}$$

is weaker than that for a single identity, which by (A.1) is $\mathbf{V}(v_H, \bar{v}, A_0) \geq \mathbf{V}(v_H, v_H, A_0+r_0) - c_0^H$. So if there is no investment with two identities, a fortiori there is none with a single one.

(b) Mixing. With a single identity, the range for mixing is given by (A.4): $\mathbf{V}(v_L, v_L, A_0) = \mathbf{V}(v_L, \hat{v}(1),\ A_0 + r_0) - c_0^L$.With two, it becomes:

$$2\mathbf{V}(v_L, v_L, A_0) = \mathbf{V}\left(v_L, \hat{v}^+(1),\ A_0 + r_0\right) + \mathbf{V}\left(v_L, \hat{v}^-(1),\ A_0\right) - c_0^L. \tag{A.18}$$

Because $\hat{v}^-(1) > v_L$ (loser's comfort), when (A.18) holds, (A.4) implies that for a given x_L

$$\hat{v}^+(1) = \frac{\chi}{\chi + (1-\chi)x_L/2} < \frac{\rho}{\rho + (1-\rho)x_L} = \hat{v}(1),$$

therefore the equilibrium x_L is higher than with a single identity: there is more investment.

(c) Separation. With a single identity, the conditions are given by (A.2)-(A.3). With two, let $v_2 \in \{v_L, v_H\}$ denote $\min\{v_i\}$ when $v_1 \equiv \max\{v_i\} = v_H$ (high type). The new condition for the high type to invest is:

$$\mathbf{V}(v_H, v_L, A_0) + \mathbf{V}(v_2, v_L,\ A_0) \leqslant [\mathbf{V}(v_H, v_H, A_0 + r_0) - c_0^H] + \mathbf{V}(v_2, \overline{v}_2, A_0) \tag{A.19}$$

where $\overline{v}_2 < \overline{v}$ is the mean of v_2 conditional on being a high type. This condition is satisfied over a wider range of parameters than with a single identity.

As to the condition for the low type not to invest, it becomes:

$$2\mathbf{V}(v_L, v_L, A_0) \geqslant [\mathbf{V}(v_H, v_H,\ A_0 + r_0) - c_0^L] + \mathbf{V}(v_L, \overline{v}_2, A_0), \tag{A.20}$$

and is harder to satisfy than with a single identity. ■

Proof of Proposition 7 Note that since $V_{12} = 0$ and the agent can invest only in A, his decision will depend only on his value of v_A, even though it will take into account the inferences to be drawn concerning v_B. Let $\phi_B(\widehat{v}_A) \equiv E[v_B \mid \hat{v}_A]$, which is an increasing (decreasing) function of $\widehat{v}_A$ under positive (negative) affiliation. As before, we restrict attention to monotonic equilibria. To facilitate the comparison with the proof of Proposition 1, we add in brackets the new terms stemming from the existence of a second identity.

(a) No investment. Type v_H is unwilling to separate if

$$V(v_H, \overline{v}, A_0) + [V(v_B, \overline{v}, B_0)] \geq V(v_H, v_H, A_0 + r_0) - c_0^H + [V(v_B, \phi_B(v_H), B_0)].$$

Note that the difference between the terms in brackets is independent of v_B, thanks to $V_{12} = 0$. For positive correlation $\phi_B(v_H) > \phi_B(\overline{v}) = \overline{v}$, so this condition becomes harder to satisfy than in the single-identity case; the higher the degree of correlation, the more so, since $\phi_B(v_H)$ rises while $\overline{v}$ remains constant.

(b) Randomization by v_H. Such an equilibrium is still Pareto-dominated. Indeed, if

$$V(v_H, \widehat{v}(0), A_0) + [V(v_B, \phi_B(\widehat{v}(0)), B_0)] = V(v_H, v_H, A_0 + r_0) - c_0^H + [V(v_B, \phi_B(v_H), B_0)],$$

with $\widehat{v}(0) < \overline{v}$, then

$$V(v_H, \overline{v}, A_0) + [V(v_B, \phi_B(\overline{v}), B_0)] > V(v_H, v_H, A_0 + r_0) - c_0^H + [V(v_B, \phi_B(v_H), B_0)].$$

Therefore, the no-investment equilibrium also exists and type v_H, as well as type v_L, is better off in that equilibrium. Together with the analysis of case (a), this implies that the region in which $x_H = 1$ expands as correlation increases.

(c) Separation. Investment by the v_H but not the v_L type requires

$$\begin{aligned} V(v_H, v_L, A_0) + [V(v_B, \phi_B(v_L), B_0)] &\leq V(v_H, v_H, A_0 + r_0) - c_0^H + [V(v_B, \phi_B(v_H), B_0)], \\ V(v_L, v_L, A_0) + [V(v_B, \phi_B(v_L), B_0)] &\geq V(v_L, v_H, A_0 + r_0) - c_0^L + [V(v_B, \phi_B(v_H), B_0)]. \end{aligned}$$

As correlation grows, $\phi_B(v_L)$ decreases and $\phi_B(v_H)$ increases, so the first becomes easier to satisfy and the latter more difficult.

(d) Mixing by v_L. The indifference condition is now :

$$V(v_L, \widehat{v}(1), A_0 + r_0) - V(v_L, v_L, A_0) + [V(v_B, \phi_B(\widehat{v}(1)), B_0) - V(v_B, \phi_B(v_L), B_0)] = c_0^L.$$

By definition of ϕ_B, the derivative of the term in brackets with respect to σ is $v_H - v_L$ times

$$V_2(v_B, \phi_B), B_0)(2\widehat{\rho}(1) - 1) + V_2(v_B, \phi_B(v_L), B_0) \geq 2\widehat{\rho}(1) V_2(v_B, \phi_B(\widehat{v}(1)), B_0) > 0,$$

since $V_{22} \leq 0$. Thus, as σ rises , the term in brackets increases for any given value of $\hat{v}(1)$. To maintain indifference, $\hat{v}(1)$ must therefore decrease, meaning that x_L rises.

(e) Full investment. The v_L type wants to pool if

$$V(v_L, v_L, A_0) + [V(v_B, \phi_B(v_L), B_0)] \leq V(v_L, \overline{v}, A_0 + r_0) - c_0^L + [V(v_B, \overline{v}, B_0)]. \qquad (A.21)$$

As $\phi_B(v_L)$ decreases with correlation, this condition becomes easier to satisfy. The analysis of cases (c), (d) and (e), taken together, implies that x_L increases (weakly) with correlation, while x_H remains equal to 1. ■

REFERENCES

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

Babcock, L., Loewenstein, G., Issacharoff, S. and Camerer, C. (1995) "Biased Judgments of Fairness in Bargaining**,"** *American Economic Review,* 85(1), 1337-1343.

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.

Becker, E. (1973) *The Denial of Death*, New York: Free Press.

Becker, G. and K. Murphy (2000) *Social Economics: Market Behavior in a Social Environment.* Bellknap 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. 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), forthcoming.

Bentham, J. (1789) *Introduction to the Principles and Morals of Legislation (1970).* London: Athlone Press.

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.

Bewley, T. (1999) *Why Wages Don't Fall During a Recession.* Harvard University Press.

Bisin, A. and T. Verdier (2000) ""Beyond The Melting Pot": Cultural Transmission, Marriage, And The Evolution Of Ethnic And Religious Traits," *The Quarterly Journal of Economics,* 115(3), 955-988.

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.

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.M., and A.E. 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.

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.

Davies, (2004) "Identity and Commitment"' Tinbergen Institute Discussion Paper 055/2, University of Amsterdam.

Dessi, R. (2005) "Collective Memory, Social Capital and Integration," Université de Toulouse mimeo.

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

Erikson, E. (1968) *Identity: Youth and Crisis.* New York: Norton.

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

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

Fordham, S., and J. Ogbu (1986) "Black Students' School Success: Coping with the Burden of Acting White," *The Urban Review*, 18: 176—206.

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

Freyer, 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.

Goffman, E. (1963) *Stigma: Notes on the Management of Spoiled Identity*, Englewood Cliffs: Prentice Hall.

Gul, F. (1991) "A Theory of Disappointment Aversion" *Econometrica,* 59(3), 667-686.

Hill, C. (2006) "What The New Economics Of Identity Has To Say To Legal Scholarship," University of Minnesota Legal Studies Research Paper No. 05-46.

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

Horst, U., Kirman, A. and M. Teschl (2006) "Changing Identity: The Emergence of Social Groups," University of British Columbia mimeo, May.

Kopczuk, W., and J. Slemrod (2005) "Denial of Death and Economic Behavior," *Advances in Theoretical Economics*, 5(1). Article 5.

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

Kiesler, C., Nisbett, R. and M. Zanna (1969) "On Inferring Ones' Beliefs from One's Behavior," *Journal of Personality and Social Psychology,* 11(4), 321-327.

Köszegi, B. (2004) "Utility from Anticipation and Personal Equilibrium," U.C. Berkeley mimeo, June.

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

—— (2002) *Social Cognition: Making Sense of People.* Cambridge, Mass.: MIT Press.

Habermas, J. (1973) *Legitimation Crisis*, Boston: Beacon 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.

Lamont, M. (2004) *The Dignity of Working Men: Morality and the Boundaries of Race, Class and Immigration.* Harvard University Press.

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

LeBoeuf, R. and Shafir, E. (2004) "Alternative Selves and Conflicting Choices: Identity Salience and Preference Consistency," Princeton University mimeo.

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. (2203) "Sexual Harassment Under Social Identity Threats: The Computer Harassment Paradigm," *Journal of Personality and*

Social Psychology, 85(5), 853-870.

Morash, M. (1980) "Working Class Membership and the Adolescent Identity Crisis," *Adolescence*, 15: 313–320.

Nisbett, R. (2003) *The Geography Of Thought: How Asians and Westerners Think Differently... and Why.* New York: The Free Press.

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

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.

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

Shayo, M. (2005) "Nation, Class and Redistribution: Applying Social Identity Research to Political Economy," Princeton University mimeo, August.

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, J. (2005) "Reputation, Social Identity and Social Conflict," Princeton University mimeo, November.

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

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

Woods, K., Lacey, J. and W. Murray (2006) "Saddam's Delusions: The View from the Inside," *Foreign Affairs,* June.

Wichardt, P. (2005) "Why and How Identity Should Influence Utility," University of Bonn mimeo, November.

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