**Conférence Jean-Jacques Laffont**

# THE ECONOMICS OF MOTIVATED BELIEFS

**Roland Bénabou**

**Dalloz** | « Revue d'économie politique »

2015/5 Vol. 125 | pages 665 à 685

ISSN 0373-2630

Article disponible en ligne à l'adresse :

---

http://www.cairn.info/revue-d-economie-politique-2015-5-page-665.htm

---

!Pour citer cet article :

---

Roland Bénabou, « The Economics of Motivated Beliefs », *Revue d'économie politique* 2015/5 (Vol. 125), p. 665-685.
DOI 10.3917/redp.255.0665

---

# Conférence Jean-Jacques Laffont
# The Economics of Motivated Beliefs

Roland Bénabou*

I present the key ideas and results from recent work incorporating "motivated" belief distortions into Economics, both at the individual level (overconfidence, wishful thinking, willful blindness) and at the social one (groupthink, team morale, market exuberance and crises). To do so I develop a flexible model that unifies much of this line of research, then relate its main assumptions and testable predictions to the relevant experimental and observational evidence.

***beliefs – wishful thinking – overconfidence – hubris – groupthink – group morale – organizational culture – market exuberance – speculative bubbles – financial crisis – cognitive biases – cognitive dissonance – motivated cognition – anticipatory utility – memory – non-bayesian updating – information aversion – willful ignorance – psychology***

## *L'Économie des croyances motivées*

Je présente les idées et résultats principaux émanant des travaux récents qui visent à incorporer les croyances motivées dans le champ de l'Économie, que ce soit au niveau individuel (excès de confiance, déni de réalité, aveuglement délibéré) ou social (pensée de groupe, moral d'équipe, exubérance et crises des marchés financiers). Pour ce faire, je développe un modèle flexible permettant d'unifier cette ligne de recherche, et confronte systématiquement ses principales hypothèses et prédictions à l'évidence empirique et expérimentale.

**croyances – illusions – excès de confiance – hubris – pensée de groupe – moral collectif – culture organisationnelle – exubérance des marchés – bulles spéculatives – crises financières – biais cognitif – cognition motivée – dissonance cognitive – utilité d'anticipation – mémoire – apprentissage non-bayésien – aversion à l'information – ignorance délibérée – psychologie**

# 1. Introduction

It is a great honor to give the Jean-Jacques Laffont *AFSE* lecture. It also allows me to acknowledge an important debt, as this work would not have seen the light of day without Jean-Jacques. Indeed it was during the two

* Princeton University, CIFAR, NBER, CEPR, IZA and BREAD. Email: rbenabou@Princeton.EDU
I thank Jean Tirole for helpful comments. Financial support from CIFAR is gratefully acknowledged.

years I spent in Toulouse at his invitation that I started working, together with Jean Tirole, on a number of topics at the boundary of economics and psychology. Jean-Jacques was somewhat of a skeptic about Behavioral Economics, fearing its potentially paternalistic implications and misuses by policymakers. Sometimes he would come into the office were Jean and I were working and we would explain how agents in our models are always unsure of their willpower, ability or morality, constantly trying to maintain positive self-images and identities. Jean-Jacques would then quip that, at our age, we really ought to have figured out by now who we were. We would answer that time and life mercilessly alters one's type over time, so that it is a never-ending quest. Of course, Jean-Jacques's great intellectual curiosity and long-term focus naturally prevailed over his doubts about where departing from standard rationality might lead economics, and he always encouraged us to pursue this unchartered course.

In this article, I will provide a brief overview of some the recent work on the economics of "motivated" belief distortions, both individual and social, much of it done with Jean Tirole. Along the way, I will also emphasize the constant back and forth from empirical puzzles to theoretical modeling and then back to empirical tests – in particular, through experiments – that has been so important to progress in this area.

## 2. Beliefs and misbeliefs

A large majority of people believe they are more likely than others to experience favorable life events and, especially, less likely to suffer adverse ones such as unemployment, serious illness, divorce, accident, etc. (*e.g.*, Weinstein [1980]). We also commonly see ourselves as better drivers, better citizens, less biased and more attractive than others.

Some widely held beliefs are just plainly implausible or demonstrably false, given publicly available knowledge. One could point here to creationism or the billions of dollars spent each year on astrology, but I will confine myself here to a more standard domain. Case and Shiller [2003] surveyed the expectations of homeowners in four major US metropolitan areas during two real-estate bubbles, in 1988 and 2003. In all cases, about 90 % of respondents thought housing prices in their city would "increase over the next several years", with an average expected gain "[for] your property... over the next ten years" of 9 to 15 % *per annum*, which is close to an overall tripling. We know the end of those stories.

People also have *persistently divergent* perceptions of the world they jointly observe. This is clearest for political and economic beliefs, and our country is a case in point. In 2005, the World Public Opinion Survey polled citizens in twenty nations on their degree of (dis)agreement with the statement: "the free enterprise system and free market economy is the best system on which to base the future of the world". Average agreement was 61 %, with the US predictably higher (71 %) but China even above (74 %). At

the very bottom were Russia (43 %), Argentina (42 %) and... France (36 %). Objective data cannot justify such widely divergent worldviews, in particular the nearly two-to one ratio between France and neighboring Germany (65 %), which have very similar economic structures. These (historically determined) ideological beliefs nonetheless have major consequences. For instance, the above numbers predict quite well the size of the State relative to the economy, whether measured by the tax-to-GDP ratio or by indices of labor and product market regulation (as shown in Bénabou [2008]).[1]

***Do they actually believe it?*** A natural first question is whether the inflated self-views and outlandish or mutually incompatible worldviews so commonly expressed are just "cheap talk", or whether people truly believe and *act* on them.

The prevalence of overoptimism and overconfidence, especially in men, has by now been documented in a large number of incentivized experiments (*e.g.*, Camerer and Lovallo [1999], Hoelzl and Rustichini [2005]). While some of the findings could be rationalized by subjects having private information from earlier experiences (Benoit and Dubra [2011]), more recent tests immune to this caveat confirm the reality of overconfidence (*e.g.*, Merkle and Weber [2011]) and, especially, of biased updating (as I shall discuss later on).

Turning to "real-world" decisions, longitudinal studies of income and consumption reveal that a significant fraction of US households have substantially inadequate life insurance, given the risks they face. Overconfident behavior with high costs has also been documented in economic domains such as individual stock trading (Barber and Odean [2001]) and corporate investment (Malmendier and Tate [2005]).

Underlying the "strange" beliefs we often hold concerning our abilities, morality and future fate are strategies and mental processes which psychologists term *motivated reasoning and cognition*, through which we defend them against threatening evidence, sometimes incurring (and inflicting) very high costs. To illustrate some of these mechanisms I will use an example concerning health, a domain where motivated beliefs are also prevalent, with very real consequences.

Huntington's disease is a degenerative brain disorder that causes an ever-worsening deterioration of physical and mental capacities and a drastic shortening of life expectancy. It is due to a mutated gene, so if a parent has it the child has a 50 % chance of inheriting it and also developing the disease. Diagnostic is based on the progression of symptoms or/and a genetic test that is fully accurate. Oster *et al.* [2013] followed 700 at-risk patients who had a parent with the gene but had not themselves been tested.

A first key finding is *failure to update to bad news*. As their "motor score" worsens and the probability of disease as assessed by clinicians rises all the way up to 99 %, a patient's own reported probability changes very little, staying close to 40 % on average. At nearly every stage of progression, there are even about 15 % who report a 0 % subjective likelihood of having the

1. On other dimensions of international differences in political beliefs, see also Alesina *et al.* [2001] and Bénabou and Tirole [2006].

disease. Is this just what participants in the study say to feel good (though it would have no reason to feel good if they did not believe it at all), or are they actually making important decisions based on it? A first crucial decision they are making is to *not* get tested: even as motor symptoms progress to an implied 99 % probability, less than 5 % of patients ever get the test. The study also tracks several important "life" decisions of the participants. Those who took the test and found out they have the gene show major adjustments compared to those for whom it was negative: they are significantly more likely to get pregnant (for women), retire early, divorce, make "big financial changes" and alter their recreational activities. In contrast, those who are "uncertain" (not having taken the test but with objective probabilities ranging from 50 % to 99 %) show no significant changes in life behavior from those who are truly at zero risk.

# 3. Motivated cognition: why and how

In thinking about these phenomena, it is useful to separate the "demand" side (*why* might people want to hold, or be drawn to, distorted beliefs?) from the "supply" side (*how* do they manage, or at least attempt to, hold such beliefs?).

Let us start with demand. In standard decision theory more accurate information is always (weakly) valuable, even when it is bad news. Yet we are all familiar with beliefs that have a direct and immediate *affective* impact such as moral self-esteem (Smith [1759]), or anticipated prospects that evoke strong feelings of fear, anxiety, hope, excitement, etc. (Akerlof and Dickens [1982], Loewenstein [1989]).

Subjective beliefs also often have an important *instrumental* value. First, confidence in one's ability and chances of success (or those of teammates) can be a powerful motivator to pursue difficult long-term goals and persevere through adversity.[2] Second, and related, being convinced of one's abilities (talent, strength, determination, honesty etc.) and sincerity can be very useful to convince others.[3]

The model presented below will incorporate both motives for departures from objective cognition: affective (feeling better) and instrumental (performing better). Depending on the context and tasks at hand, either one may be most relevant, and certain beliefs can also serve both functions. An important example of the latter is religion, which (to some) simultaneously provides self-discipline and reassurance, or consolation.

2. Consistent with this view, Puri and Robinson [2007] find, using data from the Survey of Consumer Finance, that more optimistic individuals work more, save more, expect to retire later and are more likely to remarry.

3. Von Hippel and Trivers [2011] hypothesize that this signaling value is why humans initially evolved the capacity to self-deceive. Charness *et al.* [2013] show that experimental subjects who know they will face a competitive task become overconfident when such beliefs confer a strategic advantage.

Turning now to the supply side, how are desired beliefs achieved and maintained, sometimes against strong evidence? The paths to self-deception are countless, but three main categories can be distinguished: *willful blindness*, *reality denial*, and *self-signaling*.

The first one consists in avoiding information sources that *may* hold bad news. For Huntington's disease or HIV, for instance, this means not getting the test even though it is cheap or free, accurate, and can be done anonymously. Critical decisions need to be made, yet the person's words and deeds reveal a negative *ex-ante* value for information.

In the second scenario the news are already accumulating, though not yet completely final: symptoms are worsening, the objective probability of disease is rising to 70 %, 80 %, etc., yet the patient finds ways of not internalizing the data, rationalizing it away and convincing himself that his risk is still only (say) 15 %, and behaving accordingly in most respects.

The third strategy is one where it is the agent himself who manufactures "diagnostic" signals of the desired type, which he then interprets as impartial (Quattrone and Tversky [1984], Bodner and Prelec [2003], Bénabou and Tirole [2004, 2011]). Keeping with the health example, this correspond to a person who "pushes" himself to overcome their symptoms, carrying out difficult or even dangerous activities not only for their own sake but also as "proof" that things are fine.

***Motives vs. heuristics.*** It is worth pointing out three fundamental differences between such motivated beliefs or cognitive tendencies and the more purely mechanical mistakes in inference associated to the "heuristics and biases" view (*e.g.*, Tversky and Kahneman) and typically found in most models of bounded-rationality:

1. The latter types of "errors" are automatic and undirected (an "intuitive" System I is often invoked), the former valenced (pleasant or aversive) and goal-oriented, though in general not consciously so. A clear example of the difference is that of *confirmation bias* versus *self-enhancement*, for someone who is already not very confident in their skill, attractiveness, health or other key characteristic. In the first case the person tends to interpret any ambiguous signals received as confirming and hardening their negative self-view. In the second they see the same evidence positively, as showing that things are actually pretty good, or not so bad. In practice, the great majority of people show the latter type of response, and only depressive ones the former.[4]

2. A second major difference is that people who are more analytically sophisticated, educated or numerate can actually be more prone to making distorted inferences – rationalizing away evidence and compartmentalizing knowledge to protect valued beliefs – than those with lower cognitive abilities. Moreover, such reversals of the standard bounded-rationality logic occur only when the issue at hand is *value-laden* (*e.g.*, gun control, climate change; see Kahan [2013] and Kahan *et al.* [2014]), and not when it is neutral.

4. See, *e.g.*, Alloy and Abrahamson [1979].

3. Unlike computational and statistical mistakes, motivated cognition is *emotionally charged*. This feature is revealed almost instantly by a "fighting response" (agitation, anger, outrage, hostility) whenever a cherished belief pertaining to a person's identity, morality, religion, politics, etc., is directly challenged by evidence. This view of belief formation is also consistent with the renewal of interest in emotions and their influence on decision-making currently under way in psychology and neuroscience (*e.g.* Sharot *et al.* [2012].

# 4. A portable model

The following framework brings together key elements from Bénabou and Tirole [2002] and Bénabou [2013]. A risk-neutral individual $i$ has a horizon of three periods: 0, 1 and 2. At $t=1$ he makes a decision $e^i \in \{0,1\}$, with effort cost $ce^i$, $c>0$.[5] In period 2 he will reap a final payoff $U_2^i = V(\theta, e^i, c, k_0^i | \sigma)$ that depends on the action taken, the state of the world $\sigma \in \{H,L\}$ determining its return $\theta_\sigma$, and possibly some initial endowment $k_0^i$: wealth, human or social capital, genes, etc. Period 0 is when information may be received and processed into the posterior beliefs carried into period 1.

Let us first start with cognitive distortions linked to a *self-efficacy* motive. As illustrated in **Figure 1**, the agent's effort decision at is subject to a temptation problem: whereas the cost is $c$ when evaluated *ex ante* (at $t=0$), at the moment when it must actually be incurred it is perceived as $c/\beta$, where $\beta<1$ is the usual hyperbolic discounting or "weakness of will". It can then be advantageous, in order to get oneself to persevere when the going gets tough, to hold a positive view of the (net) return to resisting temptation.[6]

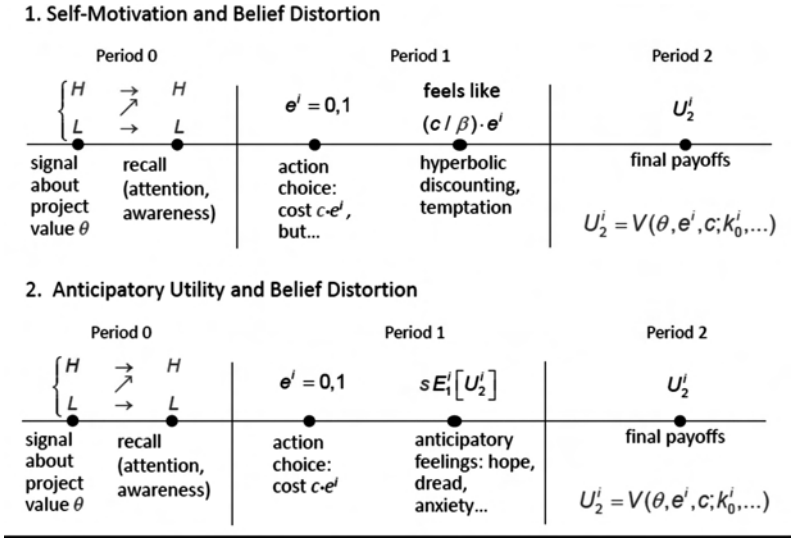
A first means to that end is strategic ignorance (Carrillo and Mariotti [2000]) or *willful blindness*: if his initial prior about $\theta_\sigma$ is good enough that he will work in period 1, the agent may prefer not to learn the true state of the world, for fear that bad news would discourage him from the *ex-ante* optimal level of effort ($c<\theta_L<c/\beta<\theta_H$). Second, and even more strikingly, if he did receive (*e.g.*, could not avoid) bad news, $\sigma=L$, he has an incentive to ignore, discount and misinterpret them. Such *ex-post denial* strategies or unconscious tendencies are represented in Figure 1 by the oblique arrow that "miscodes" state $L$ as state $H$. We thus allow the agent to process good and bad signals asymmetrically in term of *attention, interpretation, memory*

5. Note that $c>0$ is without loss of generality, by appropriate choice of the action corresponding to $e=1$.

6. This is more generally true as long as $\theta$ is a complement to effort, $V_{\theta e}>0$. In the case of substitutes (*e.g.*, when the task is to achieve some threshold level of performance), the incentive is to underestimate $\theta$ in order to guard against complacency – a form of "defensive pessimism" that can be handled very similarly (see Bénabou and Tirole [2002]).

*or awareness*. To a psychologist or neuroscientist there are very different mechanisms, but in terms of informational and behavioral outcomes they are formally equivalent: a signal of $L$ at $t=0$ is replaced at $t=1$ by some mixture (or garbling) of $L$ and $H$.

## Figures 1 and 2




The second, now *affect-driven* version of the model, is illustrated in **Figure 2**. Everything is exactly as before, except for period 1. There is no longer any hyperbolic discounting, but now the agent experiences anticipatory emotions from thinking about the future level of welfare $U^i_2 = V(\theta, e^i, c, k^i_0 | \sigma)$ he is likely to achieve in period 2: developing the disease or not, marriage succeeding or failing, firm delivering great riches or going bankrupt, house prices forever rising or crashing down. These feelings (often accompanied by somatic manifestations) constitute direct sources of (dis)utility during period 1, represented here by a flow $s \cdot E^i_1[U^i_2]$. The expectation $E^i_1[U^i_2]$ reflects what date-0 signal occurred and how it was processed, and its welfare impact is scaled by a parameter $s$ ("savoring" or "salience") that increases with the length of period 1.[7]

7. On beliefs and anticipatory feelings as direct objects of preferences see also Akerlof and Dickens [1982], Schelling [1986], Loewenstein [1987], Caplin and Leahy [2001], Brunnermeier and Parker [2005] and Köszegi [2010].

I shall now nest both versions of the model into a simple three-equation framework. In period 1, the agent will choose $e^i$ to maximize the expected PDV

$$U^i_1 = -(c/\beta)e^i + sE^i_1[U^i_2] + \delta E^i_1[U^i_2]. \quad [1]$$

Anticipating this, in period 0 he seeks, avoids or processes information with a tendency to maximize:

$$U^i_0 = -m^i/\beta + \delta E^i_0[-ce^i + sE^i_1[U^i_2]] + \delta^2 E^i_0[U^i_2], \quad [2]$$

where $m^i$ represents the direct costs (if any) of informational decisions at $t=0$: altering evidence, avoiding certain people, repressing unwelcome thoughts, etc. More interestingly, the other terms in (2) embody the *tradeoffs*, as seen from $t=0$, between the instrumental or/and affective *benefits* from optimism and the risk of costly *mistakes* (lowering $-ce^i + \delta U^i_2$ in some states of the world). Clearly, for $s=0$ equations (1)-(2) reduce to the self-efficacy case, for $\beta=1$ to the anticipatory-feelings one. More generally, the two can interact as complements or substitutes (see Bénabou and Tirole [2011] for such applications).[8]

I will now put some more structure on final payoffs, decomposing them into

$$U^i_2 = \alpha \cdot \theta_\sigma \cdot e^i + (1-\alpha) \cdot \kappa^i_\sigma, \text{ for } \sigma \in \{H, L\}. \quad [3]$$

The first term is the part of his "fate" over which the agent has control, through his date-1 action and its *return* $\theta_\sigma$ in state $\sigma$. The second, generally also state-dependent (and which can be arbitrarily correlated with $\theta_\sigma$), reflects fixed *stakes* which he cannot, or no longer, affect: age, gender, nationality, culture and any illiquid capital stocks (human, social, financial) resulting from decisions that are now *sunk*. Although $\kappa^i_\sigma$ is exogenous or predetermined for agent $i$, we shall analyze later on how it can still be endogenous at the level of a group, organization or market, reflecting how others think and behave (in equilibrium) when state $\sigma$ occurs.

***Updating.*** I turn next to belief formation, focusing here on the case of *ex-post* reality denial, which is richer than that of *ex-ante* information avoidance. Suppose that at date $t=0$ the agent learns the state of the world (or signal) $\sigma$, which is $H$ with prior probability $q$ and $L$ with probability $1-q$. The key building block (Bénabou and Tirole [2002]) is that he can respond to news with either *realism* or *denial*. Realism means objectively interpreting and remaining aware of $H$ as $H$ and $L$ as $L$. Denial corresponds to miscoding $L$ as $H$, recalling it as an ambiguous mixture of the two, or forgetting the news entirely.[9] In case of indifference the agent may randomize, which corresponds to partial or occasional awareness.

8. Note that when information is lost or distorted, the law of iterated expectations fails: $E^i_0[E^i_1[U^i_2]] \neq E^i_0[U^i_2]$.

9. I focus here on the case where agents seek to maintain optimism. Interpreting good news as neutral or even bad ones (forgetting $H$ or coding it as $L$) can also be a best response, given appropriate payoff structures. The agent may thus "lower his expectations" to guard against complacency (see footnote 6) or avoid disappointment, and a group can fall prey to collective fatalism and inaction (see Bénabou [2013]).

Let $\lambda^i \in [0,1]$ be the *equilibrium* probability with which agent $i$ attends to and correctly encodes bad news into memory, and $1-\lambda^i$ the complementary probability of self-deception. At $t=1$, if he is not aware of having received any negative news, his *posterior belief* that the state is truly $H$ is

$$r(\lambda^i) = \frac{q}{q + \chi(1-q)(1-\lambda^i)}. \quad [4]$$

For $\chi = 1$, this is simply Bayes' rule: the agent realizes that he has an average propensity $\lambda^i$ to forget or distort bad news, so at $t=1$ he corrects for it as best as he can. In particular, for $\lambda^i = 0$ the posterior remains the prior, $r(0) = q$. At the other extreme, an agent with $\chi = 0$ is fully "naïve" or, more generally, able to completely self-deceive about never being in state $L$.[10]

***Cognitive tradeoffs.*** To focus on the most interesting case, assume $\theta_H > \theta_L$ and that it is (*ex post*) optimal for the agent to exert effort at $t=1$ if his posterior on state $H$ is at least as good as his date-0 prior $q$, but not if he is aware that the state is $L$.

Consider now an agent at $t=0$, having just received news that $\sigma = L$. If, *in this particular instance*, he opts for realism, he will choose $e^i = 0$ at $t=1$ and achieve final utility $U_2^i = (1-\alpha)\kappa_L$. If he engages in denial, then at $t=1$ he will choose $e^i = 1$, expecting with probability $r(\lambda^i) \geq q$ to be in state $H$ and achieve final utility $U_2^i = \alpha\theta_H + (1-\alpha)\kappa_H$, and with probability $1 - r(\lambda^i)$ to be in state $L$ (as is really the case) and thus achieve only $U_2^i = \alpha\theta_L + (1-\alpha)\kappa_L$. Seen from $t=0$, the expected return to denial is thus

$$\Delta U_0^i(\lambda^i) \equiv U_{0,\,Denial}^i - U_{0,\,Realism}^i \quad [5]$$

$$\equiv -m/\beta - \delta[c - (\delta+s)\alpha\theta_L] + \delta s r(\lambda^i)[\alpha(\theta_H - \theta_L) + (1-\alpha)(\kappa_H - \kappa_L)].$$

The first term captures any direct cost of cognitive distortion at $t=0$, as described earlier. The second captures the objectively expected value of inducing effort at $t=1$. For a hyperbolic agent with

$$c < (\delta+s)\alpha\theta_L < c/\beta, \quad [6]$$

this represents a gain achievable through "positive thinking". For $\beta$ close enough or equal to 1, in contrast, it is a costly mistake, an investment with negative NPV. The motive for self-deception must then stem from affective reasons: a positive last term in (5) meaning that, conditional on effort, being in state $H$ is preferable to being in state $L$. The consumption value of living with more hopeful expectations during period 1 depends both on $r(\lambda^i)$, namely the extent to which the agent succeeds in persuading himself that

10. The model here is a simple three-period one, or iid repetitions of that stage game. Gottlieb [2010] shows that the main results extend to a dynamic setting in which the agent receives an infinite sequence of signals about the *same* variable $\theta_\sigma$ (e.g., talent), whether exogenously or by observing the outcomes of his actions.

the state is $H$ rather than $L$, and on $s$; the latter can be very large if, for instance, the final reckoning of date 2 is far in the future.

An *intrapersonal equilibrium* for individual $i$ (Perfect Bayesian equilibrium in the game between his date-0 and date-1 selves) is a value $\lambda^i$ such that.

$$\lambda^i \cdot \Delta U_0^i(\lambda^i) \leq 0 \leq (1-\lambda^i) \cdot \Delta U_0^i(\lambda^i). \quad [7]$$

For $\lambda^i = 1$ it corresponds to *constant realism* (and investment only in state $H$), for $\lambda^i = 0$ to *systematic denial* of state $L$ (and investment in both states), and for $0 < \lambda^i < 1$ to a mixed strategy. In general there may be multiple equilibria, corresponding to different "cognitive styles" and associated degrees of self-trust, but I will not go into this topic here (see Bénabou and Tirole [2002, 2004]). Instead I will assume that, given his environment (which may include the strategies of other players), the individual has a unique cognitive best-response.[11]

# 5. Main implications and empirical evidence

From (5) we readily derive a number of intuitive predictions, which can then be confronted to data.

*1. Asymmetric updating and information avoidance.* When self-relevant beliefs are involved, an individual will tend to process good and bad news differently – trying to ignore, discount, rationalize away or "put out of mind" those he does not like. This predicted asymmetric response will then show up in the evolution of his posteriors, as well the decisions they induce.

In Möbius *et al.* [2010] and Eil and Rao [2011], subjects are first objectively ranked by IQ (also, in the second paper, attractiveness to the other sex); this corresponds to $\theta$ in the model. They next state their prior distribution over being in each decile of the subject pool, then their updated beliefs following each of two rounds of feedback in which they learn if they ranked above or below another, randomly drawn subject. Beliefs at each stage are elicited using incentive-compatible scoring rules and the experimenter knows all the information subjects receive in-between. The key finding in both papers is a *good news / bad news asymmetry*, as predicted by the model: subjects systematically under-update to negative news, and are much closer to proper updating for positive news.[12]

11. One can always restrict parameters to ensure uniqueness. Moreover, even with multiple equilibria the comparative-statics of the equilibrium set are straightforward, and in line with the implications discussed below.

12. As shown in Bénabou [2013], when $\chi < q/(1-q)$ the model generates strict under-updating (relative to Bayes' rule) to bad news and a lesser underadjustment (possibly none) to good news.

In both studies, moreover, subjects' willingness to pay for learning their true IQ or beauty rank at the end of the experiment was positive for those who had arrived at "good posteriors", but *negative* for those who had arrived at "bad" ones. This selective aversion to information is clearly reminiscent of the non-demand for testing in patients whose family history and symptoms put them at high risk of having Huntington's disease, as well as of the behavior of the investors studied in Karlsson *et al.* [2009], who look up the value of their portfolios online much more on days when the market as a whole is up than on those when it is down.[13]

Turning to a setting of educational and career choices, Wisfal and Zafar [2015] elicit NYU students' beliefs about their own future earnings and the average earnings in different majors. Then they provide the actual figures for each major, and finally elicit subjects' updated beliefs about their own expected incomes. An underestimation of population earnings by $1 000 results in an upward revision in self-earnings of $347 (significant at 1 %), compared with a downward revision of just $159 (significant only at 10 %) for a $1 000 overestimation. On the other hand, equality of the two estimates could not be rejected.

*2. Selective memory and other mental processes.* In the model, motivated updating is represented as selective recall or awareness of past data. As explained, this is only one of several complementary and *de facto* equivalent mechanisms, but it is a relatively easy one to test.

In Thompson and Loewenstein [1992] subjects representing opposite sides in a labor negotiation later remember, from the same case file, more facts favoring their position than going the other way. The more divergent their recalls, moreover, the longer and costlier is the delay to agreement in the bargaining phase.

In Chew *et al.* [2013], subjects answer four questions from an IQ test (Raven's matrices). Two months later they are shown the same four, plus two they had never seen, together with all the answers, and *incentivized* to recall whether they answered each one correctly, incorrectly, never saw it, or just cannot remember. The probability of "remembering" having correctly answered a question which in fact one failed is, on average, six times as high as the probability of the reverse error. The probability of not remembering one's answer, or whether one saw a question, is on average twice as high if the answer had been wrong than if it had been right. As for the questions they had never seen, 56 % of subjects "remembered" answering them correctly, versus 9 % incorrectly. Furthermore, the three types of positive-attribution recall biases were highly correlated across subjects.

Neuroscientists are now also starting to explore the deep mechanisms involved in selective recall and updating. Benoit and Anderson [2012] show that subjects are able to lower their later recall rates (for word pairs) by either blocking associations as they start to resurface or by focusing on different thoughts, and that different brain networks are involved in these

13. On what types of models can or cannot explain such attitudes to information behaviors, see also Eliaz and Spiegler [2006] and Gottlieb [2014], who builds on and extends the present framework.

two processes of *voluntary forgetting*. Sharot *et al.* [2012] confirm the general finding of asymmetric updating to good and bad news and show that distinct regions of the prefrontal cortex are involved in tracking estimation errors that call for positively vs. negatively valenced updates. Furthermore, highly optimistic individuals consistently exhibit reduced tracking of estimation errors of the latter type.

*3. The "better than average" effect.* Asymmetric responses to good and bad news readily lead to a distribution of posteriors where a very high fraction of people see themselves as above average (whether mean or median), as in the examples discussed earlier.[14]

*4. Costs and salience.* As shown by the role of $s$ in (5), manipulations of salience (experimental, commercial, religious, etc.) increase the propensity to motivated thinking and related behaviors, such as self-signaling. Conversely, beliefs for which the individual cost of being wrong (term $c-(\delta+s)\alpha\theta_L$) is small are more easily distorted by emotions, desires and goals. An often-mentioned example with important aggregate implications is political views (*e.g.*, Caplan [2007]). For voters' cognitive distortions to matter, however, they must also tend to align in the same direction, rather than offset each other. The emergence of ideologies is among the model's extensions to *social* cognition, and as such will be briefly discussed at the end of Section V.

*5. Stakes-dependent beliefs.* Consider an agent who entered period 0 with initial stock $k_0^i$ of some illiquid asset (housing stock, specialized human capital, social network, OTC security) that he must hold until period 2, and whose final return will be $\theta_\sigma$ in state $\sigma=H,L$, namely the same as for any new marginal units. Thus $\kappa_\sigma^i=\theta_\sigma k_0^i$, implying that

$$\frac{\partial \Delta U_0^i}{\partial k_0^i}=\delta s(1-\alpha)(\theta_H-\theta_L)r(\lambda^i). \qquad [8]$$

The incentive to self-deceive is thus greater, the greater is the amount of "sunk" capital of a type more valuable in state $H$.[15] This is what I term *stakes-dependent beliefs*, an important and empirically testable implication of the model (especially, its self-esteem or anticipatory-utility version).

There is increasingly sophisticated evidence of this phenomenon, first demonstrated by the psychologist Kunda [1987]. In Babcock *et al.* [1995], pairs of subjects are given the same case file from a lawsuit concerning a traffic accident. One is randomly assigned to be the advocate for the plaintiff and the other for the defendant; they then bargain over a monetary settlement, with costs of delay. Both sides also (independently) make incentivized predictions as to how the judge ruled on the case and what outsiders would

14. This is true even in the case of "sophisticated" agents ($\chi=1$) where *ex-post* beliefs must average back to the true prior (and a fortiori for $\chi<1$), as Bayes' rule does not constrain skewness; see Carrillo and Mariotti [2000] and Bénabou and Tirole [2002].

15. Equation (8) is still implicit, as $\lambda^i$ is endogenous. The intuition which it provides goes through formally, however, so that the equilibrium (set of) $\lambda^i$ is decreasing in $k_0^i$; see equation (7).

deem fair. The findings are quite striking: when roles are assigned *before* subjects see the materials, their predictions on fairness and legal outcome are highly divergent and they make incompatible bargaining demands, leading to costly delay and breakdown. When roles are assigned *after* the information has been received and assimilated, by contrast, there is far less asymmetry and delay.

In Mijovic-Prelec and Prelec [2010], subjects tend to optimistically alter their (incentivized) assessments of an exogenous binary variable, once given stakes in one or the other outcome. In Mayraz [2011], subjects randomly assigned to being "farmers" or "bakers" forecast the price at which they will later trade grain. Their predictions again vary systematically and optimistically with their positions, as well as with the size of the monetary stakes involved in facing favorable terms of trade. Another example linking stakes and beliefs which I discuss later on is Cheng *et al.* [2014].

Moral judgements and decisions are particularly prone to self-serving beliefs and perceptions. I will not cover here this fast-growing segment of the literature, but simply list a few recent experimental demonstrations of self-deception over one's morality or altruism, such as Konow [2005], Dana *et al.* [2007], Di Tella *et al.* [2014] and Gneezy *et al.* [2014].

*6. Sunk-cost fallacy, escalating commitment and hedonic treadmill.* The above result is a form of endowment effect: an agent starting with enough of some illiquid, sunk type of asset has strong incentives to persuade himself of its future value. Once persuaded, he will want to invest more in this form of capital, etc. – a phenomenon psychologists refer to as *escalating commitment*. Furthermore, although the agent is optimizing at every point in time given his current preferences and beliefs, the welfare implications of such behaviors can be very negative – a loss of (*ex-ante*) intertemporal utility (Bénabou and Tirole [2011]).

This *hedonic-treadmill* result arises because, while censoring bad news or trying to offset them through self-signaling behaviors can successfully prevent a deterioration of beliefs in bad states, it also reduces the agent's confidence that good states are really what they seem to be: see (4), which embodies this "self-doubt" effect. When $\chi$ is close to 1 and beliefs enter anticipatory utility linearly as in (2) the two effects cancel out, leaving only the costs of achieving and / or acting on incorrect beliefs.[16]

# 6. Social and organizational beliefs

Large case-study literatures on corporate failures and scandals describe willful blindness and reality denial spreading within firms and other organizations.[17] How this contagion of misbeliefs can happen, and what are facilitating factors, is the question I turn to next.

16. The case of linear utility-from-beliefs is a useful benchmark. Clearly, if the functional is instead concave (resp., convex) in beliefs, the agent will gain from achieving coarser (respectively, more dispersed) posteriors. The actual shape of self-esteem or anticipatory preferences is, ultimately, an empirical question.

17. See, *e.g.*, the Online Appendix A in Bénabou [2013] for many examples.

I keep exactly the same model structure as in the basic anticipatory-utility case of Figure 1, but now allow final payoffs to explicitly reflect *social interactions*. Each agent is embedded in a firm, network or other collective endeavor where his final welfare is determined in part by his own action and in part by those of $n-1$ others, together with the state-dependent project return $\theta_\sigma$, on which everyone observes the same *public* signal $\sigma = H, L$. I focus here on the simplest interaction structure possible, both linear and symmetric:[18]

$$U_2^i = \theta_\sigma \left( \alpha \cdot e^i + (1-\alpha) \cdot \frac{1}{n-1} \sum_{j \neq i} e^j \right). \quad [9]$$

In terms of the general specification (3), this means an agent $i$'s sunk investment in the fate of the organization is $\kappa_\sigma^i = \theta_\sigma e^{-i}$, where $e^{-i}$ denotes the average action of others, making these stakes endogenous. It is important to note that $\theta_\sigma$ is the (social) return to action $e^i = 1$, *relative* to whatever is agents' next-best use of time of effort, $e^i = 0$. Since the return to the latter has been implicitly normalized to zero, the *net return* $\theta_\sigma$ can be of either sign.

I shall assume that $\alpha\theta_H > c > \alpha\theta_L$, which requires $\theta_H > 0$ but allows $\theta_L$ to be either positive or negative. In the first case, choosing $e^i = 1$ in state $L$ is individually suboptimal but constitutes a *public good* benefiting other agents. In the second, it is not only an individual mistake but also a *public bad*, inflicting losses on everyone else. This distinction will prove critical to how individuals' cognitive processes become interdependent in an organization or other network.

Suppose that, in equilibrium, a fraction $\lambda^{-i} \in [0,1]$ of agents $j \neq i$ respond to state $L$ with realism, while the remaining $1-\lambda^{-i}$ engage in denial.[19] The former will choose $e^j = 0$ and the latter $e^j = 1$, whereas when state $H$ occurs everyone exerts effort, $e^j \equiv 1$. Therefore:

$$\kappa_H - \kappa_L = [\theta_H - \theta_L (1-\lambda^{-i})] \quad [10]$$

Plugging into agent $i$'s incentive for denial computed in (5) and differentiating yields

$$\frac{\partial \Delta U_0^i}{\partial (1-\lambda^{-i})} = \delta sr(\lambda^i)(1-\alpha) \cdot (-\theta_L). \quad [11]$$

18. See Bénabou [2013] for the analysis and implications of asymmetric and/or non-separable interaction structures.

19. I am treating here $n$ as large enough to apply the Law of Large Numbers. In smaller groups $\lambda^{-i}$ and all the other expressions above would be expected values rather than deterministic outcomes; since agents are risk neutral, nothing would change.

This dependence of $\Delta U_0^i$ on $\lambda^{-i}$ makes clear how *endogenous cognitive linkages* arise whenever interacting agents form motivated beliefs, and this even though:

– All payoffs are additively separable in actions, as seen in (9).

– There is no private information that could give rise to herding or cascades.[20]

The intuition is simple: we saw earlier how each individual tends to align his beliefs with the stakes he has in different states of the world. These stakes now depend on what other people do, and hence on what they believe, in those states (the relevant one here is only $\sigma = L$ for simplicity). It follows that what is optimal for each agent to think *depends on what others think*, and vice versa. Furthermore, the nature and welfare consequences of these cognitive linkages depend very simply on the *sign of utility spillovers* (first rather than cross derivatives):

*1. Beneficial group morale:* when $\theta_L > 0$, perceptions of reality are *strategic substitutes:* in the bad state, the less others acknowledge reality, the better: they keep working, fighting and generating what is still a public good. The overoptimism of others thus makes state $\sigma = L$ more tolerable, and therefore each individual more willing to accept its reality. This case applies to relatively safe projects, team effort, political mobilization and other forms of (unconditionally) good citizenship.[21]

*2. Harmful group delusions*: when $\theta_L < 0$, perceptions of reality are *strategic complements:* people who do not recognize the reality of state $L$ and continue doing "business as usual" make things worse, not just for themselves but also for everyone else. Therefore, the more deniers there are the worse state $L$ becomes, making it more painful and scary for each agent to acknowledge the impending disaster. This case is typical of high-risk projects in which the downside is bad enough that blind persistence inflicts further expected damage on others: firm bankruptcy, layoffs, capital and reputational losses, prosecution, etc.

This result is rather perverse, yet quite robust: when denial or reality avoidance by others is socially beneficial it fails to spread, and when it is detrimental, it becomes *contagious*. When this *Mutually Assured Delusion* (MAD) effect is strong enough, moreover, multiple equilibria arise: fundamentally similar groups, firms or organizations (or the same one at different times) can operate either in a *realistic mode* where everybody faces the facts as they are, or in a *delusion mode* in which everybody engages in denial of bad news, which in turn makes those states even worse for everyone else.

***Groupthink.*** I formally show in Bénabou [2013] that such multiplicity (a positive measure of parameters $s$ or $c$ over which $\lambda^i = 1$ is the best response

20. In equation (11) $\lambda^i$ is endogenous, making the equation implicit. As can be seen from (7), it nonetheless provides the correct intuition for the formal result, which is that agent $i$*'s* best response $\lambda^i$ is increasing (resp., decreasing) in $\lambda^{-i}$ when $\theta_L < 0$ (resp., $\theta_L > 0$).

21. In a sufficiently asymmetric interaction structure, it may even be that some agent who can short-sell the project gains so much from others' denial of state $L$ that he prefers it to $H$, and as a result tends to believe in $L$ rather than $H$. This strong cognitive substitutability can lead two (sets of) agents to take opposite sides of a speculative bet on which state will realize, as in Brunnermeier and Parker [2005].

to $\lambda^{-i}=1$ and $\lambda^{i}=0$ the best response to $\lambda^{-i}=0$) arises, independently of $m$, if and only if

$$(1-q)(\theta_H-\theta_L)<(1-\alpha)(-\theta_L). \quad [12]$$

This simple formula has three important (and potentially testable) implications. First, $1-\alpha$ must be high enough: groupthink is more likely, the higher the "codependency" among members, meaning that they perceive that they share a largely common fate and have few exit options. Second and third, the adverse state of the world must be relatively rare (low $1-q$) but, when it occurs, really bad ( $\theta_L$ sufficiently negative).

Two interesting subcases can be further distinguished. When $\theta_H$ is positive but relatively low while $\theta_L<<0$, a denial equilibrium corresponds to a financial strategy of "picking pennies in front a steamroller" such as that of many hedge funds (*e.g.*, Long Term Capital Management) or, for an industrial company (Ford Pinto, BP, etc.), "saving pennies on safety" – all the while underestimating the tail risk of a disastrous outcome. When $\theta_H$ is very high and $\theta_L$ very negative, it corresponds to taking excessive amounts of two-tailed risk, *e.g.*, through oversized investments, extreme leverage, or even fraud and insider trading (*e.g.*, Enron, Global Crossing).

***Hierarchies and cognitive trickle-down.*** The *MAD* intuition embodied in (11) readily extends to asymmetric organizations and networks: an agent's propensity to realism or denial depends most on how the people whose decisions have the strongest impact on his fate (in state $L$) respond to bad news themselves. Therefore, in a hierarchy, top management's (mis)perceptions of market prospects, legal liabilities, odds of victory, etc., will tend to *trickle down* to middle echelons, and from there on to workers or troops.[22]

***Political ideologies.*** Bénabou and Tirole [2006] and Bénabou [2008] respectively embed the self-motivation and the anticipatory-utility versions of the model into simple political-economy frameworks. In the first paper, the unknown variable $\theta$ is the importance of effort (versus luck) in economic success; beliefs about it are natural complements to marginal (net-of) tax rates. In the second it is the relative efficiency of public (versus market) provision of goods and services like education, health care, insurance, etc.; beliefs about $\theta$ are then natural complements to the anticipated or/and inherited size of the public sector (stakes-dependence).[23] Thus, in both cases individual voters' beliefs about the economy's structure become strategic complements, leading to the emergence of different – broadly speaking, Left vs. Right – *dominant economic ideologies*, even across countries with the same fundamentals.

22. Another, complementary source of belief homogeneity in firms is the self-selection or deliberate screening of agents with (exogenously) differing priors, as in Van den Steen [2010].

23. The underlying intuitions can be seen from (i) a "tax-augmented" version (with $s=0$) of equation (6): $c<\delta\alpha(1-\tau)\theta_L<c/\beta$; (ii) a public-goods augmented (but here stripped-down of effort choices) version of equation (3): $U^{i}_{2}=\alpha(y^{i}-g)+(1-\alpha)\theta_{\sigma}\cdot(g-\varphi)$, where $y^{i}$ is agent $i$'s income, $g$ the provision of public goods (financed by a lump-sum tax for simplicity), $\theta_{\sigma}$ the relative efficiency of public versus private delivery, and $\varphi$ a minimum scale for government to have a net positive impact on welfare.

# 7. Wishful beliefs in financial markets

The groupthink logic of Section V also provides a *psychologically grounded* account of financial bubbles and crashes. Suppose that, following some initial good news, a continuum of investors $i \in [0,1]$ have accumulated stocks $k_0^i$ of some financial asset that is relatively illiquid, with $\int_0^1 k_0^i \, di \equiv K$.[24] Next, signals about fundamentals may stay green or turn to red, $\sigma = H, L$, and in each case investors can keep investing or stop, $e^i = 1, 0$, at cost $ce^i$. At $t = 2$, the market price $P_\sigma(K+E)$ will reflect equilibrium between demand for the asset $P_\sigma(\cdot)$ and total supply $K + E \equiv \int_0^1 (k_0^i + e_0^i) di$.

Final payoffs are thus

$$U_2^i = P_\sigma(K+E) \cdot (k_0^i + e^i), \sigma = H, L. \qquad [13]$$

In this interaction structure, agents' decisions $e^i$ are no longer additively separable as in (9) but *substitutes*, due to the standard effect of downward-sloping demand. This naturally tends to make contagion in investment harder to sustain: the more investors $j \neq i$ ignore a danger signal about fundamentals, $\sigma = L$, the higher is $E_L$, so the more $P_L$ will fall, making $e^i = 1$ even more of a costly mistake. In spite of this, investor's *responses to news at* $t = 0$ *can be strategic complements*, giving rise to the "irrationally exuberant" buildup that is the very source of the crash. Indeed, when illiquid initial positions $k_0^i$ are sufficiently large, facing reality $(\sigma = L)$ requires *recognizing* early on – in both senses of the term – major capital losses, $[P_H(K+E_H) - P_L(K+E_L)] \cdot k_0^i$, made all the worse by the blindness of others (which raises $E_L$).

***Market "exuberance" and meltdown.*** This capital-loss externality is the *MAD principle* at work again, and under appropriate conditions it can dominate the flow incentive effect from substitutability. The market is then seized by contagious overoptimism, leading to overinvestment and ultimately a deep crash. This "exuberant" equilibrium can also coexist (for the same fundamentals) with a realistic one in which investors pay attention to negative signals and thus limit the damages, for themselves and on each other. Viewed over time, this multiplicity corresponds to periodic waves of market overheating and cool-headedness.[25]

Is there evidence of such a mechanism? Focusing on the real-estate-based financial bubble of 2003-2005, Cheng *et al.* [2014] examine the *personal* housing transactions of 400 mid-level managers (traders, vice-presidents, etc.) in

24. I will treat here these initial inventories as exogenous, *e.g.*, reflecting past shocks, but they can also be the equilibrium result of a prior round of investment decisions (see Bénabou [2013]).

25. See Shiller [2005] for many examples throughout history and in recent years.

the mortgage-securitization industry, where toxic subprime loans, liar loans, etc., originated and were packaged for sale to banks and investors. Compared to equally sophisticated "outsiders" (lawyers not specializing in real estate, financial analysts covering non-housing companies) who had neither the means, private information nor the incentives for moral hazard, these Wall Street "insiders": (i) were more likely to buy a first, second or larger house at the peak of the bubble; (ii) slower to divest as prices started falling, and until the bust was well under way; (iii) consequently, worse performers in terms of the overall return on their own real-estate portfolios.[26]

Evidence that insiders bought high and sold low goes against standard moral-hazard accounts of the crisis in which agents with private information and "bad incentives" knowingly sold toxic assets to others. It also cannot be explained by large, "too-big-to-fail" banks taking on one-sided risk due to implicit bailout guarantees.[27] In contrast, it is very consistent with the model of escalating commitment and groupthink presented above, in which beliefs about future housing prices become badly distorted by personal (*e.g.*, human capital) and industry-wide stakes.

# 8. Conclusion: Bad incentives or / and bad beliefs?

In firms and organizations, the standard moral-hazard explanation for misbehavior is also often insufficient. A large literature in organizational psychology emphasizes the key roles of *moral self-deception* and overoptimistic *hubris* in many cases of corporate misconduct and financial fraud.[28] Most individuals engaging in dishonest behavior find ways to convince *themselves* that they are not doing anything wrong, and are still good persons. Transgressions most often start small and even unplanned, then gradually escalate through a series of self-serving rationalizations increasingly at odds with objective judgment and reality. Group dynamics, both of the "common fate" type analyzed in the groupthink model and linked to social norms (judging oneself relative to peers, excluding dissenters) also powerfully amplify these tendencies.

The above distinction is important and deserves more attention than it has so far received. In practice, of course, my take is that most cases involve bad incentives *and* bad beliefs, acting together as complements. This is still a

26. The authors rule out differential access to financing as a possible explanation: the housing transactions of the "insiders" and of the two control groups showed no difference in loan-to-value ratios or mortgage rates.

27. The study's sample included securitization managers and traders not only from large institutions but also from many midsize and small mortgage originators, regional lenders, hedge funds and investment firms that could not expect (and did not get) bailouts. Including firm fixed effects made no difference to the results.

28. For references and examples see, *e.g.*, Bazerman and Tenbrunsel [2011] and Bénabou [2013].

topic for ongoing and future work – a coming together of agency theory and behavioral economics which, I like to think, Jean-Jacques Laffont would have looked upon favorably.

# References

AKERLOF G., and DICKENS W. [1982], "The Economic Consequences of Cognitive Dissonance", *American Economic Review*, 72, 307-319.

ALESINA A., GLAESER E. and SACERDOTE B. [2001], "Why Doesn't the US Have a European-Type Welfare State?" *Brookings Papers on Economic Activity,* 2, 187-277.

ALLOY L. T., and ABRAHAMSON L. [1979], "Judgement of Contingency in Depressed and Nondepressed Students: Sadder but Wiser?" *Journal of Experimental Psychology: General*, 108, 441-485.

BABCOCK L., LOEWENSTEIN G., ISSACHAROFF S., CAMERER C. [1995], "Biased Judgments of Fairness in Bargaining", *American Economic Review*, 85, 1337-1343.

BARBER B. and ODEAN T. [2001], "Boys will be Boys: Gender, Overconfidence, and Common Stock Investment", *Quarterly Journal of Economics,* 116(1), 261-292.

BAZERMAN M. and TENBRUNSEL A. [2011], Blind Spots: *Why We Fail to Do What's Right and What to Do About It.* Princeton University Press.

BÉNABOU R. [2008], "Ideology", *Journal of the European Economic Association*, 6(2), 321-352.

BÉNABOU R. [2013], "Groupthink: Collective Delusions in Organizations and Markets", *Review of Economic Studies*, 80, 429-462.

BÉNABOU R. and TIROLE J. [2002], "Self-Confidence and Personal Motivation", *Quarterly Journal of Economics*, 117, 871-915.

BÉNABOU R. and TIROLE J. [2004], "Willpower and Personal Rules", *Journal of Political Economy*, 112, 848-887.

BÉNABOU R. and TIROLE J. [2006], "Belief in a Just World and Redistributive Politics", *Quarterly Journal of Economics*, 121(2), 699-746.

BÉNABOU R. and TIROLE J. [2009], "Over My Dead Body: Bargaining and the Price of Dignity", *American Economic Review*, Papers and Proceedings, 99(2), 459-465.

BÉNABOU R. and TIROLE J. [2011], "Identity, Morals and Taboos: Beliefs as Assets", *Quarterly Journal of Economics*, 126, 805-855.

BENOIT R. and ANDERSON M. [2012], "Opposing Mechanisms Support the Voluntary Forgetting of Unwanted Memories", *Neuron*, 76(2), 450-460.

BENOIT J. P. and DUBRA J. [2011], "Apparent Overconfidence", *Econometrica*, 79(5), 1591-1625.

BRUNNERMEIER M. and PARKER J. [2005], "Optimal Expectations", *American Economic Review*, 95(4), 1092-1118.

CAMERER C. and LOVALLO D. [1999], "Overconfidence and Excess Entry: An Experimental Approach", *American Economic Review*, 89, 306-318.

CAPLIN A. and LEAHY J. [2001], Psychological Expected Utility Theory and Anticipatory Feelings, *Quarterly Journal of Economics*, 116, 55-80.

CARRILLO J. and MARIOTTI T. [2000], "Strategic Ignorance as a Self-Disciplining Device", *Review of Economic Studies*, 67, 529-544.

CASE K. and SHILLER R. [2003], "Is There a Bubble in the Housing Market?" *Brookings Papers on Economic Activity*, 2, 299-342.

CHARNESS G., RUSTICHINI A. and VAN DE VEN J. [2013], "Self Confidence and Strategic Behavior", CESifo Working Paper Series No. 4517, December.

CHENG I.-A., RAINA S., and XIONG W. [2014], "Wall Street and the Housing Bubble", *American Economic Review*, 104(9), 2797-2829.

CHEW C. S., HUANG W. and XIAOJIAN Z. [2013], "Selective Memory and Motivated Delusion: Theory and Experiment", National University of Singapore mimeo, February.

DI TELLA R., GALIANI S., and SCHARGRODSKY E. [2007], "The Formation of Beliefs: Evidence from the Allocation of Land Titles to Squatters", *Quarterly Journal of Economics*, 122(1), 209-241.

DI TELLA R., PEREZ-TRUGLIA R., BABINO A. and SIGMAN [2015], "Conveniently Upset: Avoiding Altruism by Distorting Beliefs About Others' Altruism", Harvard University mimeo, May.

EIL D. and RAO A. [2011], "The Good News-Bad News Effect: Asymmetric Processing of Objective Information about Yourself", *American Economic Journal: Microeconomics*, 3(2), 114-138.

ELIAZ K. and SPIEGLER R. [2006], "Can Anticipatory Feelings Explain Anomalous Choices of Information Sources?", *Games and Economic Behavior*, 56, 87-104.

GNEEZY U., SACCARDO S., SERRA-GARCIA M. and VAN VELDHUIZEN R. [2014], "Motivated Self-Deception and Unethical Behavior", UCSD mimeo.

GOTTLIEB D. [2010], "Will You Never Learn? Self-Deception and Biases in Information Processing", Wharton School – University of Pennsylvania mimeo.

GOTTLIEB D. [2014] "Imperfect Memory and Choice under Risk", *Games and Economic Behavior,* 85, 127-158.

HOELZL E. and RUSTICHINI A. [2005], "Overconfident: Do You Put Your Money On It?" *The Economic Journal*, 115(503), 305-318.

KAHAN D. [2013], "Ideology, Motivated Reasoning, and Cognitive Reflection", *Judgment and Decision Making*, 8, 407-424.

KAHAN D., PETERS E., DAWSON E. and SLOVIC P. [2014], "Motivated Numeracy and Enlightened Self-Government", Yale Law School Cultural Cognition Project No. 107.

KARLSSON N., LOEWENSTEIN G. and SEPPI D. [2009], "The "Ostrich Effect": Selective Avoidance of Information", *Journal of Risk and Uncertainty*, 38(2), 95-115.

KONOW J. [2000], "Fair shares: accountability and cognitive dissonance in allocation decisions", *American Economic Review*, 90, 1072-1091.

KÖSZEGI B. [2010], "Utility from Anticipation and Personal Equilibrium", *Economic Theory* 44(3), 415-444.

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

LOEWENSTEIN G. [1987], "Anticipation and the Valuation of Delayed Consumption", *Economic Journal*, 97, 666-684.

MALMENDIER U. and TATE G. [2005], "CEO Overconfidence and Corporate Investment", *Journal of Finance*, 60(6), 2661-700.

MAYRAZ G. [2011], "Wishful Thinking", Oxford University Mimeo, October.

MERKLE C. and WEBER M. [2011], "True Overconfidence: The Inability of Rational Information Processing to Account for Apparent Overconfidence", *Organizational Behavior and Human Decision Processes*, 116, 262-271.

MIJOVIC-PRELEC D. and PRELEC D. [2010], "Self-Deception As Self-Signaling: A Model And Experimental Evidence", *Philosophical Transactions of the Royal Society*, B 365, 227-240.

MÖBIUS M., NIEDERLE M., NIEHAUS P. and ROSENBLAT T. [2010], "Managing Self-Confidence", Stanford University mimeo.

BARBER B. and ODEAN T. [2001], "Boys will be Boys: Gender, Overconfidence, and Common Stock Investment", *Quarterly Journal of Economics*, 116(1), 261-292.

PURI M., and ROBINSON D. [2007], "Optimism and Economic Choice", *Journal of Financial Economics*, 86(1), 71-99.

SCHELLING T. [1986], "The Mind as a Consuming Organ", in D. Bell, Raiffa H. and A. Tversky, eds., *Decision Making: Descriptive, Normative, and Prescriptive Interactions.* Cambridge, MA: Cambridge University Press.

SHAROT T., KORN C. and DOLAN R. [2012], "How Unrealistic Optimism is Maintained in the Face of Reality", *Nature Neuroscience*, 14(11): 1475-1479.

SHILLER R. [2005], *Irrational Exuberance.* Second Edition, Princeton University Press.

SMITH A. [1759], *The Theory of Moral Sentiments*. Reedited: Washington, DC, 1997, Regnery Publishing.

THOMPSON L. and LOEWENSTEIN G. [1992], "Egocentric Interpretations of Fairness and Interpersonal Conflict", *Organizational Behavior and Human Decision Processes*, 51, 17-197.

TVERSKY A. and KAHNEMAN D. [1974], "Judgment under Uncertainty: Heuristics and Biases", *Science*, New Series, 185(4157), 1124-1131.

VAN DEN STEEN E. [2010], "On the Origins of Shared Beliefs (and Corporate Culture)", *Rand Journal of Economics*, 41(4), 617-648.

VON HIPPEL W. and TRIVERS R. [2011], "The Evolution and Psychology of Self-Deception", *Behavioral And Brain Sciences*, 34, 1-56.