Over My Dead Body: Bargaining and the Price of Dignity

By ROLAND BÉNABOU AND JEAN TIROLE *

> "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 manufacturing company, in Bewley (1999)).
>
> "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".
>
> (Small business owner, in Bewley (1999)).

Concerns of pride, dignity and the desire to "keep hope" about future options often lead individuals and groups to walk away from reasonable offers, try to shift blame for failure onto others or take refuge in political utopias, leading to impasses and conflicts. Examples include trials, divorces, strikes, the scapegoating of minorities for economic hardships, and war. A key and puzzling aspect of these processes is the role played by wishful *rationalizations and delusions,* as attested by field observers (e.g., Truman F. Bewley (1999) in the context of labor relations, Kevin Woods, James Lacey and Williamson Murray (2006) in that of war) as well as controlled experiments. Leigh Thompson and George Loewenstein (1992) and Linda C. Babcock et al. (1995) thus demonstrate how subjects in bargaining situations with common knowledge spontaneously generate, through self-serving processing and recall of the same evidence, divergent beliefs about the fairness of their cause and wishful predictions of outcomes, and how these are associated to costly delays and disagreements.

To analyze these behaviors, we propose here a simple model of how anticipatory or self-esteem concerns lead to the inefficient breakdown of Coasian bargaining under *symmetric information,* as both sides seek to self-enhance by turning down "insultingly low" offers. To do so, we build on Roland Bénabou and Jean Tirole (2007), which develops a general framework for analyzing social and economic phenomena involving beliefs which people "invest in".

The underlying idea is that individuals are often uncertain or insecure about their own "deep values", abilities or worth; and that, having better, more objective access to the track record of their *actions* than to the exact mix of *motivations* that spurred them, they are rationally led to judge themselves by what they do.[1] When contemplating choices, they then factor in what kind of a person each alternative would "make them" and the desirability of those self-views. The theory is thus cognitive, as it explicitly models identity and related concepts as beliefs and emphasizes the *self-inference* process through which they operate. At the same time, the *value* of identity or dignity arises because they confer affective benefits, functional ones, or both. The first case arises when self-esteem has pure consumption value or when future prospects give rise to anticipatory feelings such as savoring or dread. The second obtains when a strong sense of self provides clear priorities and directions that help the individual mobilize energy and resist short-term temptations.

Building on these two core assumptions –self-inference and motivated beliefs– we extend here the framework to bargaining and other distributive conflicts. We consider a partnership of two individuals or

* Bénabou: Department of Economics and Woodrow Wilson School, Princeton University, Princeton, NJ 08544, CIFAR, CEPR, IZA and NBER. Tirole: Toulouse School of Economics, 21 Allées de Brienne, 31000 Toulouse, France, and MIT. We thank Muhamet Yildiz. for helpful comments. Bénabou gratefully acknowledges support from the Canadian Institute for Advanced Research.

[1] See, e.g., Leon Festinger and James M. Carlsmith (1959) on cognitive dissonance, Darryl J. Bem (1972) on self-perception and George A. Quattrone and Amos Tversky (1984) on the self-manipulation of "diagnostic" actions. For recent experiments on the strategic management of self-image through costly actions or information-avoidance, see Jason Dana, Jason X. Kuang and Roberto A. Weber(2007) and Nina Mazar, On Amir and Dan Ariely (2006).

groups (parties in a dispute, capital and labor, majority and minority populations) who must decide whether to continue together or destroy the match. Continuation always yields a positive surplus, but a low output realization means that at least one party has low ability. Moreover, whereas joint output is hard data, individual contributions to it ("who is to blame", "who is getting a raw deal") are soft signals, symmetrically observed when producing and bargaining but imperfectly recalled following a split. Agreeing to inferior or even equal contractual terms in a low-performance team then entails a loss in self image and / or anticipatory utility. Conversely, by refusing "insulting" proposals and destroying the match when they do not obtain enough of a concession, each side can try to preserve or salvage their dignity and shift the blame onto the other, taking refuge from bleak realities in feelings of self-righteousness and wishful hopes for "a better tomorrow". In equilibrium, the range of sustainable sharing rules is shown to shrink with the importance of self-image or anticipatory concerns. Beyond a point, a bargaining impasse becomes unavoidable, in spite of gains from trade and fully symmetric information.

The paper relates first to the literature on cognitive dissonance and motivated beliefs (e.g., George A. Akerlof and William T. Dickens (1982), Matthew Rabin (1994), Bénabou and Tirole (2002, 2006a), Markus Brunnermeier and Jonathan Parker (2005)), as well as the related issue of anticipatory feelings (e.g., Loewenstein (1987), Andrew Caplin and John V. Leahy (2001)). Most closely related, through the idea of self-signaling or self-reputation, are Ronit Bodner and Drazen Prelec (2003) and Bénabou and Tirole (2004, 2006b). On the experimental side, James Konow (2000) and Dana, Kuang, and Weber (2003) demonstrate that subjects making monetary allocations affecting their own payoffs engage in self-deception and information avoidance about the fairness or likelihood of other players' outcomes.

The second related body of work is that on identity (e.g., Akerlof and Rachel E. Kranton (2005), Robert J. Oxoby (2003)). In these models, agent's preferences or attitudes depend on their chosen group memberships. We instead explicitly model the management of beliefs and the cognitive mechanisms through which it occurs. This also leads to different results, such as the fact that being able to manage his own identity can often make a person worse off.

Finally, there is a recent literature on bargaining and contracting with heterogenous beliefs (e.g., Muhamet Yildiz (2004), Nageeb Ali (2006)). Its general motivation is also to understand the sources of delays and breakdowns, but its methods and focus are quite different. In particular, beliefs are exogenous and remain invariant to offers and counteroffers. On the other hand, these papers make explicit the dynamic aspect of bargaining, whereas we consider a much simpler Nash demand game.

I. Model

A. Technology

We consider a "partnership" between two risk-neutral individuals or groups –spouses, labor and management, majority and minority populations, etc. Each individual may be of high or low type, H (probability ρ) or L (probability $1-\rho$), corresponding to different levels of ability, motivation, honesty, deservedness, outside opportunities, etc. There are three periods, as illustrated in Figure 1, and we abstract from discounting. 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 both agents are of type 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.

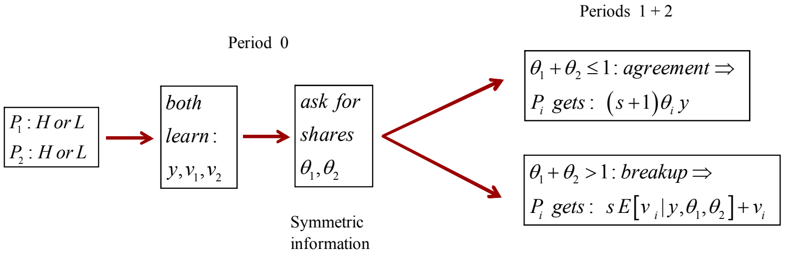
At the end of period 0, the two partners must decide whether to: (i) remain together, in which case they will continue to produce the same (expected) output in period 2 (the long run), and must bargain over how it will be shared; or (ii) split, in which case each agent i will get a reservation value determined by his type: $v^i = v_H$ for a high type and $v^i = v_L$ for a low type, with $v_H > v_L$. These outside options may correspond to producing in autarky, searching for a new match, or triggering a costly fight with the other side for control of resources.

Let parameters be such that staying together is efficient for all teams, both balanced (HH or LL) and unbalanced (HL), but 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:

$$(1) \qquad y_G > 2v_H > \; y_B \; > v_H + v_L > 2v_L.$$

When bargaining and making their stay or quit decisions at the end of period 0, the two parties are assumed to know (from recent observation) not only the

FIGURE 1: BARGAINING WITH MALLEABLE BELIEFS


Periods 1 + 2
Period 0
$P_1 : H \text{ or } L$
$P_2 : H \text{ or } L$
both learn: y, v_1, v_2
ask for shares θ_1, θ_2
Symmetric information
$\theta_1 + \theta_2 \leq 1$: agreement ⇒
P_i gets: $(s+1)\theta_i y$
$\theta_1 + \theta_2 > 1$: breakup ⇒
P_i gets: $s\,E[v_i | y, \theta_1, \theta_2] + v_i$

joint output y, but also each one's type. Such *common knowledge* will make inefficient-breakdown results all the more interesting and allow us to provide a formal model of the Babcock et al. (1995) types of findings described above.

B. *Preferences and beliefs*

In keeping with our general self-inference approach to identity, 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– represent soft, unverifiable information, which later on is only imperfectly recalled.[2] 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 ("everyone thinks they are a superperformer").

(ii) Individuals experience anticipatory feelings, such as hope and dread, from their long-run (date-2) income or consumption prospects. Alternatively, they may derive utility from pure self-esteem about their talent or worth.

We now formalize and discuss further each of these two premises.

For a person's past choices to define his sense of identity or dignity they must be *informative* about the "kind of person" he is, and therefore he must, at times, not be fully confident of his own type –deep values, abilities, etc. Similarly, if he later perfectly understood that what tipped the scales on a decision was the desire to achieve a certain self-image, such attempts would come to nil. Some form of imperfect self-knowledge (memory, accessibility) is therefore essential to understanding how people's choices can be shaped by concerns such as "being true to myself," "maintaining my integrity," "keeping my self-respect", etc. And indeed, there is extensive evidence that people's recall of their past feelings, efforts and motivations is highly imperfect and self-serving, that they judge themselves by their behavior, and consequently tailor the latter to preserve certain self-views.[3]

ASSUMPTION 1: *(Self-inference). At date* 1*, each player is aware (or reminded) of past individual contributions,* v^i, $i = 1, 2$, *only with probability* λ. *With probability* $1 - \lambda$, *he no longer recalls (has access to) these signals and uses instead the outcome of the negotiation to infer his and the other player's types.*

We denote by $\hat{\rho}^i$ individual i's date-1 belief about "what kind of a person" he is and by $\hat{v}^i \equiv \hat{\rho}^i v_H + (1 - \hat{\rho}^i) v_L$ the corresponding expected ability, either of which defines his (subjective) sense of identity. With probability λ, the posterior $\hat{v}^i$ is thus equal to the true value (or unbiased signal) v^i, and with probability $1 - \lambda$ it is equal to the conditional expectation $\hat{v}^i \in [v_L, v_H]$ that can be inferred from what offers were made and whether they were accepted or rejected. We assume that, in making these inferences at $t = 1$, players are fully rational Bayesians. Although this assumption can easily be relaxed, it is a natural benchmark and imposes discipline on the extent to which agents can chose to believe what suits

[2] Given the same information, subjects in bargaining situations systematically recall more of the evidence that favors their own side, even when roles are exogenously determined (Thompson and Loewenstein (1992)). In dictator games, they take advantage of contextual ambiguity to "persuade" themselves that they deserve more than what they judge to be the fair share when making allocations between other people (Konow (2000)).

[3] See again footnote 1. Further discussions and references can be found in Bodner and Prelec (2003) and Bénabou and Tirole (2004, 2007)

them.[4]

What suits them, in turn, depends on the affective needs and instrumental functions that identity or dignity serves for them. As discussed in Bénabou and Tirole (2002, 2007), the former include pure ego-gratification as well as remaining hopeful about one's future prospects (anticipatory utility); the latter include the motivational value of "believing in oneself" to achieve long-term goals and overcome self-control problems, as well as a possible facilitating role in signaling to others (if it is easier to persuade others of a claim, true or false, when one is convinced of it). We shall focus here on the first class of motives, namely "mental consumptions" (Thomas Schelling (1985)), but also explain in Section B how a simple variant yields a functional role for dignity, which strengthens the will to resist momentary temptations.

In what follows, we denote by E_t^i an agent i's expectations at date $t = 0, 1$.

ASSUMPTION 2: *(Motivated beliefs). Let U_2^i denote agent i's long-run income, equal to $\theta_i y$ when bargaining leads to and agreement in which i's share is θ_i and to v^i when it leads to a split. At* $t = 0$, *each agent seeks to maximize the (undiscounted) expected present value*

$$U_0^i \equiv E_0^i\left[s\,u_1^i + U_2^i\right], \tag{2}$$

where u_1^i is a utility flow received during period 1 and equal to either:
(i) $u_1^i = E_1^i[U_2^i]$, in the anticipatory-utility case
(ii) $u_1^i = E_1^i[v^i]$, in the pure self-esteem case.

As made clear by our notation, the two cases are closely related. Throughout the paper *we shall focus the exposition on (i),* which is somewhat more "consequentialist", but all the results are qualitatively identical with (ii).

C. Bargaining

We formalize the bargaining process as a standard Nash demand game. At $t = 0$, with full and symmetric information, players 1 and 2 simultaneously make demands for shares θ_1 and θ_2 of future output, y.[5] A larger share may correspond to a monetary transfer, a control right (regional autonomy, child custody, seats on the board) or a new performance measurement system that will alter the sensitivity of income shares to individual contributions. 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. We assume that offers are later remembered (having been formally recorded, submitted to an arbitrator, etc.), but the key results are similar when they are not.

We first look for a symmetric, pure-strategy Perfect Bayesian equilibrium, with agreement on shares $\theta_H^* > 1/2 > \theta_L^*$ for the high and low types respectively in an unbalanced partnership, and on a common share 1/2 in a balanced one. When no such equilibrium can be sustained we look for one (still in pure strategies) with partial efficiency, where of the two types of partnerships reaches agreement.

We restrict out-of equilibrium beliefs as follows. A pair with output y_G is unambiguously identified as HH, due to technological constraints. For pairs with output $y_{B,}$ let Θ denote the set of offers made in equilibrium.

(1) For $\theta_i \in \Theta$ and $\theta_j \notin \Theta$, player i is presumed to have played on the equilibrium path. If this identifies him as an H type, then his partner must be an L. Otherwise, we use the D1 criterion to restrict beliefs on his partner's type.

(2) If θ_i and θ_j are both in Θ but are jointly inconsistent with equilibrium, then: (i) if $\theta_i = \theta_j$ (e.g., both sides demand $\theta_H^* > 1/2$) the two players are considered equally likely to have deviated, and thus assigned the same image; (ii) 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 (such as D1), since it is always the strong type who has less to lose from breaking up the match.

[4] It also makes the model directly applicable to contexts where the two bargaining parties are signaling to an outside audience. Such social-reputational concerns, however, are "shut off" (through anonymity) in all the cited experimental evidence In many field surveys, they also seem secondary in importance to individuals' self-perceptions (see, e.g., the above quotations from Bewley (1999)). Thus, although self-reputation and social reputation are very complementary concerns, they correspond to empirically distinct phenomena and their analyses point to different mediating mechanisms –in particular, the key role of memory or retrospective accessibility in the pursuit of self-serving beliefs.

[5] We treat the allocation of period-0 output (if any) as sunk –e.g., shared *ex ante* on a 50-50 basis, before types are revealed. Since expected output is equal in both periods, allowing initial resources to be part of the bargaining would simply amount to doubling the size of the pie.

II. Results and Implications

A. Equilibrium

Let us first observe that in any equilibrium with agreement, the shares demanded by both sides must sum to 1. Otherwise, either party can ask for ε percent more and gain $(1+s)\varepsilon y$, 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. The binding constraints will thus correspond to upward deviations.

Since $(1+s)y_G/2 > (1+s)\upsilon_H$, 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:

$$(3) \qquad \theta^*_H y_B \geq \upsilon_H.$$

Otherwise he could ask for more, which would break up the team while maintaining his posterior belief $\hat{\upsilon} = \upsilon_H$ (since the other party is only asking for $\theta^*_L < 1/2$, which identifies him as an L type in a mixed pair) and achieving $(1+s)\upsilon_H > (1+s)\theta^*_H y_B$.

Next, for the weak partner (L type) to accept the bargain, it must be that:

$$(4) \qquad (1+s)\,\theta^*_L y_B \geq \upsilon_L(1+\lambda s) + s(1-\lambda)\bar{\upsilon},$$

where $\bar{\upsilon} \equiv (\upsilon_H + \upsilon_L)/2$. Otherwise, he could deviate and break the match by demanding θ^*_H (mimicking the strong partner), thus achieving with probability $1-\lambda$ the posterior self-view $\hat{\upsilon} = \bar{\upsilon}$, even though his true "worth" and outside option is only υ_L. Other deviations to $\theta' > \theta_L$ with $\theta' \neq \theta^*_H$ would still identify him as the weak type, $\hat{\upsilon} = \upsilon_L$, and be *a fortiori* unprofitable under (4).

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

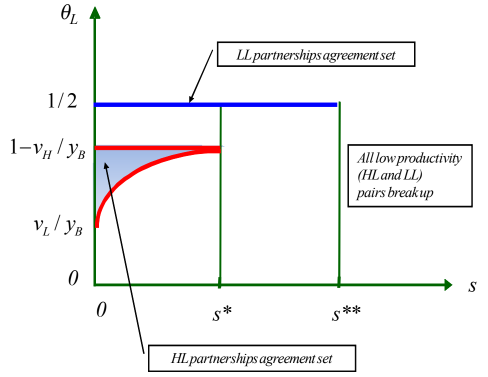
$$(5) \qquad \frac{\upsilon_L(1+s\lambda) + s(1-\lambda)\bar{\upsilon}}{1+s} \leq \theta^*_L y_B \leq y_B - \upsilon_H.$$

As illustrated in Figure 2, it shrinks as identity concerns increase, up to

$$(6) \qquad s^* \equiv \frac{y_B - \upsilon_H - \upsilon_L}{\upsilon_H + \lambda\upsilon_L + (1-\lambda)\bar{\upsilon} - y_B}$$

when the denominator is positive (otherwise, let $s^* \equiv +\infty$). Beyond this critical threshold 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 θ^*_L to be compensated. At some point this becomes more than the H type is willing to grant given his outside option, and no agreement can be reached. The two parties then split (or fight) by both demanding θ^*_H.

FIGURE 2: AGREEMENT AND BREAKDOWN REGIONS



We next turn to bargaining in an LL team. By asking for a share $\theta' > 1/2$, either side can break up the match and achieve, with probability $1-\lambda$, a self image υ_H. Therefore, the partnership remains sustainable only if $(1+s)\,y_B/2 \geq \upsilon_L + s\,[\lambda\upsilon_L + (1-\lambda)\upsilon_H]$ or $s \leq s^{**}$, where

$$(7) \qquad s^{**} \equiv \frac{y_B - 2\upsilon_L}{2\,[\lambda\upsilon_L + (1-\lambda)\upsilon_H] - y_B}$$

when the denominator is positive (if not, let $s^* \equiv +\infty$). Otherwise the match is dissolved, as each side seeks to convince himself that he is better than the other (demanding again θ^*_H), even though in reality both are equally bad.

In general, s^{**} can be above s^*, as illustrated in Figure 2, or below it. For brevity, we shall focus on the case $s^* < s^{**}$, which occurs (for all λ) if and only if $3y_B/2 < 2\upsilon_H + \upsilon_L$.[6] Together with (1), this means that $\upsilon_H + \upsilon_L < y_B < (2/3)\,(2\upsilon_H + \upsilon_L)$.

We obtain a further result by linking joint output to individual productivities. Consistent with our earlier assumptions, let HL and LL pairs both produce

[6] See the online appendix, which also provides a more detailed proof of Proposition 1 below.

$y_B = \Phi v_L$, where Φ is such that (1) holds.[7] It is then simple to verify that, as v_H/v_L rises, s^* and s^{**} both decrease and (5) becomes more stringent.

PROPOSITION 1: *(1) For $s \leq s^*$, unbalanced low-output (HL) partnerships successfully negotiate, splitting resources according to any sharing rule θ_L^* satisfying (5). This agreement range shrinks with s and, for $s > s^*$, the match is inefficiently destroyed.*
(2) For $s \leq s^{**}$, balanced low-output (LL) partnerships successfully negotiate, splitting resources equally. For $s > s^{**}$, the match is inefficiently destroyed.
(3) Let $y_B = \Phi v_L$. For any s, the bargaining set shrinks and both types of impasses become more likely, the greater the inequality v_H/v_L between high and low types' productivities.

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. It also leads to testable predictions, as both salience s and the productivity differential v_H/v_L can be manipulated experimentally. The latter can also be measured empirically in real-world contexts, where one should observe that more unequal bargaining positions reduce the likelihood of agreement.

From (6) and (7), we also have:

PROPOSITION 2: *Inefficient breakdowns of Coasian bargaining are more likely:*
(i) The more salient are agents' identity concerns (higher s).
(ii) The more malleable are their memories, and hence their beliefs (the lower λ).

B. Welfare

When HL pairs split both sides must be asking for the same $\theta_H^* > 1/2$, and when LL pairs also split the same must hold. Otherwise (by our first equilibrium refinement) one agent can deviate to θ_H^* and achieve self-reputation v_H. In any pair that splits, therefore, each side ends up with $v^i(1+s\lambda)+s(1-\lambda)\tilde{v}$, where

$$(8) \qquad \tilde{v} \equiv E\left[v \mid y_B, \theta_1 = \theta_2 = \theta_H^*\right]$$

is the average value of v over all such dissolutions, equal to $\bar{v}$ when only HL pairs dissolve and to $(\rho v_H + v_L)/(1+\rho)$ when LL pairs also split. There is thus, *in fine,* no net gain in self-esteem or anticipatory utility, only a transfer from the high to the low type within HL pairs, and from HL to LL pairs when the latter also break up. The pursuit of self-enhancement is a *zero-sum game that* leads only to a net *destruction of surplus,* equal (on average over all dissolving pairs) to $(1+s)(y_B - 2\tilde{v}) > 0$.

PROPOSITION 3: *An increase in the malleability of beliefs $1-\lambda$ always reduces ex-ante welfare. The same holds for an increase in the salience s of anticipatory-utility or identity concerns.*

In Bénabou and Tirole (2007) we show that, whereas the *positive* implications of individual belief management are very similar whether it arises from hedonic motives (self-esteem, anticipatory feelings) or instrumental ones (sense of direction, self-discipline), *normative* conclusions, by contrast, depend critically on this distinction. A similar principle applies in the present *strategic* context. Due to space constraints, we only sketch here this variant of the bargaining model that leads to a more attractive role (normatively speaking) for dignity concerns.

The only additional assumption is that, *at date* 1, each individual may need to carry out a task that:

(i) requires costly effort or perseverance, but is potentially subject to a self-control problem (e.g., due to hyperbolic discounting, $\beta < 1$);

(ii) has an expected return that increases with the agent's individual productivity v, so that perseverance and self-view $\hat{v}$ are complements.

The date-1 task may be independent of whether the agent is paired or unpaired at that time, or it could apply only to unpaired agents: searching for better opportunities, fighting, or holding out longer in costly bargaining.

In such settings, pooling by rejecting "realistic" offers boosts the v_L type's self-confidence and subsequent motivation, but weakens that of the v_H type. The first effect leads to a welfare gain, the second to a loss. Therefore when the nature of the date-1 self-control problem (value or probability distribution of β, returns to effort) makes it more of a concern for the low type than for the high one, meaning that its severity is moderate, there is a *net efficiency gain* from the malleability of beliefs ($\lambda < 1$) and the enhancement of the low types' dignity that it allows. When the self-control problem is harder, however, meaning that its affects the high types more often than the low ones, there is again a net social loss.

[7] In other words, the production technology is of the Leontieff type, $y = \Phi \min\{v^1, v^2\}$.

III. Conclusion

A simple model was proposed to analyze the role, in bargaining and other distributive conflicts, of *endogenously arising belief distortions* linked to pride, dignity or wishful thinking about future outcomes. A first set of further applications may include contracts and organizational design. A second interesting direction is the political economy of reforms such as opening to trade or liberalizing the labor market. Whereas the standard concern is whether winners can credibly commit to compensating losers, a potentially equally important one is that the latter precisely *do not want to see themselves* (and be identified by others) as losers, now dependent on "handouts" from the rest of the community.

REFERENCES

Ali, Nageeb. 2006. "Waiting to Settle: Multilateral Bargaining with Subjective Biases." *Journal of Economic Theory,* 130(1): 109–137.

Akerlof, George A. and William T. Dickens. 1982. "The Economic Consequences of Cognitive Dissonance." *American Economic Review,* 72(3): 307–319.

—— and Rachel E. Kranton. 2005. "Identity and the Economics of Organizations." *Journal of Economic Perspectives*, 19: 9–32.

Babcock, Linda C., Loewenstein, George, Issacharoff, Samuel and Colin Camerer. 1995. "Biased Judgments of Fairness in Bargaining**."** *American Economic Review,* 85(1): 1337–343.

Bem, Darryl. J. 1972. "Self-Perception Theory," in L. Berkowitz, ed., *Advances in Experimental Social Psychology*, vol. 6, 1–2. New York: Academic Press.

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

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

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

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

—– 2007. "Identity, Dignity and Taboos: Beliefs as Assets," CEPR Discussion Paper 6123, January.

Bewley, Truman F. 1999. *Why Wages Don't Fall During a Recession.* Harvard, MA: Harvard University Press: 177, 379.

Bodner, Ronit and Drazen Prelec. 2003. "Self-Signaling and Diagnostic Utility in Everyday Decision Making," in I. Brocas and J. Carrillo eds. *The Psychology of Economic Decisions*. Vol. 1: *Rationality and Well-Being.* Oxford University Press: 105–126.

Brunnermeier, Markus and Jonathan Parker. 2005. "Optimal Expectations." *American Economic Review*, 95: 1092–1118.

Caplin, Andrew and John V. Leahy. 2001. "Psychological Expected Utility Theory and Anticipatory Feelings." *Quarterly Journal of Economics*, 116: 55–80.

Dana, Jason, Kuang, Jason X., and Roberto A. Weber. 2007. "Exploiting Moral Wriggle Room: Behavior Inconsistent with a Preference for Fair Outcomes." *Economic Theory*, 33(1): 67–80.

Festinger, Leon and Carlsmith, James M. 1959. "Cognitive Consequences of Forced Compliance." *Journal of Abnormal and Social Psychology,* 58: 203–210.

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

Loewenstein, George. 1987. "Anticipation and the Valuation of Delayed Consumption." *Economic Journal*, 97: 666–84.

Mazar, Nina, Amir, On and Dan Ariely. 2006. "Mostly Honest: A Theory of Self-Concept Maintenance." MIT mimeo, December.

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

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

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

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

Thompson, Leigh and George Loewenstein. 1992. "Egocentric Interpretations of Fairness in Negotiation." *Organization Behavior and Human Decision Processes,* 51: 176–197.

Woods, Kevin, Lacey, James and Williamson Murray. 2006. "Saddam's Delusions: The View from the Inside." *Foreign Affairs,* 85(3): 2–26.

Yildiz, Muhamet. 2004. "Waiting to Persuade." *Quarterly Journal of Economics*, 119 (1): 223–249.

IV. Online Mathematical Appendix

Proof of the condition for $s^* < s^{**}$. From (6)-(7) we first easily check that, if $s^* = +\infty$, then $s^{**} = +\infty$. Next, when both $s^*$and s^{**} are finite, $s^* < s^{**}$ if and only if

$$\begin{aligned}&(y_B - 2v_L)\left[v_H + \lambda v_L + (1-\lambda)\bar{v} - y_B\right] > \\ &(y_B - v_H - v_L)\left[2(\lambda v_L + (1-\lambda)v_H) - y_B\right] \iff\end{aligned}$$

Denoting $\Delta \equiv v_H - v_L$, this becomes

$$\begin{aligned}(y_B - 2v_L)\left[2v_L + \Delta + (1-\lambda)\Delta/2 - y_B\right] &> \\ (y_B - 2v_L - \Delta)\left[2v_L + 2(1-\lambda)\Delta - y_B\right] &\iff \\ (y_B - 2v_L)\left[\Delta + (1-\lambda)\Delta/2 - 2(1-\lambda)\Delta\right] &> \\ -\Delta\left[2v_L + 2(1-\lambda)\Delta\right] &\iff \\ (y_B - 2v_L)\left[-3/2 - \lambda/2 + 2\lambda\right] + 2(1-\lambda)\Delta &> 0\end{aligned}$$

or, finally, $2v_H + v_L > (3/2)y_B$.

Proof of Proposition 1. The result for $s < s^*$ was shown in the text. The others follow from Lemmas 1 and 2 below.

LEMMA 1: *For $s > s^*$, HL pairs must split.*

(a) One cannot have both HL and LL agreeing since this requires $s \leq \min\left\{s^*, s^{**}\right\}$.

(b) One also cannot have HL agreeing and LL splitting. Otherwise, let $(\theta^*_H, \theta^*_L = 1 - \theta^*_H)$ be the shares agreed to in an HL pair and $(\theta', \theta^")$, with $\theta' + \theta" > 1$ the incompatible shares demanded in an LL pair. If neither of θ' nor $\theta"$ equals θ^*_H, by deviating to θ^*_H the L in an HL pair can achieve a gain of $s(1-\lambda)(v_H - v_L) > 0$. Therefore, it must be that $\theta^*_H \in \left\{\theta', \theta"\right\}$, say $\theta' = \theta^*_H$. But the other partner can then deviate from $\theta^"$to $1 - \theta' = \theta^*_L$, i.e. concede: he will remain identified as an L, but now achieve $(1+s)\theta^*_L y_B \geq (1+\lambda s)v_L + \lambda s\bar{v} > (1+s)v_L$, where the first inequality must hold in order for the L partner in an HL pair to agree. The deviation is thus profitable, so once again LL pairs cannot be sustained.

It follows from the Lemma that, for $s^* < s \leq s^{**}$, at most the LL matches can be sustained; and indeed, we showed in the text that in this region the shares $(1/2, 1/2)$ allow these pairs to reach agreement.

LEMMA 2: *For $s > s^{**}$, LL pairs must split.*

(a) Once again HL and LL cannot both agree, as this requires $s \leq \min\left\{s^*, s^{**}\right\}$.

(b) We also cannot have LL agreeing and HL splitting. Otherwise, let (θ^*_H, θ^*_L) with $\theta^*_H + \theta^*_L > 1$ be the incompatible shares demanded by H and L respectively in an unbalanced pair and $(\theta', 1-\theta')$ the shares agreed to in an LL pair, with $\theta' \geq 1/2$. Consider now a deviation by the partner who was getting $1 - \theta'$, to some $\theta''' > \theta'$ and $\theta''' \notin \left\{\theta^*_H, \theta^*_L\right\}$, and distinguish the following cases.

(i) If $\theta' \neq \theta^*_H$ the non-deviating partner, who is still asking for the equilibrium share θ', remains unambiguously identified as L (by the first of our refinements), and the deviating partner as an H (by the second refinement, or by D1), thus achieving $(1+\lambda s)v_L + s(1-\lambda)v_H > (1+s)y_B/2$, since $s > s^{**}$. A fortiori, this is better than his equilibrium utility $(1+s)(1-\theta')y_B$.

(ii) If $\theta' = \theta^*_H$, this implies $\theta^*_H \geq 1/2$. The LL partner receiving $1 - \theta^*_H$ in equilibrium (say, Player 1) can profitably deviate to $\theta''' > \theta^*_H$ (clearly $\theta" = 1 - \theta^*_H > 0$, otherwise LL pairs are unsustainable), with $\theta''' \neq \theta^*_L$. Indeed, by our first refinement Player 2 is then presumed to have played according to equilibrium (which stipulates θ^*_H for both H types and one side in LL pairs), while the fact that Player 1 broke the match identifies him, by $D1$, as an H type. Since $s > s^{**}$ this is again a profitable deviation.

It follows from the two Lemmas that, for $s > s^{**}$, no matches can be sustained, even through asymmetric equilibria.