Groupthink:
Collective Delusions in Organizations and Markets

Roland Bénabou*

Princeton University

This version: September 2009

Abstract

I model (individually rational) collective reality denial in groups and organizations. When others' wishful thinking is beneficial to an agent, this makes it easier for him to be a realist; when it is harmful this pushes him toward denial, which becomes contagious. This simple and general mechanism leads to multiple social cognitions of reality, independently of payoff complementarities or private signals. In hierarchies, realism and denial will trickle down from the leaders. Contagious exuberance can also seize asset markets, generating investment frenzies and deep crashes. The welfare analysis differentiates group morale from groupthink and identifies a fundamental tension in organizations' attitudes toward dissent.

Keywords: groupthink, organizational culture, overconfidence, morale, market exuberance, manias, speculative bubbles, market crashes, financial crises, toxic assets, wishful thinking, cognitive dissonance, anticipatory feelings, psychology.

JEL Classification: D03, D23, D53, D83, D84, E32, G01, Z1.

*I am particularly grateful to Jean Tirole, as the paper builds on some of our earlier joint work. I am also thankful for valuable comments to George Akerlof, Daron Acemoglu, Alan Blinder, Markus Brunnermeier, Andrew Caplin, Sylain Chassang, Rafael Di Tella, Xavier Gabaix, Bob Gibbons, Boyan Jovanovic, Alessandro Lizzeri, Glenn Loury, Kiminori Matsuyama, Ben Polak, Eric Rasmussen, Ricardo Reis, Jean-Charles Rochet, Tom Romer, Julio Rotemberg, Tom Sargent, Hyun Shin, Glen Weyl, Muhamet Yildiz and to seminar participants at Princeton, NYU, Cornell, George Mason, INSEAD, Oxford, MIT, Brown, Yale, Chicago G.S.B., Berkeley, UCSD, UCLA, LSE, PSE, Penn Duke, Northwestern and Harvard universities. Rainer Schwabe and Andrei Rachkov provided superb research assistance. Support from the Canadian Institute for Advanced Research is gratefully acknowledged.

"It appears that there are enormous differences of opinion as to the probability of a failure with loss of vehicle and of human life. The estimates range from roughly 1 in 100 to 1 in 100,000. The higher figures come from the working engineers, and the very low figures from management. What are the causes and consequences of this lack of agreement? Since 1 part in 100,000 would imply that one could put a Shuttle up each day for 300 years expecting to lose only one, we could properly ask 'What is the cause of management's fantastic faith in the machinery?' " (Richard Feynman, in Rogers Commission Report, 1986)

"We have a wealth of information we didn't have before," Joe Anderson, then a senior Countrywide executive, said in a 2005 interview. "We understand the data and can price that risk." (*BusinessWeek,* "Not So Smart," August 2007)

This paper examines how collective beliefs and delusions arise and persist in organizations such as teams, firms, bureaucracies and markets. In the aftermath of corporate and public-policy disasters, it often emerges that participants fell prey to a collective form of overconfidence and willful blindness: mounting warning signals were systematically ignored or met with denial, evidence avoided, cast aside or selectively reinterpreted, dissenters discouraged and shunned. Market bubbles and manias exhibit the same pattern of investors acting "color-blind in a sea of red flags", followed by a crash.[1]

Janis (1972), analyzing policy decisions such as the Bay of Pigs invasion, the Cuban missile crisis and the escalation of the Vietnam war, identified in those that ended disastrously a cluster of symptoms for which he coined the term "groupthink".[2] Although some later work was critical of his characterization of those episodes, the concept has flourished and spurred a large literature in social and organizational psychology. Defined in Merriam-Webster's dictionary as *"a pattern of thought characterized by self-deception, forced manufacture of consent, and conformity to group values and ethics",* groupthink was strikingly documented in the official inquiries conducted on the Challenger and Columbia space shuttle disasters. It has also been invoked as a contributing factor in the failures of companies such as Enron

[1] I borrow here the evocative title of Norris' (2008) account of Merrill Lynch's mortgage securitization debacle. For detailed accounts of market manias and crashes, see Mackay (1980), Kindleberger and Aliber (2005) and Shiller (2005). Most recently, the Inspector General's Report (2009) on the SEC's failure concerning the Madoff scheme has over 130 mentions of the words "red flags".

[2] The eight symptoms were: (a) illusion of invulnerability; (b) collective rationalization; (c) belief in inherent morality; (d) stereotyped views of out-groups; (e) direct pressure on dissenters; (f) self-censorship; (g) illusion of unanimity; (h) self-appointed mindguards. The model developed here will address (a) to (g).

and Worldcom, in some decisions relating to the second Iraq war, and most recently in the housing and mortgage-related financial crisis.[3] At the same time, one should keep in mind that the mirror opposite of harmful "groupthink" is precious "group morale" and seek to understand how they differ, even though both involve the maintenance of collective optimism in spite of negative signals.

To analyze these issues, I develop a model of (individually rational) *collective reality denial* in groups and organizations, or among participants in a market. The model, which builds on and extends the selective-awareness (attention, memory) framework of Bénabou and Tirole (2002, 2006a), allows me to ask when individual tendencies toward wishful thinking and overoptimism reinforce or dampen each other. To make clear that groupthink is entirely distinct from standard linkage mechanisms, the benchmark setup has no complementarities or substitutabilities in agent's actions, nor any private signals that could give rise to social learning or herding. What emerges is thus a novel and surprisingly simple mechanism generating interdependencies in information processing, beliefs and behavior.

The key observation is that while each person decides how to interpret objective reality, that reality –promising, disappointing, or scary– is itself shaped by the actions of others, and therefore by their subjective mindsets. Thus, whenever an agent benefits on average from other' delusions, this tends to make him more accepting of the situation; and whenever their disconnect from reality makes him worse off this pushes him toward denial, which is then contagious. This *Mutually Assured Delusion* (MAD) principle can, in particular, give rise to multiple equilibria with different "social cognitions" of the same reality.

The same general principle implies that, in organizations where some agents have a greater impact on others' welfare than the reverse (e.g., managers and workers), strategies of realism or denial will "trickle down" the hierarchy, so that subordinates will in effect *take their beliefs from the leader.* In addition to collective illusions of control, it can also account for the mirror case of collective fatalism and resignation, such as public apathy in a crisis or "looking away" from humanitarian disasters.

[3] On the shuttle accidents, see Rogers Commission (1986) and Columbia Accident Investigation Board (2003). On Enron, see Samuelson (2001), Cohan (2002), Eichenwald (2005) and Pearlstein (2006). On Iraq, see e.g., Hersh (2004), Suskind (2004) and Isikoff and Corn (2007). On self-deception and self-serving rationalizations as key enablers of corporate misconduct, see Huseman and Driver (1979), Sims (1992), Tenbrunsel and Messick (2004), Anand, Ashforth and Joshi (2005) and Schrand and Zechman (2008).

The model's welfare analysis makes clear what factors distinguish valuable group morale from harmful groupthink and leads to interesting results concerning attitudes toward dissenting speech. In particular, it explains why organizations and societies find it desirable to set up ex-ante commitment mechanisms protecting and encouraging dissent (constitutional guarantees of free speech, whistle-blower protections, devil's advocates, etc.), even when ex-post everyone would unanimously want to ignore or "kill" the messengers of bad news.

In market interactions, prices typically introduce a substitutability between agents' decisions that works against collective belief. Nonetheless I show how, in asset markets with high uncertainty and limited liquidity (new types of securities, startup firms, housing), *contagious wishful thinking* can again take hold, leading to "exuberant" investment frenzies and, ultimately, deep crashes.

The paper thus has three objectives. First, to identify a new, simple and general mechanism leading to interdependence in beliefs and actions. Second, to analyze how it interacts with different organizational and market structures, shedding light on much-discussed phenomena for which no model exists. Finally, to derive testable comparative-statics predictions.

By contrast, the paper's aim is not to make (or add to) the empirical case about collective delusions. For motivation purposes, I simply complement the references cited in the introduction with an online appendix. Appendix A thus documents certain "patterns of denial" that recur strikingly across different settings –from NASA to the FED, SEC and Fannie Mae, from Enron to major investment banks, A.I.G and individual investors. Another point it emphasizes is the inadequacy of moral hazard as the sole explanation: rather than substitutes, overoptimistic hubris, distorted rationalizations and the like are most often indissociable complements to gambling with other people's money, lives, or the law.[4]

This work ties into four literatures. The first one centers on cognitive dissonance, self-deception and other forms of belief distortion. The second, closely related, concerns anticipatory feelings.[5] Most of these papers have focused on individual rather than social beliefs,

[4]Appendix A first reviews what key actors *said:* absurd probability assessments, "this time is different" and other flawed rationales, "fantastic faith in the machinery". It then turns to what they *did,* examining both investment behaviors (failure to divest or hedge, leading to large personal losses) and informational decisions: refusal to gather or even look at available evidence, dismantling of risk-management systems both ex-ante and once alarm signals start flashing red, "normalization of deviance" in response to ever-larger anomalies. Finally, it points out the recurrent role of *forgetting* by both individuals and institutions.

[5]On cognitive dissonance, see, e.g. Akerlof and Dickens (1982), Schelling (1986), Kuran (1993), Ra-

and none has asked the question which I take up here: when does wishful thinking or reality avoidance become "infectious", when is it self-limiting, and what are the welfare implications in each case? The paper's analysis of group morale and groupthink in organizations relates it to a third strand of literature, which deals with overoptimism and heterogeneous beliefs in firms.[6] In this work beliefs are most often exogenous (reflecting different priors), whereas here they endogenously spread, horizontally or vertically, through all or part of the organization. Beyond economics, the paper relates to the management literature on corporate culture and to the work in psychology on "social cognition".

Finally, the model's application to market manias and crashes links the paper to the literatures on bubbles and herding, although the mechanism is entirely different. With informational cascades, the key problem is a failure to aggregate private signals and its cure resides in more communication. Moreover, agents display the usual "hunger" for accurate knowledge and disregard their own signal only when the choices of others are likely to embody more information.[7] In market groupthink, by contrast, investors have access to the same or very similar information, but their processing of it is distorted by wishful thinking and this pattern of thought becomes contagious. Increasing communication and information sharing among participants (e.g., through intensified media reporting) only reinforces the mania, while dissonant advice from outsiders is systematically ignored or discounted.

Section 1 presents the general model and main results. Section 2 examines the implications for welfare and the treatment of dissenting speech. Section 3 extends the analysis to asset-market manias and crashes. Section 4 considers fatalism and collective apathy in the face of crises, as well as other dimensions of robustness. Section 5 concludes. Key proofs

bin (1994), Bénabou and Tirole (2002, 2004, 2006b), Compte and Postlewaite (2004) and Di Tella et al. (2007). On anticipation, see, e.g. Loewenstein (1987), Caplin and Leahy (2001), Landier (2000), Caplin and Eliaz (2005), Brunnermeier and Parker (2005), Bernheim and Thomadsen (2005), Köszegi (2006), Eliaz and Spiegler (2006), Bénabou and Tirole (2007) and Brunnermeier et al. (2007).

[6] On the theoretical side, see, e.g. Rotemberg and Saloner (1993), Bénabou and Tirole (2003), Fang and Moscarini (2005), Van den Steen (2005), Gervais and Goldtsein (2007) and Landier, Sraer and Thesmar (2009). On the empirical side and focussing on CEO overconfidence, see, e.g. Malmendier and Tate (2005, 2008) or Camerer and Malmendier (2007). For psychology references on social cognition, see Bénabou (2008).

[7] See, e.g., Banerjee (1992), Bikhchandani, Hirshleifer and Welch (1992), Caplin and Leahy (1994), Chamley and Gale (1994) or Scheinkman and Xiong (2003). In versions of herding models with naive agents (e.g., Eyster and Rabin (2009)), the last feature needs to be qualified ("when they *think* the choices of others embody..."), but the others remain unchanged: no wishful thinking or motivated reasoning of any kind, standard demand for objective data (if informative enough) rather than information-avoidance and distortion, and elimination of the problem through public communication and information-sharing.

are gathered in the main Appendix, more technical ones in online Appendix B.

1. Groupthink in teams and organizations

"The Columbia accident is an unfortunate illustration of how NASA's strong cultural bias and its optimistic organizational thinking undermined effective decision-making." (C.I.A.B., 2003)

1.1. The benchmark model

• *Technology.* A group of risk-neutral agents, $i \in \{1, ...n\}$, are engaged in a joint project (team, firm, military unit) or other activities generating spillovers. At $t = 1$, each chooses effort $e^i = 0$ or 1, with cost $ce^i, c > 0$. At $t = 2$, he will reap expected utility

$$U_2^i \equiv \theta \left[\alpha e^i + (1-\alpha)e^{-i}\right], \tag{1}$$

where e^{-i} denotes the average effort of others,

$$e^{-i} \equiv \frac{1}{n-1}\sum_{j \neq i} e^j, \tag{2}$$

and $1-\alpha \in [0, 1-1/n]$ the degree of interdependence, reflecting the nature of the enterprise or the presence of cross-interests.[8] Depending on α, the choice of e^i thus ranges from a pure private good (or bad) to a pure public one. This linear payoff structure is maximally simple: all agents play symmetric roles, there is a fixed value to inaction $e = 0$, normalized to 0, and *no complementarity or interdependence of any kind* between agents' effort decisions.[9] These assumptions serve only to highlight the key mechanism, and will all be relaxed later on.

The productivity of the venture agents are engaged in is a priori uncertain: its expected value is $\theta = \theta_H$ in state H (probability q) and $\theta = \theta_L$ in state L (probability $1-q$), with $\Delta\theta \equiv \theta_H - \theta_L > 0$ and $\theta_H > 0$ without loss of generality. Depending on the context, θ can represent the potential value of a firm's product or business plan, the state of the market, the suitability of a political or military strategy, or the quality of a leader. Note that θ also

[8] Another source of interdependence is *altruistic concern* among agents: family or kinship ties, social identity, etc. Thus, (1) is equivalent to $U_2^i \equiv \beta\theta e^i + (1-\beta)U_2^{-i}$ with $1-\alpha \equiv (1-\beta)(n-1)/(n-\beta)$. Altruistic links are explicitly studied in Section 4.1. Note also that while (1) suggests constant returns, crowding or scale economies can be captured by dividing θ by some appropriate function of n.

[9] I intentionally abstract from complementarities and substitutabilities to demonstrate that they are *neither necessary nor sufficient* for groupthink, which involves only the interplay of *cognitive* decisions.

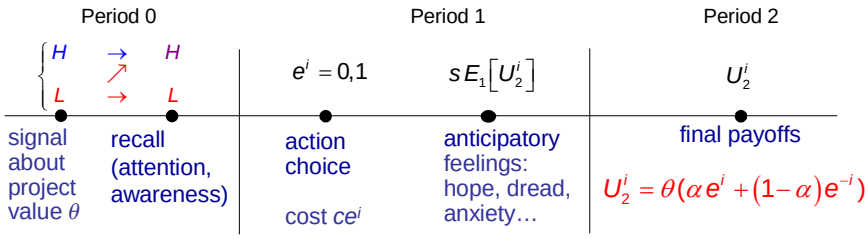

Figure 1: Timeline

corresponds to the expected social value of a choice $e^j = 1$, *relative* to what the alternative course of action would yield; the private value to the individual is $\alpha\theta - c$. If $\theta_L \geq 0$, each agent would always prefer that others choose $e = 1$ (put effort into the team or firm project, refrain from polluting, etc.). If $\theta_L < 0$, however, he would like them to pursue the "appropriate" course of action for the organization, choosing $e = 1$ in state H and $e = 0$ in state L.

While the extent to which an agent is vested in the collective outcome is here given $(1 - \alpha)$, it can in fact reflect a prior choice to join the organization.[10] Wishful beliefs about the enterprise's prospects (θ) will then also correspond to ex-post rationalizations and dissonance reduction of a sunk decision.

• *Preferences.* The flow payoffs received during period 1 include the cost of effort, $-ce^i$, plus the *anticipatory utility* experienced from thinking about one's future prospects, $sE_1^i[U_2^i]$, where $s \geq 0$ parametrizes the importance of hope, anxiety, dread, and similar emotions.[11] This parameter (s stands for "savoring" or "susceptibility") typically increases with the length of period 1, during which uncertainty remains. It may also vary across individuals, but for the moment I maintain symmetry.

At the start of period 1, agent i chooses effort to maximize the discounted value of payoffs,

$$U_1^i = -ce^i + sE_1^i\left[U_2^i\right] + \delta E_1^i\left[U_2^i\right]. \tag{3}$$

[10] Such a prior investment stage is modeled in Section 3, in the context of asset markets. I do not do so here to avoid repetition.

[11] This includes the well-documented health effects of chronic stress versus hopefulness. For models of anticipatory utility under uncertainty, see footnote 5. The linear specification $sE_1^i[U_2^i]$ avoids exogenously building into the model either information aversion or information-loving.

Given (1), his effort is determined solely by his beliefs about θ : $e^i = 1$ if $(s+\delta)\alpha E_1^i[\theta] > c$, *independently* of what any one else may be doing. I shall assume that

$$\theta_L < \frac{c}{(s+\delta)\alpha} < \frac{c}{\delta\alpha} < q\theta_H + (1-q)\theta_L. \tag{4}$$

Thus, absent credible information, an individual acting on his prior will choose $e^i = 1$, whereas one who knows for sure that the state is L will abstain.

An agent's beliefs at $t = 1$ depend on the news received at $t = 0$ and on how he processed them –accepting reality or averting his eyes from it, as specified below. In doing so, he acts to maximize the discounted utility of all payoffs

$$U_0^i = -M + \delta E_0^i\left[-ce^i + sE_1^i\left[U_2^i\right]\right] + \delta^2 E_0^i\left[U_2^i\right], \tag{5}$$

where E_t^i denotes expectations at $t = 0, 1$ and M the date-0 costs of his cognitive strategy. The tradeoff between accurate versus hopeful beliefs embodied in these preferences will manifest itself in behavior with respect to both date-0 information and date-1 choices.

• *Information and beliefs.* To represent agents' cognitive decisions or "patterns of thought", I use an extended version of the endogenous-recall or awareness technology introduced in Bénabou and Tirole (2002). At $t = 0$, agents observe a common signal that defines the relevant state of the world: $\sigma = H, L$, with probabilities q and $1 - q$ respectively.[12] Each then has some flexibility in how much attention to pay to it, how to interpret it, whether to "keep it in mind" or "not think about it", etc. Formally, he can:

(a) Accept the news realistically, thus truthfully encoding $\hat{\sigma}^i = \sigma$ into memory or awareness (his date-1 information set).

(b) Engage in denial, censoring or rationalization, thus encoding $\hat{\sigma}^i = H$ instead of $\sigma = L$, or $\hat{\sigma}^i = L$ instead of $\sigma = H$. In addition to impacting later decisions, this may entail an immediate cost $m \geq 0$.[13]

[12] As θ_σ is only the *expected* value of the project conditional on σ, a low signal does not preclude a high final realization, and vice versa. The perfect correlation of signals across individuals is also chosen for simplicity (it just needs to be positive) and to make clear that the mechanism at work here has nothing to do with herding or informational cascades, where agents with private signals make inferences from each other's behavior.

[13] Self-deception can be a deliberate strategy or an unconscious tendency, and the resources expended in the process may be material (eliminating evidence, avoiding certain people or situations, searching for and

(c) Deal in partial truths, using a mixed strategy. Equivalently, the memory process can be stochastic, with any recall probability $\lambda \in [0,1]$ achievable at cost $M = m(1-\lambda)$.

This simple informational structure captures a broad range of situations. The prior distribution $(q, 1-q)$ could itself be conditional on some other signal being good news, such as the appearance of a new technology or market opportunity (versus a status quo where θ is low for sure). This positive signal may also have warranted some initial investment in the activity, including the formation of the group itself.

• *Directed attention and inattention.* Instead of "tuning out" unwelcome news (denial), selective awareness can also take the form or investing extra resources in retaining good ones (rehearsal, preserving evidence). This corresponds to the case where attention or recall is naturally imperfect ($\lambda < 1$) but can be raised at some cost; it is like setting $m < 0$ in (b) above. Both mechanisms lead to broadly similar results, and can be combined: what matters is that there be a possibility (and a motive) for *differential awareness* of H and L, not how this is achieved. While costly recall may be a more familiar assumption, actual episodes of groupthink, market manias, etc., typically involve the more striking phenomena of willful inattention, ex-post rationalizations, refusals to face the evidence, silencing of doubters and similar forms of information disregard. For this reason, the model emphasizes "selective inattention" more than "selective attention".

A first result is that, no matter how small $m > 0$, an agent will never censor signals in both states: either $\lambda_H = 1$, or $\lambda_L = 1$. Given (1), moreover, intuition suggests that it is only in the low state L that he may do so: agents with anticipatory utility would not want to substitute bad news for good ones.[14] Verifying these claims in online Appendix B (Lemmas 3-4), I focus for the time being on cognitive decisions in state L, denoted simply

rehearsing desirable signals) or mental ones (stress from repression, cognitive dissonance, guilt). As discussed below, any arbitrarily small $m > 0$ suffices to rule out uninteresting "babbling" equilibria in which there is censoring in both states ($\lambda_L < 1, \lambda_H < 1$). Beyond this, all the paper's key results apply equally well with $m = 0$, though non-zero costs are more realistic, particularly for the welfare analysis.

[14] An agent who likes pleasant surprises and dislikes disappointments, on the other hand, may want to. Such preferences correspond (maintaining linearity) to $s = -\delta s'$, $0 < s' < 1$, so that the last two terms in (5) become $\delta^2 E_0^i \left[U_2^i - s' E_1^i \left[U_2^i\right]\right]$. All the results could be transposed to the case $s < 0$, leading to a (less empirically relevant) model of collective "defensive pessimism". By focussing on $s \geq 0$, I am implicitly assuming that the disappointment-aversion motive, if present, is dominated by anticipatory concerns. Such is the case, for instance, if the"waiting" period 1 is long enough. The potential social or evolutionary value of anticipatory concerns is discussed in Section 2.

$$\lambda \equiv \Pr\left[\hat{\sigma} = L | \sigma = L\right]. \tag{6}$$

Later on I will consider payoffs structures more general than (1), under which either state may (endogenously) be censored.[15]While agents can selectively process information, their latitude to affect beliefs remains constrained by Bayesian rationality: at $t = 1$, agent i may no longer have direct access to the original signal, but if he (as others) has a systematic tendency toward selective attention or interpretation, he will take that into account, using Bayes' rule to form posteriors. Thus, when $\hat{\sigma}^i = L$ he knows that the state is L, but when $\hat{\sigma}^i = H$ his posterior belief is only

$$\Pr\left[\sigma = H \mid \hat{\sigma}^i = H, \lambda^i\right] = \frac{q}{q + (1-q)(1-\lambda^i)} \equiv r(\lambda^i), \tag{7}$$

where λ^i is is his *equilibrium* degree of realism.[16]

To analyze the Perfect Bayesian equilibria of this game, I proceed in three steps. First, I fix everyone but agent i's awareness strategy at some arbitrary $\lambda^{-i} \in [0,1]$ and look for his "best response" λ^i.[17] Second, I identify the general principle that governs whether individual cognitions are strategic *substitutes* (the more others delude themselves, the better informed I want to be) or *complements* (the more others delude themselves, the less I also want to face the truth). Finally, I derive conditions under which groupthink arises in its most striking form, where both collective realism and collective denial constitute self-sustaining *social cognitions.*

[15]The restricted informational structure in which miscoding only occurs from L to H and never from H to L is equivalent to replacing the H state with a no-news state $\varnothing$, as in Bénabou and Tirole (2002, 2006).

[16] It is straightforward to allow for naiveté (forgetting that you forget, not realizing that you are prone to self-deception), parametrized for instance by a coefficient $\chi \leq 1$ multiplying $(1-q)(1-\lambda^i)$ in (7). This leaves all the positive results unchanged but can affect some of the welfare conclusions: see Bénabou and Tirole (2002, 2006b) for such a treatment of self-deception in different economic contexts. Finally, before making their investment decisions, agents could also have a choice of whether to seek further information, at some cost. Looking away from a red flag (coding L as H) can then imply failing to get additional data, whereas facing up to it (coding L as L) optimally triggers further investigation.

[17] With imperfect recall, each agent's problem is itself a game of strategic information transmission between his date-0 and date-1 "selves". Condition (4) and $m > 0$ will rule out any multiplicity of intrapersonal equilibria, simplifying the analysis and making clear that the groupthink phenomenon is one of *collectively sustained* cognitions. Note also that the focus on symmetric group equilibria, implicit in equating all λ^j's to a common λ^{-i}, is without loss of generality when there are many identical agents, as all best-respond to the aggregate. For finite n and/or heterogenous groups, there can also be asymmetric equilibria; see Section 1.4.

1.2. Best-response awareness

Following bad news, agents who remain aware that $\theta = \theta_L$ do not exert effort, while those who managed to ignore or rationalize away the signal have posterior $r(\lambda^j) \geq q$ and choose $e^j = 1$. Responding as a "realist" to a signal $\sigma = L$ thus leads for agent i to intertemporal expected utility (R is for "realism")

$$U^i_{0,R} = \delta(\delta + s)\left[\alpha \cdot 0 + (1-\alpha)(1-\lambda^{-i})\theta_L\right], \tag{8}$$

reflecting his knowledge that only the fraction $1 - \lambda^{-i}$ of other agents who are in denial will exert effort. If he censors, on the other hand, he will assign probabilities $r(\lambda^i)$ to the state being H, in which case everyone exerts effort with productivity θ_H, and $1 - r(\lambda^i)$ to it being really L, in which case only the other "optimists" like him are working and their output is $(1 - \lambda^{-i})\theta_L$. Hence ($D$ is for "denial"):

$$\begin{aligned} U^i_{0,D} \;=\; & -m + \delta\left(-c + \delta\left[\alpha + (1-\alpha)(1-\lambda^{-i})\right]\theta_L\right) \\ & +\delta s\left(r(\lambda^i)\theta_H + \left(1 - r(\lambda^i)\right)\left[\alpha + (1-\alpha)(1-\lambda^{-i})\right]\theta_L\right). \end{aligned} \tag{9}$$

Agent i's incentive to deny reality, given that a fraction $1 - \lambda^{-i}$ of others do so, is thus:

$$U^i_{0,D} - U^i_{0,R} = -m - \delta\left[c - (\delta + s)\,\alpha\theta_L\right] + \delta s r(\lambda^i)\left[(1-\alpha)\lambda^{-i}\theta_L + \Delta\theta\right]. \tag{10}$$

The second term is the net loss from mistakenly choosing $e^i = 1$ due to overoptimistic beliefs. The third one is the gain in anticipatory utility, proportional to the post-denial belief $r(\lambda^i)$ that the state is H and comprising two effects. First, the agent raises his estimate of the fraction of others choosing $e = 1$, from $1 - \lambda^{-i}$ to 1; at the true productivity θ_L, this contributes $(1-\alpha)\lambda^{-i}\theta_L$ to his expected welfare. Second, he believes the project's value to be θ_H rather than θ_L, so that when everyone chooses $e = 1$ his welfare is higher by $\Delta\theta = \theta_H - \theta_L$.

Agent i's incentive for denial is increasing in his own "habitual" truthfulness λ^i, ensuring a unique fixed point (personal equilibrium). This best response to how others *think* is characterized by the following properties, illustrated in Figure 2 by the dotted curves.

Proposition 1. ***(Optimal awareness and the MAD principle)*** *For any cognitive strategy λ^{-i} used by other agents, there is a unique optimal awareness rate λ^i for agent i :*
(i) $\lambda^i = 1$ for s up to a lower threshold $\underline{s}(\lambda^{-i}) > 0$, λ^i is strictly decreasing in s between $\underline{s}(\lambda^{-i})$ and an upper threshold $\bar{s}(\lambda^{-i}) > \underline{s}(\lambda^{-i})$, and $\lambda^i = 0$ for s above $\bar{s}(\lambda^{-i})$.
(ii) λ^i decreases with others' awareness rate λ^{-i} if $\theta_L > 0$, and increases with it if $\theta_L < 0$.
(iii) λ^i increases with the degree of spillovers $1 - \alpha$ if $\theta_L > 0$, and decreases if $\theta_L < 0$.

The first result is straightforward: the more important anticipatory feelings –the consumption value of beliefs– are to an agent's welfare, the more bad news will be repressed.

The second result brings to light a general insight which I shall term the *"Mutually Assured Delusion" (MAD)* principle. If others' blindness to bad news leads them to act in a way that is better for an agent than if they were well informed ($\theta_L > 0$), it makes those news not as bad, thus reducing his own incentive to engage in denial. But if their avoidance of reality makes things worse than if they reacted appropriately to the true state of affairs ($\theta_L < 0$), future prospects become even more ominous, increasing the incentive to look the other way and take refuge in wishful thinking. In the first case, individual cognitive strategies are strategic *substitutes,* in the latter they are strategic *complements.* It is also worth emphasizing that:

(a) This "psychological multiplier", less than 1 in the first case and greater in the second, arises even though agents' payoffs are separable and there is no scope for social learning. It thus represents a novel mechanism giving rise to interdependent beliefs and actions.

(b) The case in which individuals' willful blindness feeds on itself is also that in which it is worse for everyone, as it leads to the wrong course of action ($e^j = 1$ when $\sigma = L$).

Proposition 1 shows that the scope for contagion hinges critically on whether overoptimism has *positive or negative spillovers,* conditional on the bad-news state. Illustrative examples of both types of interaction are provided below.

• *Low-risk projects and public goods:* $\theta_L > 0$. The first scenario, best epitomized by a sports team, is that in which an individual's motivation and "can-do" optimism are always valuable to others: effort and quality control at work, political participation and other forms of good citizenship. More generally, it characterizes activities with a limited downside, in the sense that pursuing them remains *socially* desirable for the organization even in the low

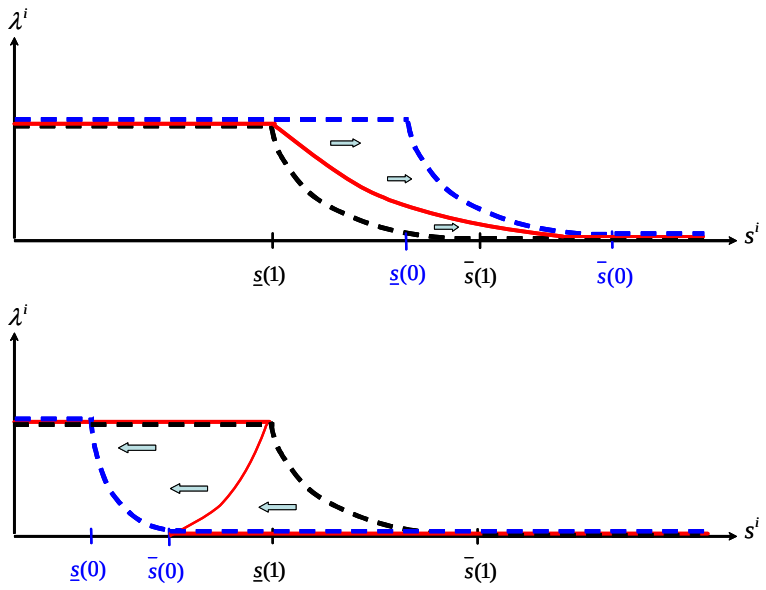

Figure 2: Group Morale ($\theta_L > 0$, upper panel) and Groupthink ($\theta_L < 0$, lower panel). The dotted lines give agent i's optimal awareness λ^i when others are realists ($\lambda^j = 1$) or deniers ($\lambda^j = 0$); the arrows indicate the shift between the two. The solid lines define the social equilibria.

state where the private return falls short of the cost. In the financial sector, this corresponds to making "plain vanilla" home loans or lending to stable brick-and mortar companies, which remains generally profitable even in a mild recession (though less than in a boom).

- *High-risk projects:* $\theta_L < 0$. The second scenario corresponds to ventures in which the downside is severe enough that persisting in this state has *negative social value* for the organization. The archetype is a firm like Enron, Bears Stearns or Lehman Brothers, whose high-risk strategy could be either extremely profitable (state H) or dangerously misguided (state L), in which case most stakeholders are likely to bear heavy losses: firm bankruptcy, layoffs, evaporated stock values, pensions and reputations, costly lawsuits or even criminal prosecution.

In such high-stakes contexts, the greater is other players's tendency (especially higher-ups, see below) to ignore danger signals and forge ahead with the strategy –accumulating yet more subprime loans and CDO's on the balance sheet, increasing leverage, setting up new off-the-books partnerships– the deeper and more widespread the losses will be if the scheme was flawed, the assets "toxic", or the accounting fraudulent and exposed as such. Therefore, when red flags start mounting, the greater is the temptation for each individual whose future

welfare is tied to the firm's fate to also look the other way, engage in rationalization, and "not think about it".[18]

The proposition's third result shows how both types of cognitive interdependencies are *amplified,* the more closely tied an individual's welfare is to the actions of others.[19] Three interesting implications ensue:

(a) Groupthink is likely to be particularly important for closed, cohesive groups whose members perceive that they largely share *a common fate* and have few exit options. This is in line with Janis' (1972) findings, but with a more precise notion of "cohesiveness".

(b) In groups with asymmetric roles, such as *hierarchies,* there will be a tendency to "follow the leader" into realism or denial. This idea is formalized in Section 1.4 below.

(c) Contagious beliefs are also more likely for *large-scale public goods*, such as those provided by a government, market, or other society-wide institutions which a single individual has little power to affect. This point is pursued in Bénabou (2008), where I study country-level ideologies concerning the relative efficacy of markets and governments.

1.3. Social cognition

I now solve for a full social equilibrium in cognitive strategies, looking for fixed points of the mapping $\lambda^{-i} \to \lambda^i$. The main intuition stems from Proposition 1 and is illustrated by the solid lines in Figure 2. First, $\lambda = 1$ is an equilibrium for $s \leq \underline{s}(1)$, as realism is the best response to realism; similarly, $\lambda = 0$ is an equilibrium for $s \geq \bar{s}(0)$, where denial is the best response to denial. Second, when $\theta_L > 0$ (cognitive substitutes), the thresholds $\underline{s}$ and $\bar{s}$ are both decreasing in λ^{-i}, so $\underline{s}(1) < \bar{s}(1) < \bar{s}(0)$ and the two pure equilibria correspond to distinct ranges. When $\theta_L < 0$ (cognitive complements), on the other hand, both thresholds are increasing in λ^{-i}, and if that effect is strong enough one can have $\bar{s}(0) < \underline{s}(1)$, creating a range of overlap.

[18] Enron's employees, whose pension portfolios had on average 58% in company stock, could have moved out at nearly any point, but most never did (Samuelson (2001)). At Bears Stearns, 30% of the stock was held until the last day by employees –with presumably easy access to diversification and hedging instruments– who thus lost their capital together with their job. CEO James Cayne alone owned an unusually high 6% and went from billionaire to small millionaire in the process (spending most of the intervening months away playing golf and bridge). The pattern was similar at Lehman Brothers and many other financial institutions.

[19] This intuition is reflected in (10) through the term $(1-\alpha)\lambda^{-i}\theta_L$. A lower α also increases the cost of suboptimal effort when $\theta_L > 0$ and raises it when $\theta_L < 0$, reinforcing this effect (term $c - \alpha(\delta + s)\alpha\theta_L$).

Proposition 2. ***(Groupthink)*** *(1) If the following condition holds,*

$$(1-q)(\theta_H - \theta_L) < (1-\alpha)(-\theta_L), \tag{11}$$

then $\bar{s}(0) < \underline{s}(1)$ *and for any* s *in this range, both realism* ($\lambda = 1$) *and collective denial* ($\lambda = 0$) *are equilibria, with a mixed-strategy equilibrium in between. Under denial agents always choose* $e^j = 1$*, even when it is counterproductive.*
(2) If (11) is reversed, $\underline{s}(1) < \bar{s}(0)$*. The unique equilibrium is* $\lambda = 1$ *to the left of* $(\bar{s}(1), \underline{s}(0))$*, a declining function* $\lambda(s)$ *inside the range, and* $\lambda = 0$ *to the right of it.*

Equation (11) reflects the MAD principle at work. The left-hand side is the basic incentive to think that actions are highly productive (θ_H rather than θ_L) when there are no spillovers ($\alpha = 1$) or, equivalently, when fixing everyone else's behavior at $e = 1$ in both states. The right-hand corresponds to the expected losses –relative to what the correct course of action would yield– inflicted on an individual by others' delusions, and which he can (temporarily) avoid recognizing by denying the occurrence of the bad state altogether. These endogenous losses, which *transform reality from second best* to *third best,* must be of sufficient importance relative to the first, unconditional, motive for denial.

• *Comparative statics.* The proposition also yields several testable predictions. First, there is the reversal in how agents respond to others' beliefs (or actions) depending on the sign of θ_L, with the very different equilibrium patterns that result. Second, and focusing on the more interesting case where (11) holds:

(a) The more vested in the group outcome are its members, the more likely is collective denial: as shown in the Appendix, both thresholds $\bar{s}(0)$ and $\underline{s}(1)$ decrease with $1-\alpha$.

(b) A more desirable or more plausible high state (higher θ_H or q) has the same effects.

(c) A worse low state (lower $\theta_L < 0$), arising for instance from a more risky project, has more subtle effects. On the one hand, it makes a realistic equilibrium easier to sustain ($\underline{s}(1)$ increases): the cost of making the wrong decision rises, while there is no harmful delusion of others to "escape from". When others are in denial, on the other hand, a lower θ_L makes it even worse. If $1/\alpha - 1/q$ (which must be positive by (11)) is relatively small, the first effect dominates and $\bar{s}(0)$ increases: sufficiently bad news will lead people to "snap out" of their collective delusion. With a sufficiently common fate or high priors ($1/\alpha - 1/q$ large enough),

on the other hand, the second effect dominates and $\bar{s}(0)$ decreases. The range over which multiplicity occurs thus widens, and a worsening of bad news can now cause a previously realistic group to take refuge in groupthink.

The types of enterprises that are most prone to collective delusions are thus:

(a) Those involving new technologies, products, markets or policies that combine a *highly profitable upside* and a potentially *disastrous downside.* High-powered *incentives,* when prevalent throughout the organization (e.g., performance bonuses affected by common market uncertainty) have a similar effect, as do highly leveraged investments.

(b) Those in which participants have only *limited exit options* and, consequently, a lot riding on the soundness or folly of other's judgements. Such dependence typically arises from irreversible or illiquid *prior investments:* specific human capital, professional reputation or network, company pension plan, etc. Alternatively, it could reflect the large-scale nature of the problem: state of the economy, quality of the government, global warming, etc.

The model also shows how a propensity to "can-do" optimism (high s) can be very beneficial at the entrepreneurial stage –starting a business, mobilizing energies around a new project ($\theta_L > 0$), but turn into a source of real danger once the organization has grown and is involved in more high-stakes ventures (e.g., a mean-preserving spread in θ, with $\theta_L > 0$).[20]

1.4. Asymmetric roles: hierarchies and corporate culture

> "And if the blind lead the blind, both shall fall into the ditch." (Matthew 15:14)

I now demonstrate the generality of the MAD principle by relaxing all the symmetry assumptions, as well as the state-invariance of the payoff to "inaction" ($e = 0$). I then use this more general framework to show how, in hierarchical organizations, denial and realism will "trickle down". Let the payoff structure (1) be extended to

$$U_2^i \equiv \sum_{j=1}^{n} \left(a_\sigma^{ji}\ e^j + b_\sigma^{ji}\ (1 - e^j) \right), \quad \text{for all } i = 1, \ldots n \text{ and } \sigma \in \{H, L\}. \tag{12}$$

[20] Similarly, through most of human history collective activities (hunting, foraging, fighting, cultivation) were typically characterized by $\theta_L > 0$, making group morale valuable and susceptibility to optimism (a high s or low m) an evolutionary advantageous trait. Modern technology and finance (e.g., leverage) now involve many high-stakes activities ($\theta_L << 0 << \theta_H$), for which those same traits can be a source of periodic trouble.

Each agent j's choice of $e^j = 1$ thus creates a state-dependent value a_σ^{ji} for agent i, while $e^j = 0$ generates value b_σ^{ji}; for $i = j$, these correspond to agent i's private returns to action and inaction. All payoffs remain linearly separable for the same expositional reason as before, but complementarities or substitutabilities are easily incorporated, as shown in Section 1.5. Agents may also differ in their preference and cognitive parameters c^i, m^i, δ^i, their proclivity to anticipatory feelings s^i or even their priors q^i. The generalization of (4) is then

$$a_L^{ii} - b_L^{ii} < \frac{c^i}{s^i + \delta^i} < q^i \left(a_H^{ii} - b_H^{ii}\right) + (1 - q^i)\left(a_L^{ii} - b_L^{ii}\right), \tag{13}$$

while the generalization of $\theta_H > \theta_L$ (H is the better state, conditional on everyone taking the optimal action), is

$$\sum_{j=1}^{n} a_H^{ji} > \sum_{j=1}^{n} b_L^{ji}. \tag{14}$$

Focussing here on pure-strategy equilibria, one can again compare an agent i's incentive to ignore a signal $\sigma = L$ when surrounded by deniers ($\lambda^j \equiv 0$) and by realists ($\lambda^j \equiv 1$). The condition for complementarity, generalizing $\theta_L < 0$, is now $\sum_{j=1}^{n}(a_L^{ji} - b_L^{ji}) < 0$, for all $i = 1, \ldots n$. In line with the MAD principle, it means that others' delusions, leading them to choose $e^j = 1$ when $\sigma = L$, are *on average* harmful to agent i. Multiple equilibria occur when this expected loss is sufficiently large relative to the "unconditional" incentive to deny:

$$(1 - q)\sum_{j=1}^{n}\left(a_H^{ji} - a_L^{ji}\right) < \sum_{j \neq i}\left(b_L^{ji} - a_L^{ji}\right). \tag{15}$$

Proposition 3. ***(Organizational cultures)*** *Let (13)-(15) hold for all $i = 1, \ldots n$. There exists a non-empty range $[\bar{s}^i(0), \underline{s}^i(1)]$ for each i, such that if $(s^1, \ldots s^n) \in \Pi_{i=1}^{n}[\bar{s}^i(0), \underline{s}^i(1)]$, then both collective realism ($\lambda^i \equiv 1$) and collective denial ($\lambda^i \equiv 0$) are equilibria.*

• *Directions of cognitive influence.* Going beyond multiplicity, interesting results emerge for organizations in which members play asymmetric roles. Indeed, the thresholds $\bar{s}^i(0)$ and $\underline{s}^i(1)$, given in appendix, confirm the intuition that each agent's optimal awareness is most sensitive to how the people whose decisions have the greatest impact on his welfare (the largest contributors to the right-hand-side of (14)) deal with unwelcome news.

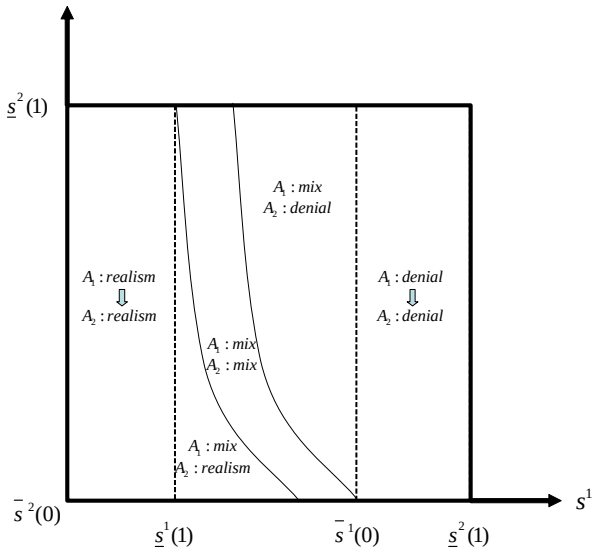

Figure 3: "Trickle down" of realism and denial in a hierarchy

As an application, consider the simplest form of hierarchy: two agents, 1 and 2, such as a manager and worker. If $a_L^{12} - b_L^{12}$ is sufficiently negative while $|a_L^{21} - b_L^{21}|$ is relatively small, agent 2 suffers a lot when agent 1 loses touch with reality, while the converse is not true. Workers thus risk losing their job if management makes overoptimistic investment decisions, whereas the latter has little to lose (perhaps the reverse) if workers put in more effort than realistically warranted. When the asymmetry is sufficiently pronounced it leads to a testable pattern of predominantly *top-down cognitive influences,* illustrated in Figure 3.

Formally (see online Appendix B), $[\underline{s}^1(1), \bar{s}^1(0)] \subset [\bar{s}^2(0), \underline{s}^2(1)] \equiv S$ and for all $(s^1, s^2) \in S \times S$ there is a *unique equilibrium,* such that

(a) The qualitative nature of the manager's cognitive strategy –complete realism, complete denial, or mixing– depends only on her own s^1, not on the worker's s^2.

(b) If the manager behaves as a systematic denier (respectively, realist), so does the worker: where $\lambda^1 = 1$ it must be that $\lambda^2 = 1$, and similarly $\lambda^1 = 0$ implies $\lambda^2 = 0$.

(c) Only when both agents are in partial denial (between the two loci in Figure 3) does the worker's degree of realism also influence that of the manager.

Let agent 2 now be replicated into $n-1$ identical workers, each with influence $[a_\sigma^{j1} e^j + b_\sigma^{j1}(1 - e^j)]/(n-1)$ over the manager or leader, but subject to the same influence from him as before, $a_\sigma^{1j} e^1 + b_\sigma^{1j}(1 - e^1)$. Figure 3 then remains operative, showing how *the leader's*

attitude toward reality tends to *spread to all his subordinates,* while being influenced by theirs only in a limited way, and over a limited range.

This result has clear applications to corporate and bureaucratic culture, explaining how people will contagiously invest excessive *faith in a leader's "vision".*[21] Likewise, in the political sphere, a dictator who is secure in his power need not exert constant censorship or constraint to implement his policies, as crazy as they may be: he can rely on people's mutually reinforcing tendencies to rationalize as "not so bad" the regime they (endogenously) have to live with.

The model is of course an oversimplified representation of an organization; yet the same general principles will carry over to more realistic hierarchies with multiple tiers, control rights, transfer payments, losers and gainers from the delusions of others, etc. I leave such extensions to future work and return below to the symmetric focus of Section 1.1.

1.5. Strategic interactions

To highlight the model's new source of interdependence in beliefs and behaviors, I have until now focussed attention on public-goods-like settings in which an agent's welfare level depends on others' actions, but his return to acting does not. Strategic complementarities in payoffs will, quite intuitively, reinforce the tendency for contagion, whereas substitutabilities will work against it.[22]

To see this, let agent i' expected payoff in state $\sigma = H, L$ now be $\Pi^i_\sigma(e^i, \mathbf{e}^{-i})$, where $\mathbf{e}^{-i}$ denotes the vector of others' actions; his incentive to act is then $\pi^i_\sigma(\mathbf{e}^{-i}) \equiv \Pi^i_\sigma(1, \mathbf{e}^{-i}) - \Pi^i_\sigma(0, \mathbf{e}^{-i})$. In state L, the differential in i's anticipatory value of denial that results from others' "blind" persistence, previously given by $-s(1-\alpha)\theta_L$, is now $-s\sum_{j\neq i}[\Pi^i_L(1, \mathbf{0})$

[21] In Rotemberg and Saloner (1993), a manager's "vision" (prior beliefs or preferences that favor some activities over others) serves as a commitment device to reduce workers' concerns about ex-post expropriation of their innovations. In Prendergast (1993), managers' use of subjective performance evaluations to assess subordinates' effort at seeking new information leads the latter to distort their reports in the direction of the manager's (expected) signal. Both mechanisms thus lead workers to "conform" their behavior to managers' prior beliefs. Unlike here, however, in neither case do they actually espouse those beliefs, nor would the manager ever want them to report anything but the truth. In Hermalin (1998), a leader with private information about the return to team effort works extra-hard to convince his coworkers to do so; the resulting separating equilibrium shifts up the whole profile of efforts (ameliorating the free-rider problem) but involves no mistaken belief by anyone.

[22] At the same time, without anticipatory feelings or some other "non-standard" role for beliefs, no amount of complementarity can generate results similar to the model's: agents with standard preferences always have (weakly) positive demand for information and thus never engage in denial or ex-post rationalizations.

$-\Pi_L^i(1,\mathbf{1})]$, which embodies the same MAD intuition as before. The new ingredient is that others' persistence now also changes the material value of investing in state L (previously a fixed $\alpha\theta_L$), by an amount equal to $\sum_{j\neq i}\left[\pi_L^i(\mathbf{1})-\pi_L^i(\mathbf{0})\right]$, with sign governed by $\Sigma_{\ j\neq i}\ \partial^2\Pi_L^i/\partial e^i\partial e^j$. When actions are complements, delusion is thus less costly if others are also in denial, whereas with substitutes it is more costly. Rather than restate general results with nonseparable payoffs, which would not add much insight, I shall focus in Section 3 on an important concrete application: how, in spite of investments being substitutes, asset markets can be seized by collective "manias", ultimately leading to a crash.

2. Welfare, Cassandra's curse and free speech protections

Are agents in collective denial worse or better off than if they squarely faced the truth –as an alternative equilibrium, or possibly by means of some collective commitment mechanism? Conversely, can they benefit from preserving a high morale if everyone is able commit to ignoring bad news?

Consider first state $\sigma = L$, which occurs with probability $1-q$. When agents are realists (setting $\lambda^j = 1$ in (8)), equilibrium welfare is $U_{L,R}^* = 0$. When they are deniers (setting $\lambda^j = 0$ in (9)), it is given by:

$$U_{L,D}^*/\delta = -m - c + \delta\theta_L + sq\theta_H + s\,(1-q)\,\theta_L. \tag{16}$$

As illustrated in Figure 4, whether collective denial of bad news is harmful or beneficial thus depends on whether s lies below or above the threshold

$$s^* \equiv \frac{m/\delta + c - \delta\theta_L}{q\theta_H + (1-q)\,\theta_L}. \tag{17}$$

Proposition 4. *Welfare following bad news (state L):*
(1) If $\theta_L < 0$, then $s^* > \max\{\bar{s}(0), \underline{s}(1)\}$. Whenever realism ($\lambda = 1$) is an equilibrium, it is superior to denial ($\lambda = 0$). Moreover, there exists a range in which realism is not an equilibrium but, if it can be achieved through collective commitment, yields higher welfare.
(2) If $\theta_L > 0$, then $s^* < \bar{s}(0)$. The equilibrium involves excessive realism for $s \in (s^*, \bar{s}(0))$ and excessive denial for $s \in (\underline{s}(1), s^*)$, when this interval is nonempty.[23]

[23]Part (1) follows directly from (17) and (A.4)-(A.5) in the Appendix. In Part (2), it is easily seen that

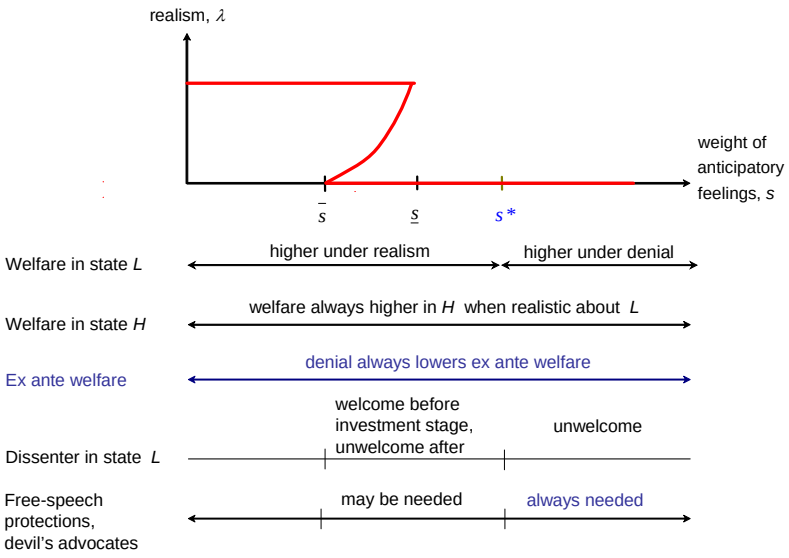

Figure 4: Welfare and dissenting speech (groupthink case)

Given how damaging collective delusion is in state L with $\theta_L < 0$, it makes sense that when realism can also be sustained as an equilibrium it dominates, and that when it cannot the group may try to commit to it. Conversely, with $\theta_L > 0$, *boosting morale* in state L *ameliorates the free-rider problem,* so the group would want to commit to ignoring bad signals when $s \geq s^*$ but the only equilibrium involves realism.[24]

Consider now welfare in state H, which occurs with probability q: given (4), everyone chooses $e^i = 1$ in both equilibria. Under denial, however, agents *can never be sure* of whether the state is truly H, or it was really L and they censored the bad news. As a result of this "spoiling" effect, welfare is only

$$U^*_{H,D}/\delta = -c + \delta\theta_H + s\left[q\theta_H + (1-q)\,\theta_L\right] < -c + (\delta + s)\,\theta_H = U^*_{H,R}/\delta. \tag{18}$$

Averaging over the two states, finally, the mean belief about θ remains fixed (by Bayes' rule), so the net welfare impact of denial, $\Delta W \equiv q\left(U^*_{H,D} - U^*_{H,R}\right) + (1-q)\left(U^*_{L,D} - U^*_{L,R}\right)$, is just

$$\Delta W \equiv (1-q)\left[(\delta + s)\,\theta_L - c - m/\delta\right], \tag{19}$$

$s^* < \bar{s}(0)$, but $s^* < \underline{s}(1)$ requires $(1-q)\,\Delta\theta[m/\delta + c - \delta\alpha\theta_L] < \delta\,(1-\alpha)\,\theta_L\theta_H$, which can go either way.

[24] If θ_L is high enough that $\delta\theta_L > c + m/\delta$, then $s^* < 0$: denial in state L is socially beneficial even absent anticipatory emotions ($s = 0$). The best example may be team morale in sports.

realized in state L. In assessing the overall value of social beliefs, one can thus focus on *material* outcomes and ignore anticipatory feelings, which are much more difficult to measure but wash out across states of nature.

Proposition 5. *(1) Welfare following good news (state H) is always higher, the more realistic agents are when faced with bad news (the higher is λ).*
(2) If $\theta_L \leq 0$, denial always lowers ex-ante welfare. If $\theta_L > 0$, it improves it if and only if $(\delta + s)\,\theta_L > c + m/\delta$.

These results, also illustrated in Figure 4, lead to a clear (and potentially testable) distinction between two types of collective beliefs and the settings that give rise to them.

• *Beneficial group morale.* When $\theta_L > 0$, $e = 1$ is socially optimal even in state L, but since $\alpha(s + \delta)\theta_L < c$ it is not privately optimal. If agents can all manage to ignore bad news at relatively low cost, either as an equilibrium or through commitment, they will thus be better off not only ex-post but also ex-ante: $\Delta W > 0$. This is in line with a number of recent results showing the functional benefits of overoptimism (achieved through information manipulation or appropriate selection of agents by a principal) in settings where agents with the correct beliefs would underprovide effort.[25]

• *Harmful groupthink.* The novel case is the one in which contagious delusions can arise, $\theta_L < 0$, and it also leads to a more striking conclusion: not only can such reality avoidance greatly damage welfare in state L, but even when it improves it those gains are always dominated by the losses induced in state H : $\Delta W < 0$.[26] This normative result has positive implications for how organizations deal with dissenters, revealing an interesting form of *time inconsistency* between ex ante and ex post attitudes. In carrying out this discussion, I shall refer interchangeably to "the group" and to "society", as in the case of political ideologies.

[25] In a team or firm context see, e.g., Bénabou and Tirole (2003), Fang and Moscarini (2005), Van den Steen (2005) and Gervais and Goldtstein (2007). In a self-control context, see Bénabou and Tirole (2002) and Battaglini et al. (2005). Also closely related to the present framework is Dessi (2008), who shows how one generation may want to collude in order to paint to the next one an overly optimistic picture of the benefits of cooperation. Dessi studies only the social-planner solution achieved through centralized control of beliefs (e.g., by an all-powerful state), not what equilibria arise from parents' individual child-rearing and indoctrination decisions, or their own ideological choices.

[26] The "shadow of doubt" cast over the good state by the censoring of the bad state could also distort some decisions in state H, although in this simple example it does not. Conversely, departing from Bayesian updating, (e.g., parametrizing naiveté as in footnote 16) would attenuate the losses in state H and thus allow ex-ante gains. See Bénabou and Tirole (2002, 2006c) for examples of both effects.

• *The curse of Cassandra.* Let $\theta_L < 0$ (more generally, $(\delta + s)\,\theta_L < c + m/\delta$) and consider a denial equilibrium, as illustrated in Figure 4. Suppose now that, in state L, an individual or subgroup with a lower s or a different payoff structure attempts to bring back the facts to everyone's attention. If this occurs after agents have have sunk in their investment it simply amounts to deflating expectations in (3), so they will refuse to pay attention, or may even try to "kill the messenger" (pay a new cost to forget). Anticipating that others will behave in this way, in turn, allows everyone to more confidently invest in denial at $t = 0$. To avoid this deleterious outcome, organizations and societies will find it desirable to set up *ex-ante guarantees* such as whistle-blower protections, devil's advocates, constitutional rights to free speech, independence of the press, etc. These will ensure that bad news will most likely "resurface" *ex-post* in a way that is hard to ignore, thus lowering the *ex-ante* return of investing in denial.

Similar results apply if the dissenter brings her message at an interim stage, after people have censored but before investments are made. For $s < s^*$ they should welcome the opportunity to correct course and return to reality. In practice, this can be hard to achieve: it may not be an equilibrium (case $\theta_L > 0$), or require full coordination (case $\theta_L < 0$). With payoff heterogeneity, dissenters' motives will also be suspect, making it hard to convince others. The conclusion is even starker if people value maintaining hope or dislike anxiety-sufficiently that $s > s^*$. In that case, bringing (back) the bad news about the state really being L will hurt everyone, leading to a *universal unwillingness to listen* and rejection –the curse of Cassandra. Yet free-speech guarantees and mechanisms encouraging dissent *remain desirable ex-ante,* because they avoid welfare losses in state H and, on average, save the organization or society from wasting resources on denial (including killing messengers). There is now a strong tension between ex-ante and ex-post incentives to tolerate dissenting speech, illustrations of which abound in corporations, bureaucracies, and polities.

3. Contagious market exuberance

"Why did the company's chief, who routinely warned of his rivals' lax lending practices well before the mortgage market cracked, ultimately allow Countrywide to ardently embrace those practices? According to... a former banking analyst and founder of a New York investment fund, 'The biggest self-inflicted wound here is they should have pulled back in '05 and '06 when you had these com-

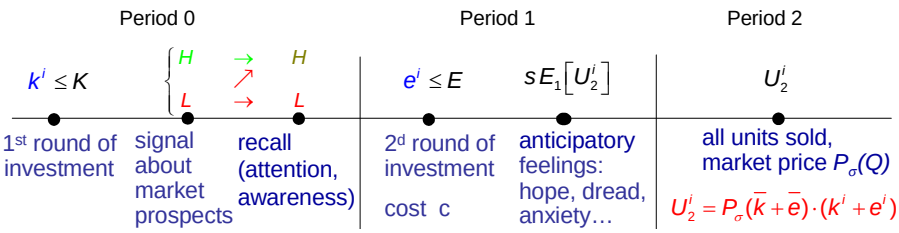

Figure 5: The market game

petitors doing all sorts of crazy things. Angelo [Mozilo] talked about the danger but somehow went for the market share gains anyway.'" (Morgenson and Fabrikant, 2007)

"I don't think it's a bubble, David M. Rubenstein of Carlyle Group told the Financial Times in December 2006. I think really what's happening now is that people are beginning to use a different investment technique, and this investment technique, private equity, adds real value." (BusinessWeek, 2007)

3.1. The dynamics of manias and crashes

I now extend the model to asset markets, adding an ex-ante investment stage and deriving final payoffs from equilibrium prices: see Figure 5. A large number (continuum) of firms or investors i can each produce $k^i \le K$ units of a good or asset (housing, office space, mortgage-backed security, internet startup) in period 0 and an additional $e^i \le E$ units in period 1, where K and E reflect capacity constraints or "time to build" technological limits. The cost of production in period 0 is set to 0 for simplicity, while in period 1 it is equal to c. All units are sold at $t = 2$, at which time the *expected* market price $P_\sigma(Q)$ will reflect total supply $Q \equiv \bar{k} + \bar{e} \in [0, K + E]$ as well as stochastic market conditions θ_σ, with $\sigma = H, L$ and $P'_\sigma(Q) < 0$. Between the two investment phases agents all observe the signal σ, then decide how to process it, with the same information structure and preferences as before.

To take recent examples, θ_H may correspond to a "new economy" in which high-tech startups will flourish and their prospects are best assessed using "new metrics"; to a permanent rise in housing values; or to any other positive and lasting shift in fundamentals. Conversely, θ_L would reflect an inevitable return to "old" economy and valuations; the presence of a bubble that will ultimately burst; or the unsustainability for many households of meeting future payments on their adjustable-rate mortgages, stated-income loans and other subprime

	Lehman Brothers	Bear Stearns	
Level 1	96 (35.6%)	39 (17.7%)	Trade in active markets with readily available prices
Level 2	152 (56.3%)	163 (74.1%)	"Mark to model"
Level 3	22 (8.2%)	18 (8.2%)	"Reflect management's best estimates of what market participants would use in pricing the assets"
Total ($ billions)	270	220	

Figure 6: Financial assets on balance sheet, 2d fiscal quarter of 2007. Source: Reilly (2007).

debt. Finding reasons to believe in H even as evidence of L accumulates corresponds to what Shiller (2005) terms "new-era thinking", and of which he relates many examples. I provide in this section the first analytical model of this phenomenon.[27]

The absence of an interim or futures market before date 2 is a version (chosen for simplicity) of the kind of "limits to arbitrage" commonly found in the finance literature. Specifically, I assume that: (i) goods produced in period 0 cannot be sold before period 2, for instance because they are still work-in-progress whose quality or market potential is not verifiable: startup company, unfinished residential development or office complex, new type of financial instrument; (ii) short sales are not feasible.

Limited liquidity and arbitrage possibilities are *empirically descriptive* of the types of markets which the model aims to analyze.[28] In the recent financial crisis, a dominant fraction of the assets held by major U.S. investment banks did not have an active trading market and objective price, but were instead valued according to the bank's own model and projections, or even according to management's "best estimates". Figure 6 shows the figures for Lehman Brothers and Bear Stearns, constructed from Reilly (2007).[29] Worldwide, the notional value

[27] As explained in the introduction, neither "rational bubbles" nor informational cascades involve any element of wishful thinking, distorted information processing or motivated rationalization. In both cases, all investors act exactly as an impartial, information-seeking statistician would advise (or allow) them to.

[28] Shiller (2003) cites several studies documenting the fact that short sales have never amounted to more than 2% of stocks, whether in number of shares or value. Gabaix et al. (2007) provide specific evidence of limits to arbitrage in the market for mortgage-backed securities.

[29] The share of Level 3 assets, whose valuations Reilly describes as "little more than management's guesses", was as high as 10% when Goldman Sachs and J.P. Morgan were included, and around 6% when Merrill Lynch was added. Concerning Level 2 assets, the major trading houses commonly used computer programs designed for "plain vanilla" loans to value novel and highly complex securities (Hansell, (2008)).

of outstanding Collateralized Debt Obligation (CDO) tranches stood in 2008 at about $2 trillion and that of Credit Default Swaps (CDS) around $50 trillion; and yet for most of them there was (and still is) no established, centralized marketplace where they could easily be traded. These are instead *very illiquid* ("buy and hold") and *hard-to-price* assets: originating in private deals, highly differentiated and exchanged only over-the-counter. In housing, finally, regional-index futures (Case-Shiller) are a very recent innovation and their market is still small and fairly illiquid.

Suppose (for now) that, ex-ante, the market is sufficiently profitable that everyone invests up to capacity at the start of period $0: k^j = \bar{k} = K$. Moreover, following (4), let

$$P_L(K) < \frac{c}{s+\delta} < \frac{c}{\delta} < qP_H(K+E) + (1-q)P_L(K+E). \tag{20}$$

It is thus a dominant strategy for an agent at $t = 1$ to invest the maximum $e^i = E$ if his posterior is no worse than the prior q, and to abstain if he is sure that the state is L.

Consider now the market subgame that unfolds when agents observe the signal L at the end of period 0. The optimality of first-stage investment $k^j = K$ (given expected profits in both states) is proved in the appendix and taken here as given, for expositional simplicity.

• *Realism.* If market participants acknowledge and properly respond to bad news ($\lambda^j \equiv 1$) they will not invest further at $t = 1$, so the price at $t = 2$ will be $P_L(K)$. For an individual investor i with stock k^i, the net effect of ignoring the signal is thus

$$U^i_{0,D} - U^i_{0,R} = -m + \delta\left[(\delta+s)P_L(K) - c\right]E + \delta sr(\lambda^i)\left[P_H(K+E) - P_L(K)\right](k^i+E). \tag{21}$$

The second term reflects the expected losses from producing at $t = 1$, while the last one represents the value of maintaining hope that the market is strong or will eventually recover, in which case total output will be $K+E$ and the price $P_H(K+E)$. Realism is an equilibrium if $U^i_{0,D} \leq U^i_{0,R}$ for $\lambda^i = 1$ and $k^i = K$, or

$$s \leq \frac{m/\delta + [c - \delta P_L(K)]\,E}{[P_H(K+E) - P_L(K)]\,(K+E) + P_L(K)E} \equiv \underline{s}(1). \tag{22}$$

• *Denial.* If the other participants remain bullish in spite of adverse signals, they will keep investing at $t = 1$, *causing the already weak market to crash:* at $t = 2$, the price will

fall to $P_L(K+E) < P_L(K)$. The net value of denial for investor i is now

$$U^i_{0,D} - U^i_{0,R} = -m + \delta\left[(\delta+s)P_L(K+E) - c\right]E + \delta s r(\lambda^i)\left[P_H(K+E) - P_L(K+E)\right](k^i+E). \tag{23}$$

In the second term, the expected losses from overinvestment are higher than when other participants are realists. Through this channel, which reflects the usual *substitutability* of investments in a market interaction, each individual's cost of delusion increases when others are deluded. On the other hand, the third term makes clear that the psychological value of denial is also greater, since acknowledging the bad state now requires *recognizing an even greater capital loss* on any preexisting holdings. This is again the MAD principle at work.

Denial is an equilibrium if $U^i_{0,D} \geq U^i_{0,R}$ for $\lambda^i = 0$ and $k^i = K$, or

$$s \geq \frac{m/\delta + \left[c - \delta P_L(K+E)\right]E}{q\left[P_H(K+E) - P_L(K+E)\right](K+E) + P_L(K+E)E} \equiv \bar{s}(0). \tag{24}$$

In such an equilibrium, each investor keeps optimistically accumulating assets that have in fact become "toxic", both to his own balance sheet and to the market at large.

When does other participants' exuberance make each individual more likely to also be exuberant? Intuitively, such contagion occurs when the substitutability effect, which bears on the *marginal* units E produced in period 1, is dominated by the capital-loss effect on the *outstanding position* K inherited from period 0. Formally, $\bar{s}(0) < \underline{s}(1)$ requires that K be large enough relative to E, though not so large as to preclude (22).

Proposition 6. ***(Market manias and crashes)*** *If*

$$P_H(K+E)\left(1 + E/K\right) < c/\delta < P_H(K+E), \tag{25}$$

there exists $q^* < 1$ such that, for all $q \in [q^*, 1]$, there is a non-empty interval for s in which both realism and blind "exuberance" in the face of adverse news are equilibria, provided m is not too large. Contagious exuberance leads to overinvestment, followed by a deep crash.

The model provides a microfounded and psychologically-based account of market groupthink, investment frenzies and ensuing crashes. It also identifies some key features of the markets that are prone to such cycles.

First, there must be a "story" about shifts in fundamentals that is minimally plausible a priori (q must not be too low): technology, demographics, globalization, etc. The key result is that investors's beliefs in the story can then quickly become resistant to any contrary evidence.[30] Second, when the new opportunity first appears (q rising above the threshold), there is an initial phase of investment buildup and rising price expectations.[31] Finally, the assets in question must involve both significant uncertainty and limited liquidity, as discussed earlier. These conditions are typical of assets tied to new technologies or financial instruments, whose potential will take a long time to be fully revealed.

The model's comparative statics also shed light on other puzzles. From (21)-(24), we have:

(a) *Escalating commitment* at the individual level: the more an agent has invested to date (k^i), the more likely he is to continue even in the face of bad news, thus displaying a form of the *sunk cost fallacy.*[32] Moreover, while k^i represents here an outstanding inventory or financial position, any other illiquid asset with market-dependent value, such as sector-specific human capital (e.g., in banking or finance), clearly has the same effect.[33]

(b) *Market momentum:* the larger the total market buildup ($k^{-i} = K$), the more likely is each agent to continue investing in spite of bad news, under a simple condition on the price sensitivity of demand. In a denial equilibrium the incentive to discount bad news stems from prospective capital losses proportional to $P_H(K+E) - P_L(K+E)$, which increases in K when $\partial^2 P/\partial Q \partial \theta > 0$.[34] This occurs for instance when good fundamentals correspond to a

[30] By contrast, in standard models of stochastic bubbles everyone realizes that they are trading a "hot potato" whose value does not reflect any fundamentals, must eventually collapse and can do so at any instant.

[31] In the interim period there is no objective market price, but all participants' "mark to model" or "best estimates" values remain at $qP_H(K+E) + (1-q)P_L(K+E)$, which reflects only the increased prior q, instead of falling the very low $P_L(K+E)$ actually warranted by the red flags which they are ignoring. Note also that the most economically important aspect of market manias is not price dynamics per se but the misallocation of resources, which is what the present analysis focuses on.

[32] This effect is closely related to the escalating commitment studied in Bénabou and Tirole (2007), but arises there from a somewhat different mechanism (self-signaling).

[33] Having an initial stake in the market raises an agent's propensity to engage in wishful thinking and invest in spite of red flags, but it is *not* a precondition. Equation (21) or (23) can be positive (for $\lambda^i = 0$) even with $k^i = 0$, especially for people with high sensitivity to anticipatory feelings, s^i. Moreover, the presence of participants with large k^j's, resulting in a high market buildup, will push even those with low (but positive) k^i toward delusion, through the "momentum " result below. (Similar externalities occur between agents with different s^i's). At the market level, finally, (24) can also hold with $K = 0$. With no overhang the equilibrium is unique, however: (22) and (24) become mutual opposites. Depending on s, a given asset market should then, unrealistically, never or always fall prey to exuberance. Multiplicity, by contrast, signifies a "fragility" to recurrent manias, and more generally the potential to greatly amplify small shocks.

[34] For simplicity I focus here on the benefit side of denial, leaving aside the cost. A higher K always raises

scarcity of some close substitute and market demand is concave: $P_\sigma(Q) = \mathcal{P}(Q + Z(\theta_\sigma))$, with $Z', \mathcal{P}', \mathcal{P}'' < 0$.[35] A greater market buildup K then tends to makes denial easier to sustain and realism more difficult, thus raising the likelihood of continued momentum.

This simple asset-market model could be extended in several ways. First, in a dynamic context, outstanding stocks will result stochastically from the combination of previous investment decisions and demand realizations. Second, one could relax the relatively strong form of "limits to arbitrage" imposed here through the assumption that trades occur only at $t = 2$ (no forward market). Such early trades could instead involve transactions costs, risk due to limited market liquidity or, for large positions, an adverse price impact.[36]

3.2. Regulators, politicians and other indirect stakeholders

The preceding analysis showed how an agent's propensity to respond to danger signals with a "suspension of disbelief" increases with his initial investment position k^i, market-correlated human capital or any other asset that cannot easily be sold off or hedged. Other, more indirect stakes have similar effects, both contributing to and feeding on the propagation of collective blindness (and ultimate losses) to broad parts of the economy.

Thus, if indicators point to a state of the world in which the housing sector is headed for a crash and the economy for a recession, all three major assets of households are at risk: their job, the value of their house and their pension –the latter especially is some of it is invested in company stock. The worse the potential downturn is made by other agents' feeding of the market frenzy, the greater is the incentive not to acknowledge these risks (dismiss or rationalize away the signal $\sigma = L$). And, as a result, the greater the likelihood that the household will itself contribute to the excessive buildup of debt, housing, or undiversified stock holdings.

Another set of key actors with "value at risk" are politicians and regulators, whose career and reputation will be badly damaged if the disaster scenario (state L, worsened by market

---

the latter, but if E is small relative to K this effect will be dominated. Similarly, in a realistic equilibrium, capital losses are proportional to $P_H(K+E) - P_L(K) = [P_H(K+E) - P_L(K+E)] + [P_L(K+E) - P_L(K)]$. The first term is the same as before, and it will dominate the second one if E is not too large relative to K.

[35] Another example is the linear market demand $Q(P, \theta) = \theta(a - bP)$, leading to $P = b^{-1}(a - Q/\theta)$.

[36] Trying to sell (or sell short) in period 1 could also be self-defeating, as it would reveal again to the market that the state is L, generating an immediate price collapse. For a model of how market thinness generates endogenous limits to arbitrage and delays in trade, see Rostek and Weretka (2008).

participants' manic overinvestment) occurs. This should normally make them try to dampen the market's enthusiasm, but if the buildup has proceeded far enough (high K) that large, economy-wide losses are are unavoidable in the bad state, they will also become "believers" in a rosy future or smooth landing. Consequently, they will fail to take the measures that could have limited (tough not avoided) the damage, and thus further enable the investment frenzy and subsequent crash.[37] Public officials and some academics may also have a more general ideological stake in (say) the virtues of unfettered financial markets: a severe crisis publicly proving such faith to be excessive would reduce the general credibility of laissez-faire arguments and increase demand for public regulation in other parts of the economy.

4. Other applications and extensions

4.1. Collective apathy and fatalism

The form of denial considered so far has been a collective "illusion of control" or overconfidence, leading an organization or market to persist in a costly course of action in spite of widely available evidence that it is doomed. The opposite case is collective apathy: rather than acknowledge a crisis that could be partly remedied through timely action, everyone pretends that things, though perhaps not great, "could be worse", and that "nothing can be done" to improve them anyway. One can think of an ethnic group subject to discrimination and threat by another one, but whose members pessimistically deem it useless to fight back, try to escape or otherwise improve their lot.[38] Another example is global-warning denial. A third one, examined below, is that of "tuning out" the distress of others.

To capture these ideas, I simply extend (1) to

$$U_2^i = \theta \left[\alpha e^i + (1-\alpha)e^{-i} - \kappa\right], \quad \text{where} \quad \kappa \gtrless 0. \tag{26}$$

• When $\kappa < \min\{1, \theta_H/\Delta\theta\}$, state H remains (conditional on $e = 1$) a more favorable state than L, and one can show that for κ below a certain threshold all the results of the

[37] Asked in a 2007 Congressional testimony whether he was "at all concerned... that if one of these huge institutions fails, it will have a horrendous impact on the national and global economy", former FED Chairman Alan Greenspan replied: "No, I'm not," "I believe that the general growth in large institutions have occurred in the context of an underlying structure of markets in which many of the larger risks are dramatically –I should say, fully– hedged" (Goodman (2008). For other instances of blindness to red flags and active information-avoidance by the FED and other regulators, see SEC (2008, 2009) and Appendix A.

[38] On minorities' "acquiescence" to a discriminatory system, see Cialdini (1984) and Hochschild (1996).

case $\kappa = 0$ carry over with little change. In particular, if $-\kappa > 0$ it plays a role very similar to an individual's outstanding market position k^i in the previous section.

• When $\kappa > \max\{1, \theta_H/\Delta\theta\}$, state H corresponds to a *crisis state:* action is called for, but even when carried out effectively ($e^j \equiv 1$) it will not suffice to offset the shock, leaving agents worse off than in state L. Intuition now suggests that an equilibrium in which agents respond appropriately to crises can coexist with one in which they systematically censor such signals and remain passive, even when they actually have individual "agency".[39]

Indeed, this problem is closely related to the original one, once recast in terms of the relative effectiveness of *inaction.* Formally, let $\tilde{\theta}$ take values $\tilde{\theta}_{\tilde{H}} \equiv -\theta_L$ in state $\tilde{H} \equiv L$ and $\tilde{\theta}_{\tilde{L}} \equiv -\theta_H < 0$ in state $\tilde{L} \equiv H$, with respective probabilities $\tilde{q} \equiv 1 - q$ and $1 - \tilde{q}$; similarly, let $\tilde{c} \equiv -c$. Using these transformed variables, it is then easy to obtain "parallels" to Propositions 2 to 5. In particular, condition (4) is replaced by

$$q\theta_H + (1-q)\,\theta_L < \frac{c}{\alpha\,(s+\delta)} < \frac{c}{\alpha\delta} < \theta_H, \tag{27}$$

and the equilibrium strategies and thresholds are obtained by replacing $\Delta\theta$ with $-\kappa\Delta\theta$ and $\theta_H,\ \theta_L,\ q$, and c with their "tilde" analogues.

Proposition 7. *Assume (27) and $\kappa > \max\{1, \theta_H/\Delta\theta\}$. All the results in Proposition 2 remain, but with denial ($\lambda < 1$) now occurring in state H only and leading to inaction. Facing up to crises and fatalistic inertia are both social equilibria if and only if $q\,(\kappa\Delta\theta) < (1-\alpha)\,\theta_H$.*

The left-hand side of this modified MAD condition reflects the action-independent gain from being in the no-crisis state, while the right-hand side measures the endogenous losses inflicted by all those who, denying that a crisis has occurred, *fail to act.*

• *Helping others or tuning out.* Studies of how people respond to the distress of others –victims of accidents, wars, natural disasters, famine, etc.– display two important puzzles. First, they show a greater willingness to help when the number of those perceived to be in need is small than when it is large. Slovic (2007) discusses many experiments documenting

[39] Furthermore, there is now no equilibrium in which agents censor the signal $\sigma = L$, just like when $\kappa = 0$ (or κ sufficiently below $\min\{1, \theta_H/\Delta\theta\}$ more generally) there is no equilibrium in which they censor $\sigma = H$. See Lemma 4 and the proof of Proposition 7 in the appendix, with $\Delta\gamma \equiv -\kappa\Delta\theta$.

such "psychic numbing" (lowered affective reactions and donations) in response to even small absolute increases in the size of the at-risk group. He further argues for the importance of this phenomenon in accounting for public inertia in the face of humanitarian disasters, poverty and genocide. A second regularity, common to most public-goods situations, is that people give and help more when they know or expect that others are doing so.[40]

The above results can help understand both phenomena. Let K be the number of people in need, or emphasized as being in need, and let θ be the severity of their situation. At cost c, each individual $i = 1, ... n$ can help up to a victims ($e^i = 1$), and he experiences an empathic disutility equal to the total amount of suffering,

$$U_2^i = -\theta \left[K - a\Sigma_{j=1}^n e^j \right]. \tag{28}$$

Note that this *does not assume* that people intrinsically undervalue "statistical lives" or actions that represent only "a drop in the ocean". Instead, *this will be a result.* Indeed, (28) corresponds to (26) with $\alpha = 1/n$, $\kappa = K/na$ and θ simply replaced by θna. Therefore, as K increases beyond a critical threshold:

(a) The loss in utility from acknowledging $\theta = \theta_H$ overtakes an individual's ability to remedy it, causing him to switch from helping to "tuning out" the problem altogether. Thus, he effectively censors from awareness and recall all painful evidence of the crisis: turning the page of the newspaper, switching the channel, rationalizing the situation as not so bad, etc.

(b) The level at which an individual switches from response to non-response depends on how many others he believes are helping or also tuning out: what matters to i is $K - a\Sigma_{j \neq i}^n e^j$. Hence, within some range of K, both *collective generosity* and *collective apathy* –what Slovic terms the "collapse of compassion"– are social equilibria, even though charitable giving involves, realistically, no increasing returns.

(c) Vivid, memorable images of the *intensity* of individual suffering θ (but *not* the number, K, which has the opposite effect) make the crisis more difficult to put "out of mind" and thus reduce the scope of apathy. In the multiplicity range, one small such example, widely

[40] The first phenomenon is distinct from (but combines with) the "identifiable victim effect". Small et al. (2007) thus found that donations to a specifically identified Malawian child facing the risk of starvation decreased by more than a half when information about the child was complemented with background statistics documenting the scale of food shortages in Africa. An alternative explanation for the second set of findings is one of social or personal norms; see Bénabou and Tirole (2006a).

publicized, can trigger a large equilibrium shift.

4.2. Alternative information preferences and technologies

To highlight general-equilbrium effects, the model used the simplest possible (linear) specification of individual utility from beliefs. On the cognitive side, it emphasized ex-post information processing –selective attention, interpretation, recall, in a word: *thinking*– rather than ex ante information acquisition or avoidance. The MAD principle is much more general, however, and robust to alternative ways of modeling the psychological motives and cognitive operations underlying individual belief distortion.[41] All it requires is that: (a) agents have some motive and scope to process signals or probabilities subjectively; (b) processing all information objectively implies recognizing that some other actors may not be doing so.

This idea provides, for instance, a template for generating "social cognition" results from ex-ante attitudes toward *risk.* Consider agents with a preference for late resolution of uncertainty (Kreps and Porteus (1978)), or some other value function concave in beliefs. If an agent's remaining ignorant of the state of the world leads him to increase the risks borne by others, this will push them toward also wanting to delay finding out about the future. Thus, if information avoidance or lack of costly attention generates adverse spillovers (now in the variance rather than mean of payoffs, but still without need for exogenous complementarities), it will be contagious. Conversely, if its risk spillovers are favorable, it will be self-dampening.

5. Conclusion

This paper has developed a general model of how wishful thinking and reality denial spread through organizations and markets. The underlying mechanism does not rely on built-in complementarities, agents' herding on a subset of private signals, or exogenous biases in inference. It is widely applicable, helping to explain corporate cultures characterized by dysfunctional groupthink or valuable group morale, why delusions flow down hierarchies, and the emergence of market manias sustained by "new-era" thinking, followed by deep crashes.

[41] In particular, the presence or lack of Bayesian sophistication is largely inessential (see footnote 16), as is the use of a binary state space (with a richer space, wishful thinking would take the form of a partitional coarsening of signals, as is standard in models of strategic communication).

In each of these applications, the institutional and market environment was kept very simple, so as to make clear the workings of the underlying "Mutually Assured Delusion" principle. Enriching these context-specific features of the model would be quite valuable and permit new applications. For hierarchical organizations in particular, richer payoff and information structures could be incorporated, along with greater heterogeneity of interests among agents. Potential applications include the spread of organizational corruption (e.g., Anand et al. (2005)), corporate politics (e.g. Zald and Berger (1998)) and organizational-design questions such as the optimal mix of agents, network structure and communication mechanisms (e.g. Calvó-Armengol et al. (2009)).

"Fantastic faith" and immunity to evidence are also clearly at work in political ideology. In Bénabou (2008) I embed the model into a political economy setting and analyze society-wide beliefs concerning the relative merits and proper scope of the state versus the market. A common principle is thus shown to help explain reality distortions in both organizational and political culture. A further application to politics could be the spread and persistence of conspiracy theories (e.g., "dependency theory", as described by North (1990)).

A somewhat different class of collective delusions are mass panics and hysterias. While the model generates not only hard-reality crashes but also episodes of excessive doubt and overcautiousness, the latter seem too mild to capture what goes on in a full-fledged panic.[42] Understanding the sources and transmission mechanisms that underlie delusional group pessimism, rather than optimism, is an interesting question for further research.

[42]Recall first that, when agents censor bad news, they never fully believe in the good state ($\sigma = H$) even when it actually occurs: they cannot avoid suspecting that there could have been danger signals which they (and others) looked away from. Second, investors who fear (perhaps from having been burned once) falling prey to the next wave of collective overoptimism will shy away from even positive expected-value investments (this occurs when condition (A.22) in the appendix is reversed).

Appendix

In the proofs given here, I maintain the text's focus on cognitive decisions in state L, fixing everyone's recall strategy in state H to $\lambda_H = 1$. In Appendix B (Lemmas 3 and 4), I show that this is not a binding restriction: with the payoffs (1) there is no equilibrium with $\lambda_H < 1$ and no profitable individual deviation to $\lambda_H^i < 1$ from an equilibrium with $\lambda_H = 1$.[43] These results, as well as Proposition 7, are proved using the more general specification

$$U_2^i \equiv \theta \left[\alpha e^i + (1-\alpha)e^{-i}\right] + \gamma, \tag{A.1}$$

where γ, like θ, is now also state-dependent and $\Delta\gamma \equiv \gamma_H - \gamma_L$ can be of either sign.

Proof of Proposition 1. (i) Let $\equiv \Psi(\lambda^i, s|\lambda^{-i})$ denote the right-hand side of (10). Since it is increasing in λ^i, agent i's optimal awareness strategy is uniquely determined as follows

(a) $\lambda^i = 1$ if $\Psi(1, s|\lambda^{-i}) \leq 0$. By (10), and noting that $\alpha\theta_L + \Delta\theta + (1-\alpha)\lambda^{-i}\theta_L \geq \min\{\Delta\theta, \theta_H\} > 0$, this means

$$s \leq \frac{m/\delta + c - \delta\alpha\theta_L}{\alpha\theta_L + \Delta\theta + (1-\alpha)\lambda^{-i}\theta_L} \equiv \underline{s}(\lambda^{-i}). \tag{A.2}$$

(b) $\lambda^i = 0$ if $\Psi(0, s|\lambda^{-i}) \geq 0$. By (10), and noting that $\alpha\theta_L + q\left[\Delta\theta + (1-\alpha)\lambda^{-i}\theta_L\right] \geq \min\{q\Delta\theta, q\theta_H + (1-q)\theta_L\} > \min\{q\Delta\theta, c/(s+\delta)\} > 0$, this means

$$s \geq \frac{m/\delta + c - \delta\alpha\theta_L}{\alpha\theta_L + q\left[\Delta\theta + (1-\alpha)\lambda^{-i}\theta_L\right]} \equiv \bar{s}(\lambda^{-i}). \tag{A.3}$$

Moreover, $\underline{s}(\lambda^{-i}) < \bar{s}(\lambda^{-i})$, since $\Delta\theta + (1-\alpha)\lambda^{-i}\theta_L \geq \Delta\theta + (1-\alpha)\lambda^{-i}\min\{\theta_L, 0\} \geq \Delta\theta + \min\{\theta_L, 0\} = \min\{\theta_H, \Delta\theta\} > 0$.

(c) $\lambda^i \in (0,1)$ is the unique solution to $\Psi(\lambda^i, s|\lambda^{-i}) = 0$ for $\Psi(0, s|\lambda^{-i}) < 0 < \Psi(1, s|\lambda^{-i})$, which corresponds to $\underline{s}(\lambda^{-i}) < s < \bar{s}(\lambda^{-i})$.

(ii) and (iii) follow from the monotonicity of Ψ in θ_L and α. Note that no assumption of symmetry in strategies was imposed (λ^{-i} could, a priori, be the mean of heterogenous recall rates); therefore, the only equilibria are the symmetric ones described in the proposition. ■

[43] Under the very weak condition that each agent encodes his own information (for future recall) in a cost-effective manner, which Lemma 3 shows can always be ensured. This is seen most clearly for $\lambda_H^i = \lambda_L^i = 0$, which is informationally equivalent to $\lambda_H^i = \lambda_L^i = 1$ but wastes m in each state.

Proof of Proposition 2. By Proposition 1, $\lambda = 1$ is an equilibrium when $\Psi(1, s|1) \leq 0$, or

$$s \leq \frac{m/\delta + c - \delta\alpha\theta_L}{\alpha\theta_L + \Delta\theta + (1-\alpha)\theta_L} = \frac{m/\delta + c - \delta\alpha\theta_L}{\theta_H} \equiv \underline{s}(1), \tag{A.4}$$

and $\lambda = 0$ is an equilibrium when $\Psi(0, s|0) \geq 0$, or

$$s \geq \frac{m/\delta + c - \delta\alpha\theta_L}{\alpha\theta_L + q\Delta\theta} \equiv \bar{s}(0). \tag{A.5}$$

Finally, $\lambda \in (0,1)$ is an equilibrium if and only if $\Psi(\lambda, s|\lambda) = 0$. Now, from (10) and (7),

$$\Psi(\lambda, s|\lambda) = -m/\delta - c + (\delta + s)\,\alpha\theta_L + sq\left(\frac{\Delta\theta + (1-\alpha)\lambda\theta_L}{q + (1-q)(1-\lambda)}\right). \tag{A.6}$$

This function is either increasing or decreasing in λ, depending on the sign of $(1-\alpha)\theta_L + (1-q)\,\Delta\theta$. One can also check, using (A.2)-(A.3), that the same expression governs the sign of $\underline{s}(1) - \bar{s}(0)$. The equilibrium set is therefore determined as follows:

(a) If (11) does not hold, $\Psi(\lambda, s|\lambda)$ is increasing, so $\Psi(0, s|0) < \Psi(1, s|1)$, or equivalently $\underline{s}(1) < \bar{s}(0)$ by (A.2)-(A.3). There is then a unique equilibrium, equal to $\lambda = 1$ if $\Psi(1, s|1) \leq 0$, interior if $\Psi(0, s|0) < 0 < \Psi(1, s|1)$, and equal to $\lambda = 0$ if $0 < \Psi(0, s|0)$.

(b) If (11) does hold, $\Psi(\lambda, s|\lambda)$ is decreasing, so $\Psi(1, s|1) < \Psi(0, s|0)$, or equivalently $\bar{s}(0) < \underline{s}(1)$ by (A.2)-(A.3). Then: (i) $\lambda = 1$ is the unique equilibrium for $\Psi(0, s|0) \leq 0$, meaning that $s \leq \bar{s}(0)$, while $\lambda = 0$ is the unique equilibrium for $\Psi(1, s|1) \geq 0$, meaning that $s \geq \underline{s}(1)$; for $\Psi(1, s|1) < 0 < \Psi(0, s|0)$, or $\bar{s}(0) < s < \underline{s}(1)$, both $\lambda = 1$ and $\lambda = 0$ are equilibria, together with the unique solution to $\Psi(\lambda, s|\lambda) = 0$, which is interior. ■

Proof of Proposition 3. Following the same steps as in the symmetric case and denoting Λ^{-i} the vector of other agent's strategies, it is easy to show that

$$\underline{s}^i(\Lambda^{-i}) \equiv \frac{m^i/\delta^i + c^i - \delta^i\left(a_L^{ii} - b_L^{ii}\right)}{\Sigma_{j=1}^n\left(a_H^{ji} - a_L^{ji}\right) + \Sigma_{j\neq i}\,\lambda^j\left(a_L^{ji} - b_L^{ji}\right) + a_L^{ii} - b_L^{ii}}, \tag{A.7}$$

$$\bar{s}^i(\Lambda^{-i}) \equiv \frac{m^i/\delta^i + c^i - \delta^i\left(a_L^{ii} - b_L^{ii}\right)}{q\left[\Sigma_{j=1}^n\left(a_H^{ji} - a_L^{ji}\right) + \Sigma_{j\neq i}\,\lambda^j\left(a_L^{ji} - b_L^{ji}\right)\right] + a_L^{ii} - b_L^{ii}}. \tag{A.8}$$

Setting $\lambda^j \equiv 1$ in the first equation and $\lambda^j \equiv 0$ in the second yields the result. ■

Proof of Proposition 6. Assume or now that at $t = 0$, everyone else invests $k^{-i} = K$. Since investing (respectively, abstaining) at $t = 1$ is a dominant strategy given posterior

$\mu^j = r(\lambda^j) \geq q$ (respectively, $\mu^j = 0$), the price in state L will be $P_L(K + (1 - \lambda^{-i})E)$ and the date-0 expected utilities of realism and denial equal to

$$\text{(A.9)} \qquad U_{L,R}(\lambda^i, \lambda^{-i}; k^i)/\delta = (\delta + s)P_L(K + (1 - \lambda^{-i})E)k^i,$$

$$\begin{aligned}\text{(A.10)} \qquad U_{L,D}(\lambda^i, \lambda^{-i}; k^i_,)/\delta &= -m/\delta + (\delta + s)P_L(K + (1 - \lambda^{-i})E)(k^i + E) - cE \\ &\quad + sr(\lambda^i)\left[P_H(K + E) - P_L(K + (1 - \lambda^{-i})E)\right](k^i + E).\end{aligned}$$

The net incentive for denial, $\Delta U_L \equiv U_{L,D} - U_{L,R}$, is thus given by

$$\begin{aligned}\text{(A.11)} \qquad [\Delta U_L(\lambda^i, \lambda^{-i}; \bar{k};) + m]/\delta &= \left[(\delta + s)P_L(K + (1 - \lambda^{-i})E) - c\right]E, \\ &\quad + sr(\lambda^i)\left[P_H(K + E) - P_L(K + (1 - \lambda^{-i})E)\right](k^i + E).\end{aligned}$$

Setting $r(\lambda^i) = 1$, realism is a (personal-equilbrium) best response to λ^{-i} for an agent entering period 1 with stock k^i if

$$\begin{aligned}\text{(A.12)} \qquad m/\delta &\geq \left[(\delta + s)P_L(K + (1 - \lambda^{-i})E) - c\right]E \\ &\quad + s\left[P_H(K + E) - P_L(K + (1 - \lambda^{-i})E)\right](k^i + E).\end{aligned}$$

Conversely, denial ($r(\lambda^i) = q$) is a (personal-equilibrium) best response for i if

$$\begin{aligned}\text{(A.13)} \qquad m/\delta &\leq \left[(\delta + s)P_L(K + (1 - \lambda^{-i})E) - c\right]E \\ &\quad + sq\left[P_H(K + E) - P_L(K + (1 - \lambda^{-i})E)\right](k^i + E).\end{aligned}$$

For given k^i and λ^{-i}, these two conditions are mutually exclusive. When neither holds, there is a unique $\lambda^i \in (0, 1)$ that equates ΔU_L to zero, defining a mixed-strategy (personal equilibrium) best-response. The next step is to solve for (symmetric) social equilibria.

1. Realism. From (A.12), $\lambda^i = \lambda^{-i} = 1$ is an equilibrium in cognitive strategies if

$$\text{(A.14)} \qquad [(\delta + s)P_L(K) - c]E + s\left[P_H(K + E) - P_L(K)\right](k^i + E) \leq m/\delta.$$

This condition holds for all $k^i \leq K$ if and only if

$$\text{(A.15)} \qquad s \leq \frac{m/\delta + [c - \delta P_L(K)]E}{\left[P_H(K + E) - P_L(K)\right](K + E) + P_L(K)E} \equiv \underline{s}(1; K).$$

Moving back to the start of period 0, one now verifies that it is indeed an equilibrium for everyone to invest $k^i = K$. Since agents will respond to market signals $\sigma = H, L$, the expected

price is $qP_H(K+E)+(1-q)P_L(K)>0$, whereas the cost of period-0 production is 0 (more generally, sufficiently small). Thus, it is optimal to produce to capacity.

2. Denial From (A.13), $\lambda^i=\lambda^{-i}=0$ is a cognitive equilibrium if

$$[(\delta+s)P_L(K+E)-c]\,E+sq\,[P_H(K+E)-P_L(K+E)]\,(k^i+E)\geq m/\delta. \tag{A.16}$$

This condition holds for $k^i=K$ if

$$s>\frac{m/\delta+[c-\delta P_L(K+E)]\,E}{q\,[P_H(K+E)-P_L(K+E)]\,(K+E)+P_L(K+E)E}\equiv\bar{s}(0;q,K). \tag{A.17}$$

An agent with low k^i, however, has less incentive to engage in denial. In particular, for $s<\underline{s}(1;K)$, (A.14) for $k^i=0$ precludes (A.16) from holding at $k^i=0$. Let $\bar{k}(s,q)$ therefore denote the unique solution in k^i to the linear equation

$$[(\delta+s)P_L(K+E)-c]\,E+sq\,[P_H(K+E)-P_L(K+E)]\,(k^i+E)=m/\delta. \tag{A.18}$$

Subtracting the equality obtained by evaluating (A.16) at $s=\bar{s}(0;q,K)$ yields

$$\begin{aligned} & sq\,[P_H(K+E)-P_L(K+E)]\,(K-\bar{k}) \\ = \;& (s-\bar{s})P_L(K+E)E+(s-\bar{s})q\,[P_H(K+E)-P_L(K+E)]\,(K+E), \end{aligned}$$

where the arguments are dropped from $\bar{k}$ and $\bar{s}$ when no confusion results. Thus,

$$K-\bar{k}=\frac{s-\bar{s}}{s}\times\left(\frac{qP_H(K+E)+(1-q)P_L(K+E)}{q\,[P_H(K+E)-P_L(K+E)]}E+K\right)>\frac{s-\bar{s}}{s}\times(K+E)\,. \tag{A.19}$$

Note that $\bar{k}\leq K$ (and is thus feasible) if and only if $s\geq\bar{s}$. One can now examine the optimal choice of k^i at $t=0$, which will be either $k^i=K$ or some $k^i\leq\bar{k}$.

(a) For $k^i>\bar{k}(s,q)$, (A.18) implies that denial is the unique best response to $\lambda^{-i}=0$, leading agent i to produce $e^i=E$ in both states at $t=1$. These units and the initial k^i will be sold at the expected price $\bar{P}_q(K+E)\equiv qP_H(K+E)+(1-q)P_L(K+E)>0$. Therefore, producing K in period 0 is optimal among all levels $k^i>\bar{k}(s,q)$, and yields ex-ante utility

$$U_D(0,K,K)/\delta=(\delta+s)\bar{P}_q(K+E)(K+E)-cE-(1-q)m/\delta. \tag{A.20}$$

(b) For $k^i\leq\bar{k}(q;s)$, on the other hand, agent i's continuation (personal-equilibrium)

strategy is some $\lambda^i = \lambda(k^i) \geq 0$: in state L he weakly prefers to be a realist, achieving

$$\begin{aligned}(\text{A.21}) \quad U(\lambda^i, 0, k^i_{,} K)/\delta &= (\delta+s)\bar{P}_q(K+E)\left(k^i+E\right) - cE \\ &\quad - (1-q)\left\{\left(1-\lambda^i\right) m/\delta - \lambda^i\left[c-(\delta+s)P_L(K+E)\right]E\right\}.\end{aligned}$$

The agent prefers $k^i = K$ to any $k^i \leq \bar{k}(q;s)$ if $U_D(0,K,K) > U(\lambda^i, 0, k^i_{,} K)$, or

$$(\text{A.22}) \qquad (\delta+s)\bar{P}_q(K+E)(K-k^i) > (1-q)\,\lambda^i\left\{m/\delta + \left[c-(\delta+s)P_L(K+E)\right]E\right\}.$$

Using (A.19) and $\lambda^i \leq 1$, it suffices that

$$(\text{A.23}) \qquad \left(\frac{s-\bar{s}(0;q,K)}{s}\right)\left(\frac{\bar{P}_q(K+E)\,(K+E)}{1-q}\right) \geq \frac{m}{\delta\,(\delta+s)} + \left(\frac{c}{\delta+s} - P_L(K+E)\right)E.$$

Since $\bar{P}_q(K+E)$ tends to $P_H(K+E)$ as q tends to 1, (A.23) will hold for q close enough to 1, provided $s - \bar{s}(0;q,K)$ remains bounded away from 0. Lemmas 1 and 2 (in online Appendix B) formalize this idea, showing that there exist a threshold $q^*(K) < 1$ and a nonempty interval $S^*(K)$ such that, for all $q > q^*(K)$: $S^*(K) \subset (\bar{s}(0;q,K), \underline{s}(1;K))$ and (A.23) holds for all $s \in S^*(K)$. Consequently, when $q > q^*(K)$ both $(k^i = K, \lambda^i = 1)$ and $(k^i = K, \lambda^i = 0)$ are equilibria of the two-stage market game, for any $s \in S^*(K)$. Indeed, we showed that: (i) for $s < \underline{s}(1;K)$, when others play $(k^{-i} = K,\ \lambda^{-i} = 1)$ agent i finds it optimal to also invest $k^i = K$ and then be a realist; (ii) for $s > \bar{s}(0;q,K)$, when others play $(k^{-i} = K,\ \lambda^{-i} = 0)$ he finds it optimal to invest K in period 0 even though he knows that this will cause him to engage in denial if state L occurs. ■

REFERENCES

Akerlof, G., and W. Dickens (1982) "The Economic Consequences of Cognitive Dissonance," *American Economic Review,* 72, 307-19.

Anand, V., Ashforth, B. and J. Mahendra (2005) "Business as Usual: The Acceptance and Perpetuation of Corruption in Organizations," *Academy of Management Executive,* 19, 9-23.

Banerjee, A.(1992) "A Simple Model of Herd Behavior," *Quarterly Journal of Economics,* 107(3), 797-817.

Battaglini, M., Bénabou, R., and J. Tirole (2005) "Self-Control in Peer Groups," *Journal of Economic Theory,* 123, 105–34.

Bénabou, R. (2008) "Ideology," *Journal of the European Economic Association,* 6(2), 321–52.

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

Bénabou, R. and J. Tirole (2004) "Willpower and Personal Rules," *Journal of Political Economy,* 112, 848–887.

Bénabou, R. and J. Tirole (2006a) "Incentives and Prosocial Behavior," *American Economic Review,* 96(5), December , 1652-78.

Bénabou, R. and J. Tirole (2006b) "Belief in a Just World and Redistributive Politics," *Quarterly Journal of Economics,* 121(2), May, 699-746.

Bénabou, R. and J. Tirole (2007) "Identity, Dignity and Taboos: Beliefs as Assets," CEPR Discussion Paper no. 6123, February.

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

Bikhchandani, S. Hirshleifer, D., and I. Welch (1992) "A Theory of Fads, Fashion, Custom, and Cultural Change as Informational Cascades," *Journal of Political Economy,* 100(5), 992-1026.

Brunnermeier, M. and J. Parker (2005) "Optimal Expectations," *American Economic Review,* 90, 1092-118. .

Brunnermeier, M., Gollier, C. and J. Parker (2007) "Optimal Beliefs, Asset Prices, and the Preference for Skewed Returns," *American Economic Review,* 97(2), 159-65.

Calvó-Armengol, A., de Martí, J. and A. Prat (2009) "Endogenous Communication in Complex Organization," LSE mimeo, March.

Camerer, C. and U. Malmendier (2007) "Behavioral Economics of Organizations," in *Behavioral Economics and Its Applications,* P. Diamond and H. Vartiainen (eds.), Princeton University Press.

Caplin, A. and J. Leahy (1994) "Business as Usual, Market Crashes, and Wisdom After the Fact," *American Economic Review,* 84(3), 548-65.

Caplin, A. and J. Leahy (2001) "Psychological Expected Utility Theory and Anticipatory Feelings," *Quarterly Journal of Economics,* 116, 55–79.

Chamley, C. and D. Gale (1994) "Information Revelation and Strategic Delay in a Model of Investment," *Econometrica,* 62(5), 1065-85.

Cialdini, R. (1984) *Influence: The Psychology of Persuasion.* HarperCollins Publishers.

Cohan, J. (2002) " 'I Didn't Know' and 'I Was Only Doing My Job': Has Corporate Governance Careened Out of Control? A Case Study of Enron's Information Myopia". *Journal of Business Ethics,* 40, 275-99.

Columbia Accident Investigation Board (2003) *CIAB Final Report,* especially Chapters 6, 7 and 8. Available at http://caib.nasa.gov/.

Compte, O. and Postlewaite, A. (2004) "Confidence-Enhanced Performance," *American Economic Review,* 94(5), 1536-1557.

Dessi, R. (2008) "Collective Memory, Cultural Transmission and Investments," *American Economic Review,* 98(1), 534-560.

Di Tella, R., Galiani, S., and E. Schargrodsky, (2007) "The Formation of Beliefs: Evidence from the Allocation of Land Titles to Squatters," *Quarterly Journal of Economics,* 122(1), 209-41.

Eichennwald, K. (2005) *Conspiracy of Fools: A True Story. New York, NY:* BroadwayBooks.

Eliaz, K. and R. Spiegler (2006) "Can Anticipatory Feelings Explain Anomalous Choices of Information Sources?" *Games and Economic Behavior*, 56 (1), 87-104.

Eyster, E. and Rabin, M. (2009) "Rational and Naive Herding", LSE mimeo, June.

Fang, H., and Moscarini, G. (2005) "Morale Hazard," *Journal of Monetary Economics,* 52(4), 749-78.

Gabaix, X., Krishnamurthy, A. and O. Vigneron (2007) "Limits of Arbitrage: Theory and Evidence from the Mortgage-Backed Securities Market", *Journal of Finance,* 62(2), 557-595,

Gervais, S. and Goldtsein, I. (2007) "The Positive Effects of Self-Biased Perceptions in

Teams," *Review of Finance,* 11(3), 453-96.

Goodman, P. (2008) "The Reckoning: Taking Hard New Look at a Greenspan Legacy," *The New York Times,* October 8.

Hansell, S. (2008) "How Wall Street Lied to Its Computers," *The New York Times,* September 18.

Haslam, A. (2004). *Psychology in Organizations: The Social Identity Approach* (2nd ed.). London, UK & Thousand Oaks, CA: Sage.

Hermalin, B. (1998) "An Economic Theory of Leadership: Leading by Example," *The American Economic Review,* 88(5), 1188-206.

Hersh, S. (2004) *Chain of Command.* New York, NY: HarperCollins Publishers.

Huseman, R. and R. Driver (1979) "Groupthink: Implications for Small Group Decision Making in Business," in *Readings in Organizational Behavior: Dimensions of Management Action,* R. Richard Huseman and Archie Carral, eds.. Boston, MA: Allyn and Bacon.

Isikoff, M. and D. Corn (2007) *Hubris.* New York, NY: Three Rivers Press.

Janis, I. (1972) *Victims of Groupthink: Psychological Studies of Policy Decisions and Fiascoes.* Boston, MA: Houghton Mifflin Company.

Karlsson, N., Loewenstein, G. and D. Seppi (2005) "The 'Ostrich Effect': Selective Attention to Information about Investments," Carnegie Mellon University mimeo, May.

Kindleberger, C. and R. Aliber (2005) *Manias, Panics, and Crashes: A History of Financial Crises.* Hoboken, NJ: John Wiley and Sons.

Köszegi, B. (2006) "Emotional Agency," *Quarterly Journal of Economics,* 21(1), 121-56.

Kreps, D. and Porteus, E. (1978), "Temporal Resolution of Uncertainty and Dynamic Choice Theory," *Econometrica,* 46(1), 185–200.

Kuran, T. (1993) "The Unthinkable and the Unthought," *Rationality and Society,* 5, 473-505.

Landier, A. (2000 "Wishful Thinking: A Model of Optimal Reality Denial," MIT mimeo.

Landier, A., Sraer, D. and D. Thesmar (2009) "Optimal Dissent in Organizations," *Review of Economic Studies,* 76, 761-794.

Mackay, C. (1980) *Extraordinary Popular Delusions and the Madness of Crowds.* New York, NY: Three Rivers Press.

Malmendier, U. and G. Tate (2005) "CEO Overconfidence and Corporate Investment," *Journal of Finance,* 60 (6), 2661-700.

Malmendier, U. and G. Tate (2008) "Who Makes Acquisitions? CEO Overconfidence and the Market's Reaction," *Journal of Financial Economics,* forthcoming.

Morgenson, G. and G. Fabrikant (2007) "Countrywide's Chief Salesman and Defender," *The New York Times,* November 2007.

Norris, F. (2008) "Color-Blind Merrill in a Sea of Red Flags." *New York Times,* May 16.

North, D. (1990) *Institutions, Institutional Change and Economic Performance.* Cambridge, UK: Cambridge University Press.

Pearlstein, S. (2006) "Years of Self-Deception Killed Enron and Lay," *The Washington Post,* July 8.

Prendergast, C. (1993) "A Theory of 'Yes Men'," *American Economic Review,* 83(4), 757-70.

Reilly, D. (2007) "Marking Down Wall Street." *The Wall Street Journal,* September 14, C1.

Rogers Commission (1986). *Report of the Presidential Commission on the Space Shuttle Challenger Accident.* http://history.nasa.gov/rogersrep/genindex.htm.

Rostek, M. and M. Weretka (2008) "Dynamic Thin Markets," University of Madison-Wisconsin mimeo, December.

Rotemberg, J. and G. Saloner (2000) "Visionaries, Managers, and Strategic Direction," *Rand Journal of Economics* 31, Winter, 693-716.

Samuelson, R. (2001) "Enron's Creative Obscurity." *The Washington Post,* December 19.

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.

Scheinkman, J. and W. Xiong (2003) "Overconfidence and Speculative Bubbles," *Journal of Political Economy* 111, 1183-219.

Schrand, C. and S. Zechman (2008) "Executive Overconfidence and the Slippery Slope to Fraud," Wharton School mimeo, University of Pennsylvania, December.

Securities and Exchange Commission (2008) *SEC's Oversight of Bears Stearns and Related Entities: Consolidated Supervised Entity Program.* Inspector General's Report, Office of Audits, Report No. 446-. September 25, viii-ix. Available at http://www.sec-oig.gov.

Securities and Exchange Commission (2009) *Investigation of Failure of the SEC To Uncover Bernard Madoff's Ponzi Scheme.* Office of Investigations. Case No. OIG-509, August 31. Available at http://www.sec.gov/news/studies/2009/oig-509.pdf.

Shiller, R. (2003) "From Efficient Markets Theory to Behavioral Finance," *Journal of Economic Perspectives,* 17(1), 83-104

Shiller, R. (2005) *Irrational Exuberance.* Second Edition, Princeton University Press.

Sims, R. (1992) "Linking Groupthink to Unethical Behaviors in Organizations," *Journal of Business Ethics,* 11, 651-62.

Slovic, P. (2007) "If I Look at the Mass I will Never Act: Psychic Numbing and Genocide," *Judgment and Decision-Making,* 2(2), 79-95.

Small, D., Loewenstein, G. and Slovic, P. (2007) "Sympathy and Callousness: The Impact of Deliberative Thought on Donations to Identifiable and Statistical Victims," *Organizational Behavior and Human Decision Processes,* 143–53.

Suskind, R. (2004) "Without a Doubt," *The New York Times,* October 17.

Tenbrunsel, A. and D. Messick (2004) "Ethical Fading: The Role of Self-Deception in Unethical Behavior," *Social Justice Research,* 17(2), 223-62.

Van den Steen, E. (2005) "On the Origins of Shared Beliefs (and Corporate Culture)," MIT Sloan School Working Paper, August.

Zald, M. and M. Berger (1978) "Social Movements in Organizations: Coup d'Etat, Insurgency, and Mass Movements," *The American Journal of Sociology,* 83(4), 823-61.

SUPPLEMENTARY MATERIAL
Online Appendix A: Patterns of Denial

This appendix highlights certain patterns (in both words and deeds) that recur across most instances of organizational and market meltdown, from the Space Shuttle disasters to the recent financial crisis.[44]

1. Preposterous probabilities. Feynman's simple reasoning cited in the introduction makes clear that NASA management's risk estimates –one thousand times lower than those of their own engineers– made no statistical sense. The housing-related bubble and buildup to the current financial crisis abound in even more extreme statements of confidence –nothing short of probability one. In an August 2007 conference with analysts, Joseph Cassano, head of A.I.G. Financial Services, asserted

"It is hard for us, without being flippant, to even see a scenario within any kind of realm of reason that would see us losing one dollar in any of those transactions...".[45]

As late as 2008, in a meeting with investors,

"Lehman's chief financial officer, Erin Callan,... exuded confidence... With firms like Citigroup and Merrill raising capital, an investor asked, why wasn't Lehman following suit? Glaring at her questioner, she said that Lehman didn't need more money at the time –after all, it had yet to post a loss during the credit crisis. The company had industry veterans in the executive suite who had perfected the science of risk management, she said. "This company's leadership has been here so long that they know the strengths and weaknesses... We know when we need to be worried, and when we don't." (Anderson and Duhig (2008))

Are such statements by top executives only cynical attempts to deceive investors and analysts about the quality of their balance sheet? While there is surely an element of moral hazard, this explanation falls short on several counts. First, absurd claims of zero risk in highly turbulent times are simply not credible, and thus more likely to be read as negative signals about the executive's grasp of reality than reassurance about fundamentals. In fact,

[44] In what follows, all the quotes concerning NASA come from The Rogers Commission Report (1986) and the Columbia Accident Investigation Board Final Report (2003).

[45] Cited in Morgenson (2008). Not coincidentally, this is the London unit (which he founded) that sank the company after selling over $500 billion in credit default swaps that could not be covered.

they typically do nothing to bolster a company's share price, credit rating or prevent a run (see Sorkin (2008) for many examples).

Second, knowingly deceiving investors often leads to criminal prosecution and prison, as well as ruinous civil lawsuits and loss of reputation. A key aspect of self-delusion in such cases involves the expectation of "getting away" with fraud and cover-up, rather than ultimately sharing the fate of predecessors at Drexel Burnham Lambert, Enron, Worldcom, and many others.[46] Even abstracting from legal liability, selective blindness and collective rationalizations about the unethical nature of an organization's practices are key elements in the process that leads otherwise respectable citizens to take part in those practices (e.g., Sims (1992), Cohan (2002), Tenbrunsel and Messick (2004), Anand et al. (2005), Schrand and Zechman (2008)).

Third, identical claims of zero risk are made in settings where no large financial gain is involved and the downside can be truly catastrophic –as with NASA mission managers and financial regulators. Former FED Chairman Alan Greenspan's certainty that the new risks taken on by financial institutions were "dramatically –I should say, fully hedged" thus turned to "shocked disbelief" when the disaster scenario materialized a few months later.

2. New paradigms: this time is different, we are smarter and have better tools. Every case also displays the typical pattern of hubris, based on claims of superior talent or human capital. For A.I.G.'s Joseph Cassano, losses being simply unimaginable,

"The question for us is, where in the capital markets can we gain the best opportunity, the best execution for the business acumen that sits in our shop?".

What Feynman termed "fantastic faith in the machinery" is also often vested in computer models and statistical data. Subprime lenders and the banks purchasing the derived CDO's could thus rely on "a wealth of information we didn't have before" (Countrywide), fed to sophisticated computer programs:

" 'It's like having a secret sauce; everyone had their own best formulas," says Edward N. Jones, CEO of ARC Systems, which sold [underwriting and risk-pricing] technology to HSBC... and many

[46] In 2007 alone the FBI made over 400 arrests in subprime-related cases (including top fund managers at Lehman Brothers) and had ongoing criminal investigations into 26 major financial companies including Countrywide Financial, A.I.G., Lehman Brothers, Fannie Mae and Freddie Mac. These companies and their top executives (e.g., most of those cited in this appendix) are also being sued by several State attorney generals, in addition to countless shareholders groups, investors and borrowers.

of their rivals." (BusinessWeek (2007))

Closely related is the argument that previous rules of accounting, risk management or economics no longer apply, due to some radical shift in fundamentals. Shiller (2005) documents how such "new era thinking", variously linked to railroads, electricity, internet, demography or deregulation, was involved in nearly all historical episodes of financial bubbles and manias. Section 3 mentioned its latest incarnation –private equity as "a different investment technique... [that] adds real value." One can also see it at work in government:

"The [senior White House] aide said that guys like me were "in what we call the reality-based community," which he defined as people who "believe that solutions emerge from your judicious study of discernible reality." I nodded and murmured something about enlightenment principles and empiricism. He cut me off. "That's not the way the world really works anymore," he continued." We're an empire now, and when we act, we create our own reality. And while you're studying that reality – judiciously, as you will – we'll act again, creating other new realities, which you can study too, and that's how things will sort out." (Suskind (2004))

3. Escalation, failure to diversify, divest or hedge. Wishful beliefs show up not only in words but also in deeds. Enron's CEO Ken Lay resisted selling his shares throughout the long downfall, pledging other assets to meet collateral requirements, even buying stock back later on and ending up ruined well before his legal troubles began (Eichenwald (2005), Pearlstein (2006)). The company's employees, whose pension portfolios had on average 58% in Enron stock, could have moved out at nearly any point, but most never did (Samuelson (2001)). At Bears Stearns, 30% of the stock was held until the last day by employees – with presumably easy access to diversification and hedging instruments– who thus lost their capital together with their job. CEO James Cayne alone owned an unusually high 6% and went from billionaire to small millionaire in the process (spending most of the intervening months away playing golf and bridge). The pattern is similar at Lehman Brothers and other financial institutions.

Without looking to such extremes, Malmendier and Tate (2005, 2008) document many CEO's tendency to delay exercising their stock options and how this measure of overconfidence is a predictor of overinvestment. Studying individual investors, finally, Karlsson, Loewenstein and Seppi (2006) find that many more go online to check the value of their

portfolios on days when the market is up than when it is down.

Some of the most interesting evidence comes from cases in which an official inquiry or trial was conducted following a public- or private-sector disaster. Extensive records of meeting notes, memos, emails and sworn depositions reveal how key participants behaved, in particular with respect to information.

4. Information avoidance, repainting red flags green and overriding alarms. The most literal case of willful blindness occurred after the Columbia mission sustained a large foam strike to its wing's thermal shield:

"At every juncture of [the mission], the Shuttle Program's structure and processes, and therefore the managers in charge, resisted new information. Early in the mission, it became clear that the Program was not going to authorize imaging of [damage to] the Orbiter because, in the Program's opinion, images were not needed. Overwhelming evidence indicates that Program leaders decided the foam strike was merely a maintenance problem long before any analysis had begun."

Similar "head-in the sand" behavior was extensively documented at the Securities and Exchange Commission, even before its decade-long ignorance of Bernard Madoff's giant Ponzi scheme was revealed. The Inspector General's Report (S.E.C. (2008)) thus states:

"The audit found that [the Division of] Trading and Markets became aware of numerous potential red flags prior to Bear Stearns' collapse, regarding its concentration of mortgage securities, high leverage, shortcomings of risk management in mortgage-backed securities and lack of compliance with the spirit of Basel II standards, but did not take actions to limit these risk factors."

Instead, as reported in Labaton (2008), "the commission assigned [only] seven people to examine [the major investment banks] –which last year controlled... combined assets of $4 trillion. Since March 2007, the office has not had a director. And as of last month, the office had not completed a single inspection since it was reshuffled by Mr. Cox [the SEC chairman] more than a year and a half ago."

Similarly, at the FED...

"Edward M. Gramlich, a Federal Reserve governor... warned nearly seven years ago that a fast-growing new breed of lenders was luring many people into risky mortgages they could not afford. But when Mr. Gramlich privately urged Fed examiners to investigate mortgage lenders affiliated with national banks, he was rebuffed by Alan Greenspan... Mr. Greenspan and other

Fed officials repeatedly dismissed warnings about a speculative bubble in housing prices... The Fed was hardly alone in not pressing to clean up the mortgage industry. When states like Georgia and North Carolina started to pass tougher laws against abusive lending practices, the Office of the Comptroller of the Currency successfully prohibited them from investigating local subsidiaries of nationally chartered banks." (Morgenson and Fabrikant (2007))

... and the Treasury:

"In 1997, the Commodity Futures Trading Commission,... led by a lawyer named Brooksley E. Born... was concerned that unfettered, opaque trading could "threaten our regulated markets or, indeed, our economy without any federal agency knowing about it," she said in Congressional testimony. She called for greater disclosure of trades and reserves to cushion against losses. Ms. Born's views incited fierce opposition from Mr. Greenspan and Robert E. Rubin, the Treasury secretary then. Treasury lawyers concluded that merely discussing new rules threatened the derivatives market... In the fall of 1998, the hedge fund Long Term Capital Management nearly collapsed, dragged down by disastrous bets on, among other things, derivatives. Despite that event, Congress froze the Commission's regulatory authority for six months. The following year, Ms. Born departed. In November 1999, senior regulators –including Mr. Greenspan and Mr. Rubin– recommended that Congress permanently strip the C.F.T.C. of regulatory authority over derivatives." (Goodman (2008))

To avoid having to override alarms systems, it is sometimes simplest to turn them off from the start:

"The Commission was surprised to realize after many hours of testimony that NASA's safety staff was never mentioned... No one thought to invite a safety representative or a reliability and quality assurance engineer to the [prelaunch] January 27, 1986, teleconference between Marshall [Space Center] and Thiokol. Similarly, there was no representative of safety on the Mission Management Team that made key decisions during the countdown on January 28, 1986. The Commission is concerned about the symptoms that it sees."

Similarly, at Fannie Mae:

"Between 2005 and 2007, the company's acquisitions of mortgages with down payments of less than 10% almost tripled... For two years, Mr. Mudd operated without a permanent chief risk officer to guard against unhealthy hazards. When Enrico Dallavecchia was hired for that position

in 2006, he told Mr. Mudd that the company should be charging more to handle risky loans. In the following months to come, Mr. Dallavecchia warned that some markets were becoming overheated and argued that a housing bubble had formed... But many of the warnings were rebuffed... Mr. Dallavecchia was among those whom Mr. Mudd forced out of the company during a reorganization in August." (Duhig (2008))

The cavalier misuse of computerized models and simulations beyond their intended purposes is also mirrored between the engineering and financial worlds. Thus,

"Even though [Columbia's] debris strike was 400 times larger than the objects [the computer program] Crater is designed to model, neither Johnson engineers nor Program managers appealed for assistance from the more experienced Huntington Beach engineers, who might have cautioned against using Crater so far outside its validated limits. Nor did safety personnel provide any additional oversight."

In the subprime-credit boom,

"Some trading desks [at major banks] took the most arcane security, made of slices of mortgages, and entered it into the computer as if it were a simple bond with a set interest rate and duration... But once the mortgage market started to deteriorate, the computers were not able to identify all the parts of the portfolio that might be hurt." (Hansell, 2008)

5. Normalization of deviance, changing standards and rationales.

How do organizations react when what was not supposed to happen does, with increasing frequency and severity?

"This section [of the report] gives an insider perspective: how NASA defined risk and how those definitions changed over time for both foam debris hits and O-ring erosion. In both cases, engineers and managers conducting risk assessments continually "normalized" the technical deviations they found... Evidence that the design was not performing as expected was reinterpreted as acceptable and non-deviant, which diminished perceptions of risk throughout the agency... Engineers and managers incorporated worsening anomalies into the engineering experience base, which functioned as an elastic waistband, expanding to hold larger deviations from the original design. Anomalies that did not lead to catastrophic failure were treated as a source of valid engineering data that justified further flights... NASA documents show how official classifications of risk were downgraded over time."

The same pattern of normalizing close calls with disaster shows up as a precursor to corporate scandals and financial meltdowns. Several years before Ken Lay failed to heed V.P. Sherron Watkins' urgent plea that he and the CAO "sit down and take a good, hard, objective look at what is going to happen to Condor and Raptor [ventures] in 2002 and 2003", lest the company "implode in a wave of accounting scandals", he had refused to fire two high-revenue-generating oil traders after learning that they had stolen millions from the company and forged financial documents to hide it. A year later, those very same "rogue" traders used again falsified books to make huge unauthorized bets on oil prices, which went sour and exposed the company to several hundred millions dollars of potential losses (Eichenwald (2005)). In a near repeat scenario, in 2004 AIG Financial Services caused the parent company to be fined $126 million for helping clients engage in tax and accounting fraud. Yet the same manager (J. Cassano) remained in charge and was even put on the newly formed committee in charge of quality and risk control –until his unit blew up the company four years later.

6. Reversing the burden of proof. At the Beech-Nut Corporation in late 1970's, tests by the main food scientist suggested that the apple concentrate from a new (and cheaper) major supplier was probably adulterated. Top management responded by telling scientists that the company would not switch suppliers unless they could absolutely prove that it was. At the same time, they made it more difficult for them to conduct inspections.[47] Similarly, at NASA,

"When managers... denied the team's request for imagery, the Debris Assessment Team was put in the untenable position of having to prove that a safety-of-flight issue existed without the very images that would permit such a determination... Organizations that deal with high-risk operations must always have a healthy fear of failure – operations must be proved safe, rather than the other way around. NASA inverted this burden of proof..."

Similar reversals of evidentiary standards and shifting rationales were also documented in the decision process leading to the second Iraq war, particularly on the issue of weapons of mass destruction (Hersh (2004), Isikoff and Corn (2007)).

[47] The product was later shown to be 100% artificial. Beech-Nut was convicted and paid several million in fines and class-action settlements, while the CEO and the former Vice-President of manufacturing were sentenced to jail (Sims (1992)).

7. Malleable memories: forgetting the lessons of history. The commission investigating the Columbia accident was struck by how the same patterns had repeated themselves six years after Challenger:

"The Board found that dangerous aspects of NASA's 1986 culture, identified by the Rogers Commission, remained unchanged... Despite the constraints that the agency was under, prior to both accidents NASA appeared to be immersed in a culture of invincibility, in stark contradiction to post-accident reality. The Rogers Commission found a NASA blinded by its "Can-Do" attitude... which bolstered administrators' belief in an achievable launch rate, the belief that they had an operational system, and an unwillingness to listen to outside experts."

In the financial and regulatory worlds, the lessons of LTCM were also quickly forgotten, as were those of the internet bubble a few years later. Such failures of individual and collective memory are recurrent. They were even pointed out (and then forgotten) by a key observer and participant:

"An infectious greed seemed to grip much of our business community... The trouble, unfortunately, is that the shock of what has happened will keep malfeasance down for a while. But human nature being what it is –and memories fade– it will be back. And it is important that at that time appropriate legislation be in place to inhibit activities that we would perceive to be inappropriate." (Greenspan (2002)).

REFERENCES

Anderson, J. and C. Duhig (2008) "Death and Near-Death Experiences on Wall Street," *The New York Times,* September 21.

Andrews, E. (2007) "Fed and Regulators Shrugged as the Subprime Crisis Spread". *The New York Times,* December 18.

Duhig, C. (2008) "Pressured to Take More Risks, Fannie Mae Reached Tipping Point," *The New York Times,* October 5.

Greenspan, Alan. (2002). Testimony to the United States House Financial Services Committee, July 17.

Labaton, S. (2008) "Agency Rule Let Banks Pile Up Debt," *The New York Times,* Oct. 3.

Morgenson, G. (2008) "Behind Insurer's Crisis, Blind Eye to a Web of Risk," *The New York Times,* September 28.

Securities and Exchange Commission (2008) *SEC's Oversight of Bears Stearns and Related Entities: Consolidated Supervised Entity Program.* Inspector General's Report, Office of Audits, September 25, viii-ix. Available at http://www.sec-oig.gov.

Sorkin, A. (2008) "What Goes on Before a Fall? On Wall Street, Reassurance," *The New York Times,* September 30.

Suskind, R. (2004) "Without a Doubt," *The New York Times,* October 17.

SUPPLEMENTARY MATERIAL
Online Appendix B: Additional Proofs

Proof of the claims following Proposition 3. I establish here claims (a)-(b) concerning the "trickle-down" equilibrium illustrated in Figure 3. To make things simple, let $m^1 = m^1$, $c^1 = c^2$, $\delta^1 = \delta^2$, $a_H^{11} = a_H^{22}$, $a_L^{11} = a_L^{22}$ and $a_H^{11} - a_L^{11} = a_H^{22} - a_L^{22} \equiv a > 0$; finally, set $b^{ij} = 0$ for all i, j. The asymmetry in roles is then captured by $X \equiv (a_H^{12} - a_L^{12})/a > (a_H^{21} - a_L^{21})/a \equiv x$ and, especially, $Y \equiv -(a_L^{12} - b_L^{12})/a > -(a_L^{21} - b_L^{21})/a \equiv y$. I shall first provide conditions ensuring

$$\bar{s}^2(0) < \underline{s}^1(0) < \underline{s}^1(1) < \bar{s}^1(0) < \bar{s}^1(1) < \underline{s}^2(1), \tag{B.1}$$

which implies $[\underline{s}^1(1), \bar{s}^1(0)] \subset [\bar{s}^2(0), \underline{s}^2(1)] \equiv S$, as illustrated in Figure 3. From (A.7)-(A.8), the middle inequality is equivalent to $y < (1-q)(1+x)$, which can always be ensured given $q < 1$. The inequalities $\underline{s}^1(0) < \underline{s}^1(1)$ and $\bar{s}^1(0) < \bar{s}^1(1)$ hold for all $y > 0$ (complementarity). Turning finally to the two outer conditions, we have $\bar{s}^2(0) < \underline{s}^1(0)$ if

$$q\left(a_H^{12} - a_L^{12} + a_H^{22} - a_L^{22}\right) > a_H^{21} - a_L^{21} + a_H^{11} - a_L^{11},$$

or $qX > x + 1 - q$, while $\bar{s}^1(1) < \underline{s}^2(1)$ if

$$q\left[a_H^{21} - a_L^{21} + a_H^{11} - a_L^{11} + a_L^{21} - b_L^{21}\right] > a_H^{12} - a_L^{12} + a_H^{22} - a_L^{22} + a_L^{12} - b_L^{12},$$

or $Y > qy + X - qx + 1 - q$; both are clearly satisfied for X sufficiently larger than x and Y sufficiently larger than X. I can now prove the claims (a)-(c) made in the text.

(a) The result follows from the fact that $\bar{s}^2(0) \leq s \leq \underline{s}^2(1)$ and the definitions of these two thresholds in Proposition 1.

(b) The same definitions imply that an equilibrium with $(\lambda^1, \lambda^2) = (1, 1)$ (respectively, $(\lambda^1, \lambda^2) = (0, 0)$) exists if and only if $s^2 \leq \underline{s}^2(1)$ and $s^1 \leq \underline{s}^1(1)$ (respectively, $s^2 \geq \bar{s}^2(0)$ and $s^1 \geq \bar{s}^1(0)$), which corresponds to the left (respectively, right) region in Figure 3. In the middle region one must therefore have $\lambda^1 = \lambda_1^*(s^1; \lambda^2) \in (0, 1)$, where λ_1^* is the mixed-strategy best-response characterized in Proposition 1. It is decreasing in s^1 and increasing (respectively increasing) in λ^2 since for $a_L^{21} - b_L^{21} = -ya < 0$.

(c) Consider now the boundary loci within the middle region. An equilibrium with $(\lambda^1, \lambda^2) = (\lambda_1^*(s^1;1), 1)$ exists if and only if $s^1 \in [\underline{s}^1(1), \bar{s}^1(1)]$ and $s^2 \leq \underline{s}^2(\lambda_1^*(s^1;1))$. This is a decreasing function of s^1, which declines from $\underline{s}^2(\lambda_1^*(\underline{s}^1(1);1)) = \underline{s}^2(1)$ at $s^1 = \underline{s}^1(1)$ to $\underline{s}^2(\lambda_1^*(\bar{s}^1(0);1))$ at $s^1 = \bar{s}^1(0)$. For $|a_L^{21} - b_L^{21}|/a = y$ small enough, $\lambda_1^*(\bar{s}^1(0);\lambda_2)$ is very insensitive to the value of λ_2, so $\lambda_1^*(\bar{s}^1(0);1) \approx \lambda_1^*(\bar{s}^1(0);0)) = 0$ and hence $\underline{s}^2(\lambda_1^*(\bar{s}^1(0);1)) \approx \underline{s}^2(0) < \bar{s}^2(0)$. Therefore the curve $\underline{s}^2(\lambda_1^*(s^1;1))$ cuts the lower boundary of S_2 at a point $s_1 < \bar{s}^1(0)$, as on Figure 3.

Similarly, with $(\lambda^1, \lambda^2) = (\lambda_1^*(s^1;0), 0)$ exists if and only if $s^1 \in [\underline{s}^1(0), \bar{s}^1(0)]$ and $s^2 \geq \bar{s}^2(\lambda_1^*(s^1;0))$. This is a decreasing function of s^1, which declines to $\bar{s}^2(\lambda_1^*(\bar{s}^1(0);0)) = \bar{s}^2(0)$ at $s^1 = \bar{s}^1(0)$, from $\bar{s}^2(\lambda_1^*(\underline{s}^1(1);0)$ at $s^1 = \underline{s}^1(1)$. For y small enough, $\lambda_1^*(\underline{s}^1(1);\lambda_2)$ is very insensitive to the value of λ_2, so $\lambda_1^*(\underline{s}^1(1);0) \approx \lambda_1^*(\underline{s}^1(1);1) = 1$ and hence $\bar{s}^2(\lambda_1^*(\underline{s}^1(1);1)) \approx \bar{s}^2(1) > \underline{s}^2(0)$. Therefore, the curve $\bar{s}^2(\lambda_1^*(s^1;0))$ cuts the upper boundary of S_2 at a point $s_1 > \underline{s}^1(1)$, as in Figure 3. Finally, for $a_L^{21} - b_L^{21} = 0$,

$$\underline{s}^2(\lambda_1^*(s^1;1)) = \underline{s}^2(\lambda_1^*(s^1;0)) < \bar{s}^2(\lambda_1^*(s^1;0)) = \bar{s}^2(\lambda_1^*(s^1;1)), \tag{B.2}$$

since agent 1's behavior is independent of that of agent 2. For y small enough, it remains the case that $\underline{s}^2(\lambda_1^*(s^1;1)) < \bar{s}^2(\lambda_1^*(s^1;1))$, by continuity. These properties of the two curves imply that equilibria of the form $(\lambda^1, \lambda^2) = (\lambda_1^*(s^1;1), 1)$, $(\lambda^1, \lambda^2) = (\lambda_1^*(s^1;0), 0)$ and $(\lambda^1, \lambda^2) = (\lambda_1^*(s^1;\lambda_2), \lambda_2^*(s^2;\lambda_1))$ exist only in the three respective regions indicated in Figure 3. The equilibrium is therefore unique, except possibly in the middle region where both agents mix. But since it is unique for $x = y = 0$, by continuity it remains so for x and y small enough. ■

Lemmas for the proof of Proposition 6. I prove here the claims made following equation (A.23) in the paper's main appendix.

Lemma 1. *Under (25), there exists $\tilde{q}(K) < 1$ such that, for all $q \in [\tilde{q}(K), 1]$, $\bar{s}(0;q,K) < \underline{s}(1;K)$.*

Proof. By (A.15)-(A.17), $\bar{s}(0;q,K) < \underline{s}(1;K)$ means that

$$\begin{aligned} &\frac{m/\delta + [c - \delta P_L(K+E)]\,E}{q\,[P_H(K+E) - P_L(K+E)]\,(K+E) + P_L(K+E)E} \\ < \; &\frac{m/\delta + [c - \delta P_L(K)]\,E}{[P_H(K+E) - P_L(K)]\,(K+E) + P_L(K)E}. \end{aligned} \tag{B.3}$$

If (B.3) holds for $m = 0$, the first denominator must be greater than the second, as $P_L(K + E) < P_L(K)$. Therefore, (B.3) holds for all $m \geq 0$ if and only if it holds for $m = 0$, or

$$\frac{c - \delta P_L(K+E)}{c - \delta P_L(K)} < \frac{q\left[P_H(K+E) - P_L(K+E)\right](K+E) + P_L(K+E)E}{P_H(K+E)(K+E) - P_L(K)K}$$
$$= \frac{P_H(K+E)(K+E) - P_L(K+E)K}{P_H(K+E)(K+E) - P_L(K)K} - (1-q)\frac{\left[P_H(K+E) - P_L(K+E)\right](K+E)}{P_H(K+E)(K+E) - P_L(K)K},$$

that is,

$$(1-q)\frac{\left[P_H(K+E) - P_L(K+E)\right](K+E)}{P_H(K+E)(K+E) - P_L(K)K}$$
$$< \frac{P_H(K+E)(K+E) - P_L(K+E)K}{P_H(K+E)(K+E) - P_L(K)K} - \frac{c - \delta P_L(K+E)}{c - \delta P_L(K)}.$$

Finally, the condition takes the form

$$1 - q < \left(\frac{cK/(K+E) - \delta P_H(K+E)}{c - \delta P_L(K)}\right)\left(\frac{P_L(K) - P_L(K+E)}{P_H(K+E) - P_L(K+E)}\right). \tag{B.4}$$

Condition (25) ensures that $cK/(K+E) > \delta P_H(K+E)$, hence the result. ■

Lemma 2. *Assume (25). For any $\eta \in (0, 1/2)$ define $s_\eta(0;1,K) \equiv (1-\eta)\bar{s}(0;1,K) + \eta\underline{s}(1;K)$. There exists $q^*_\eta(K) < 1$ such that, for all $q \in (q^*_\eta(K), 1]$ condition (A.23) holds for all s in the nonempty interval $S_{2\eta}(K) \equiv (s_{2\eta}(0;1,K)), \underline{s}(1;K))$.*

Proof. For q close to 1 $\bar{s}(0;q,K)$ is close to $\bar{s}(0;1,K)$, so there exists $\hat{q}_\eta(K) \in (\tilde{q}(K), 1]$ such that, for all $q \in (\hat{q}_\eta(K), 1]$:

$$\bar{s}(0;q,K) < (1-\eta)\bar{s}(0;1,K) + \eta\underline{s}(1;K) \equiv s_\eta(0;1,K) < \underline{s}(1;K) \tag{B.5}$$

This implies, for any $s \in S_{2\eta}(K)$:

$$1 - \frac{\bar{s}(0;q,K)}{s} > \frac{s_{2\eta}(0;1,K) - s_\eta(0;1,K)}{\underline{s}(1;K)} = \eta\left(\frac{\underline{s}(1;K) - \bar{s}(0;1,K)}{\underline{s}(1;K)}\right) = \eta\left(1 - \frac{\bar{s}(0;1,K)}{\underline{s}(1;K)}\right).$$

Therefore, condition (A.23) holds provided that

$$1 - q \leq \eta\left(1 - \frac{\bar{s}(0;1,K)}{\underline{s}(1;K)}\right)\left(\frac{\bar{P}_q(K+E)\,(K+E)}{m/\left[\delta\,(\delta+s)\right] + \left[c/(\delta+s) + \bar{s}(0;1,K) - P_L(K+E)\right]E}\right),$$

which will be the case for all q in some nonempty subinterval $(q^*_\eta(K), 1]$ of $(\hat{q}_\eta(K), 1]$.

From Lemmas 1 and 2, the last step in the proof of Proposition 6 stated in the main Appendix follows: pick any $\eta \in (0, 1/2)$, e.g., $\eta > 0$ and very small, then define $S^*(K) \equiv S_{2\eta}(K)$ and $q^* = q^*_\eta(K)$. ■

Proofs for Proposition 7 and the restriction to $\lambda^i_H = 1$ in Proposition 1 A strategy profile for agent i at $t = 0$ (his "self 0") is a pair $\lambda^i = (\lambda^i_H, \lambda^i_L)$ of probabilities with which he truthfully encodes $\hat{\sigma}^i = \sigma$ in each state $\sigma^i = H, L$. A strategy profile for the same agent at $t = 1$ (his "self 1") is a pair $\xi^i = (\xi^i_H, \xi^i_L)$ of probabilities with which he chooses $e^i = 1$ in each recall state $\hat{\sigma}^i = H, L$. An *intrapersonal* equilibrium consists of a quadruplet $(\lambda^i_H, \lambda^i_L; \xi^i_H, \xi^i_L)$ and posterior beliefs (r^i_H, r^i_L) in each recall state that together constitute a Perfect Bayesian Equilibrium for agent i (keeping fixed the strategies of all $j \neq i$):

(i) The posterior beliefs (or "reliability") of each recall state are given by Bayes' rule:

$$r^i_H \equiv \Pr\left[\sigma^i = H \mid \hat{\sigma}^i = H\right] = \frac{q\lambda^i_H}{q\lambda^i_H + (1-q)(1-\lambda^i_L)}, \tag{B.6}$$

$$r^i_L \equiv \Pr\left[\sigma^i = L \mid \hat{\sigma}^i = L\right] = \frac{(1-q)\,\lambda^i_L}{(1-q)\,\lambda^i_L + q(1-\lambda^i_H)}. \tag{B.7}$$

(ii) Date-1 actions are optimal: $\xi^i_\sigma = 1$ if $\alpha E[\theta \mid \hat{\sigma}_i] > c$ and $\xi^i_\sigma = 0$ if $\alpha E[\theta \mid \hat{\sigma}_i] < 0$.

(iii) At $t = 0$, the agent in each state $\sigma = H, L$ optimally chooses (or randomizes between) which $\hat{\sigma} = H, L$ to encode, taking (i) and (ii) as given.

Lemma 3. *Let $m > 0$ and fix any strategies $(\lambda^{-i}_H, \lambda^{-i}_L)$ (whether equilibrium or not) of players $j \neq i$. If $(\lambda^i_H, \lambda^i_L)$ is an intrapersonal equilibrium for i such that $\max\{\lambda^{-i}_H, \lambda^{-i}_L\} < 1$, then $(1, 1)$ is also an equilibrium and it makes him strictly better off in both states.*

Proof. I shall omit time-0 subscripts for simplicity. For any $(\sigma, \hat{\sigma}) \in \{L, H\}^2$, let $V^i_{\sigma\hat{\sigma}}$ denote the date-0 expected value of U^i_1 that agent i could achieve in state σ by encoding it as $\hat{\sigma}$, *if* his behavior at date 1 was guided by "naive" posteriors, i.e. $\xi^i = 1$ when $\hat{\sigma} = H$ and $\xi^i = 0$ when $\hat{\sigma} = L$. The $V^i_{\sigma\hat{\sigma}}$'s do not depend on any actual or conjectured mixing probabilities used by the agent at $t = 0$. Next, define $U^i_{\sigma\hat{\sigma}}$ from the same encoding choices as $V^i_{\sigma\hat{\sigma}}$, but anticipating that beliefs at $t = 1$ will be derived from $(\lambda^i_H, \lambda^i_L)$ using (B.6)-(B.7). Finally, let U^i_σ be the date-0 expected utility achieved in state σ by following the mixing strategy $(\lambda^i_H, \lambda^i_L)$. Thus, for all $\sigma, \hat{\sigma}$ and $\tilde{\sigma} \neq \sigma$,

$$U^i_{\sigma\hat{\sigma}} \equiv r^i_{\hat{\sigma}} V^i_{\sigma\hat{\sigma}} + (1 - r^i_{\hat{\sigma}}) V^i_{\sigma\tilde{\sigma}}, \tag{B.8}$$

$$U^i_{\sigma} = \lambda^i_{\sigma\sigma} U^i_{\sigma\sigma} + (1 - \lambda^i_{\sigma\sigma}) \left(U^i_{\sigma\tilde{\sigma}} - m \right). \tag{B.9}$$

For any alternative candidate strategy $(\lambda'^i_H, \lambda'^i_L)$ I use the same notations but with "primes" on all the variables. I first show that

$$U^i_H = U^i_{HL} < U'^i_{HH} \iff \left(1 - r'^i_H - r^i_L\right) \left(V^i_{HH} - V^i_{HL}\right) < m, \tag{B.10}$$

$$U^i_L = U^i_{LH} < U'^i_{LL} \iff \left(1 - r'^i_L - r^i_H\right) \left(V^i_{LL} - V^i_{LH}\right) < m. \tag{B.11}$$

In each case the equality comes from the fact that $\lambda^i_\sigma < 1$, so that denial is an optimal strategy in state σ, and the equivalence between inequalities then follows from (B.8) applied to both $(\lambda^i_H, \lambda^i_L)$ and $(\lambda'^i_H, \lambda'^i_L)$. Next, note that for $(\lambda^i_H, \lambda^i_L)$ to be a personal equilibrium the inequalities in (B.10)-(B.11) must be reversed when $(\lambda'^i_H, \lambda'^i_L) = (\lambda^i_H, \lambda^i_L)$, meaning that

$$\left(1 - r^i_H - r^i_L\right) \min \left\{ V^i_{HH} - V^i_{HL},\ V^i_{LL} - V^i_{LH} \right\} \geq m. \tag{B.12}$$

Suppose first that $r^i_L + r^i_H \leq 1$, implying $V^i_{HH} - V^i_{HL} > 0$ and $V^i_{LL} - V^i_{LH} > 0$. Consider then $(\lambda'^i_H, \lambda'^i_L) \equiv (1, 1)$, which by (B.6)-(B.7) leads to $(r'^i_H, r'^i_L) = (1, 1)$. Equations (B.10)-(B.11) are clearly satisfied, and the same is true if r^i_H and r^i_L are both replaced by 1. Therefore, systematic truthfulness leads to higher expected utility in each state than the original $(\lambda^i_H, \lambda^i_L)$ and it is also an equilibrium.

Suppose next that $r^i_L + r^i_H > 1$. From (B.6)-(B.7), we have

$$r^i_L + r^i_H > 1 \Leftrightarrow \lambda^i_H + \lambda^i_L > 1. \tag{B.13}$$

Since $\max\{\lambda^i_H, \lambda^i_L\} < 1$, this implies $\left(\lambda^i_H, \lambda^i_L\right) \in (0, 1)^2$: the agent mixes in both states, so $V^i_{HH} - V^i_{HL} = V^i_{LL} - V^i_{LH} = m / \left(1 - r^i_H - r^i_L\right) < 0$. However, by definition of the $V^i_{\sigma\hat{\sigma}}$'s,

$$\left(V^i_{HH} - V^i_{HL}\right) / \delta = (s + \delta) (\alpha\theta_H - c) + s \left(W^i_H - W^i_L\right), \tag{B.14}$$

$$\left(V^i_{LL} - V^i_{LH}\right) / \delta = \alpha(s\theta_H + \delta\theta_L) - c + s \left(W^i_H - W^i_L\right), \tag{B.15}$$

where $W^i_\sigma \equiv (1 - \alpha)\xi^{-i}_\sigma \theta_\sigma + \gamma_\sigma$ is the true final payoff that agent i will receive in state σ

from the (aggregate) effort decisions ξ_σ^{-i} of the other players, and exogenously (last term). The two expressions differ by $\alpha\delta(\Delta\theta) > 0$, so $(\lambda_H^i, \lambda_L^i)$ cannot be an equilibrium. ■

Intuitively, any strategy with distortion or memory censoring in both states represents an inefficient way of encoding information, wasting $m > 0$ with positive probability. It does not corresponds to a best response to others' behavior since the agent can, *on his own,* improve upon it (under the very weak assumption that he can coordinate his "self 0" and "self 1" on a Pareto-superior intrapersonal equilibrium, which always exists). I therefore restrict attention, throughout the paper, to *efficient encoding strategies*, meaning that $\lambda_H^i = 1$ or $\lambda_L^i = 1$ for every i. This also implies, by (B.6)-(B.7),

$$r_H^i \geq q \geq 1 - r_L^i \quad \text{and} \quad \xi_H^i = 1 \geq \xi_L^i. \tag{B.16}$$

Finally, as explained in footnote 17, I generally restrict attention to *symmetric equilibria* (except in Section 1.4, or when there is a large number ($n \to +\infty$) of identical agents, as in Section 3). These two conditions will be implicit in the use of the word "equilibrium".

Lemma 4. *(1) For $\Delta\gamma \geq -(1-\alpha)\min\{\theta_H, \Delta\theta\}$ there can be no equilibrium with $\lambda_H = 0$, and no profitable individual deviation to $\lambda_H^i < 1$ from any equilibrium in which $\lambda_H = 1$.*
(2) For $\Delta\gamma > -\min\{(1-\alpha)\theta_H, (1-\alpha)\Delta\theta, \kappa^*(s)\Delta\theta)\}$, where $\kappa^*(s) > 0$ is given by (B.21) below, there can be no equilibrium with $\lambda_H < 1$. Thus, the results of Propositions 2-5 remain unchanged, up to the substitution of $\Delta\gamma + \Delta\theta$ for $\Delta\theta$ everywhere.

Proof. Following the same reasoning as in text (or directly from (B.8)-(B.9)) and omitting time subscripts to lighten the notation, the incentive to misinterpret or misremember H as L (gross of the cost m) is given by

$$\begin{aligned}\left(U_{HL}^{i,} - U_{HH}^{i,} + m\right)/\delta = {} & s\left(1 - r_L^i - r_H^i\right)(\gamma_H - \gamma_L) + \left(\xi_H^i - \xi_L^i\right)[c - \delta\alpha\theta_H] \\ & + s\alpha\left\{\left[\left(1 - r_L^i\right)\xi_L^i - r_H^i\xi_H^i\right]\theta_H - \left[\left(1 - r_H^i\right)\xi_H^i - r_L^i\xi_L^i\right]\theta_L\right\} \\ & + s(1-\alpha)\left(1 - r_L^i - r_H^i\right)\left\{\left[\lambda_H^{-i}\xi_H^{-i} + \left(1 - \lambda_H^{-i}\right)\xi_L^{-i}\right]\theta_H\right. \\ & \left. - \left[\lambda_L^{-i}\xi_L^{-i} + \left(1 - \lambda_L^{-i}\right)\xi_H^{-i}\right]\theta_L\right\}.\end{aligned} \tag{B.17}$$

The incentive to miscode L as H is given by the same expression, with H and L switched:

$$
\begin{aligned}
(\text{B.18}) \quad \left(U^{i,}_{LH} - U^{i,}_{LL} + m\right)/\delta &= s\left(1 - r^i_H - r^i_L\right)(\gamma_L - \gamma_H) + \left(\xi^i_L - \xi^i_H\right)[c - \delta\alpha\theta_L] \\
&\quad + s\alpha\left\{\left[\left(1 - r^i_H\right)\xi^i_H - r^i_L\xi^i_L\right]\theta_L - \left[\left(1 - r^i_L\right)\xi^i_L - r^i_H\xi^i_H\right]\theta_H\right\} \\
&\quad + s(1-\alpha)\left(1 - r^i_H - r^i_L\right)\left\{\left[\lambda^{-i}_L\xi^{-i}_L + \left(1 - \lambda^{-i}_L\right)\xi^{-i}_H\right]\theta_L\right. \\
&\quad \left. - \left[\lambda^{-i}_H\xi^{-i}_H + \left(1 - \lambda^{-i}_H\right)\xi^{-i}_L\right]\theta_H\right\}.
\end{aligned}
$$

From Lemma 3 and (B.16) we know that $\lambda^i_H = 1$ or $\lambda^i_L = 1$ and that in either case, $\xi^i_H = 1$, so in a symmetric equilibrium, $\xi^{-i}_H = \xi^i_H = 1$.

1. *Equilibria with* $\lambda_H = 1$. This implies $r^i_L = 1$, so $\xi^i_L = 0 = \xi^{-i}_L$ and (B.17) becomes

$$
\begin{aligned}
\left(U^{i,}_{HL} - U^{i,}_{HH} + m\right)/\delta &= -sr^i_H(\gamma_H - \gamma_L) + [c - \delta\alpha\theta_H] \\
&\quad - s\alpha\left[r^i_H\theta_H + \left(1 - r^i_H\right)\theta_L\right] - sr^i_H(1-\alpha)\left[\theta_H - \left(1 - \lambda^{-i}_L\right)\theta_L\right] \\
&= -[(\delta + s)\alpha(r^i_H\theta_H + (1 - r^i_H)\theta_L) - c] - sr^i_H\Delta\gamma \\
&\quad - \Delta\theta[\delta\alpha(1 - r^i_H) + sr^i_H(1-\alpha)] - sr^i_H(1-\alpha)\lambda^{-i}_L\theta_L.
\end{aligned}
$$

The first term is negative since $r^i_H \geq q$, so it suffices that

$$
(\text{B.19}) \qquad sr^i_H\Delta\gamma \geq -\Delta\theta[\delta\alpha(1 - r^i_H) + sr^i_H(1-\alpha)] - sr^i_H(1-\alpha)\lambda^{-i}_L\theta_L.
$$

This inequality is linear in r^i_H and holds for $r^i_H = 0$. For $r^i_H = 1$, it takes the form $\Delta\gamma \geq -(1-\alpha)\left[\Delta\theta + \lambda^{-i}_L\theta_L\right]$, which holds whatever the sign of θ_L when $\Delta\gamma \geq -(1-\alpha)\min\{\Delta\theta, \theta_H\}$. Thus, an individual deviation to miscoding H as L is never profitable. As to miscoding L as H, (B.18) becomes

$$
\begin{aligned}
\left(U^{i,}_{LH} - U^{i,}_{LL} + m\right)/\delta &= -[c - \delta\alpha\theta_L] + s\alpha\left[\left(1 - r^i_H\right)\theta_L + r^i_H\theta_H\right] \\
&\quad + s(1-\alpha)r^i_H\left[\theta_H - \left(1 - \lambda^{-i}_L\right)\theta_L\right] + sr^i_H(\gamma_H - \gamma_L) \\
&= -[c - (\delta + s)\alpha\theta_L] + sr^i_H\left[\Delta\theta + \Delta\gamma + (1-\alpha)\lambda^{-i}_L\theta_L\right],
\end{aligned}
$$

which is identical to (10) except that $\Delta\theta$ is replaced by $\Delta\theta + \Delta\gamma$. Therefore, all the previous results and formulas shown for $\Delta\gamma = 0$ and imposing $\lambda^i_H \equiv 1$ remain the same, provided $\Delta\theta + \Delta\gamma >$ replaces $\Delta\theta$ wherever it appears.

2. Ruling out equilibria with $\lambda_H < 1 = \lambda_L$. If $\lambda_H^i < 1$ then $\lambda_L^i = 1$ by Lemma 3, so $r_H^i = 1$ and hence $\xi_H^i = 1 = \xi_H^{-i}$. Therefore, (B.17) simplifies to:

$$\begin{aligned}\left(U_{HL}^{i,} - U_{HH}^{i,} + m\right)/\delta = &-\left(1-\xi_L^i\right)\left[\left(\delta+s\right)\alpha\theta_H - c\right] \\ &- sr_L^i\left\{\Delta\theta\left[\alpha\xi_L^i + \left(1-\alpha\right)\xi_L^{-i}\right] + \Delta\gamma + \left(1-\alpha\right)\lambda_H^{-i}\left(1-\xi_L^{-i}\right)\theta_H\right\}.\end{aligned}$$

In (symmetric) equilibrium $\xi_L^i = \xi_L^i$ and $\lambda_H^i = \lambda_H^{-i}$, so this expression is strictly negative and no equilibrium with $\lambda_H^i < 1$ exists, when

$$\xi_L^i\Delta\theta + \left(1-\xi_L^i\right)\lambda_H^i\left(1-\alpha\right)\theta_H + \Delta\gamma \geq 0. \tag{B.20}$$

For $\Delta\theta + \Delta\gamma \geq 0$, we can rule out any equilibrium with $\xi_L^i = 1$, and in particular any equilibrium with $\lambda_H^i = 0$ (which implies $r_L^i = 1-q$, so $\xi_L^i = 1$). As to an equilibrium with $\xi_L^i < 1$, given $\lambda_L^i = 1$ this requires that λ_H^i not be below the critical value that makes an agent indifferent to working or not, given $\hat{\sigma}^i = L : \theta_L + \left[1 - r_L\left(\lambda_H, 1\right)\right]\Delta\theta \leq c/\alpha\left(s+\delta\right)$, or

$$\lambda_H^i\left(1-\alpha\right)\left(\frac{\theta_H}{\Delta\theta}\right) \geq \left(1-\alpha\right)\left(\frac{\theta_H}{\Delta\theta}\right)\left[1 - \left(\frac{1-q}{q}\right)\left(\frac{c/\alpha\left(s+\delta\right) - \theta_L}{\theta_H - c/\alpha\left(s+\delta\right)}\right)\right] \equiv \kappa^*(s). \tag{B.21}$$

Therefore, by (B.20), any equilibrium with $\xi_L^i < 1$ is ruled out for $\Delta\gamma \geq -\Delta\theta\min\left\{1, \kappa^*(s)\right\}$; hence the result. Note, moreover, that since $\kappa^*(s)$ is increasing, if the second inequality in (4) is strengthened to $q\theta_H + \left(1-q\right)\theta_L > c/\alpha\delta$, then $\kappa_H^*(0) > 0$ and such equilibria are ruled out for *any* s if $\Delta\theta\min\left\{1, \kappa^*(0)\right\} + \Delta\gamma > 0$. ■

Proof of Proposition 7. I again show the result for the more general specification (A.1), under which $\kappa \geq \max\{1, \theta_H/\Delta\theta\}$ is a special case of $\Delta\gamma \leq -\max\{\Delta\theta, \theta_H\}$.Note first that since $1 - r_L^i \leq q$, (27) implies that $\xi_L^i = 0$ and thus, in a equilibrium, $\xi_L^{-i} = \xi_L^i = 0$.

1. Ruling out equilibria with $\lambda_L^i < 1 = \lambda_H^i$. If $\lambda_L^i < 1$ then $\lambda_H^i = 1 = \lambda_H^{-i}$ in equilibrium by Lemma 3 and symmetry, so $r_L^i = 1$ and $\xi_L^i = 0 = \xi_L^{-i}$. Therefore, (B.18) simplifies to:

$$\begin{aligned}\left(U_{LH}^{i,} - U_{LL}^{i,} + m\right)/\delta &= sr_H^i\Delta\gamma - \xi_H^i\left[c - \delta\alpha\theta_L\right] + s\alpha\xi_H^i\left[\left(1-r_H^i\right)\theta_L + r_H^i\theta_H\right] \\ &\quad + sr_H^i\left(1-\alpha\right)\xi_H^{-i}\left[\lambda_H^{-i}\theta_H - \left(1-\lambda_L^{-i}\right)\theta_L\right] \\ &= -\xi_H^i\left[c - \left(s+\delta\right)\alpha\theta_L\right] + sr_H^i\Delta\gamma + \xi_H^i\left[\Delta\theta + \left(1-\alpha\right)\lambda_L^i\theta_L\right]\end{aligned}$$

Since $\Delta\gamma+\xi_H^i\left[\Delta\theta+(1-\alpha)\,\lambda_L^{-i}\theta_L\right]\leq\Delta\gamma+\xi_H^i\left[\Delta\theta+\max\left\{0,\theta_L\right\}\right]<0$, the previous expression is strictly negative, and no equilibrium with $\lambda_L^i<1$ exists.

2. Equilibria with $\lambda_L=1$. This implies $r_H^i=1$, so $\xi_H^i=1=\xi_H^{-i}$ and (B.18) becomes

$$\begin{aligned}\left(U_{LH}^{i,}-U_{LL}^{i,}+m\right)/\delta &= -sr_L^i\left(\gamma_L-\gamma_H\right)-\left[c-\delta\alpha\theta_L\right]+s\alpha\theta_H+sr_L^i\left(1-\alpha\right)\lambda_H^{-i}\theta_H\\ &= -[c-(\delta+s)\alpha(r_L^i\theta_L+(1-r_L^i)\theta_H)]\\ &\quad+sr_L^i\Delta\gamma-(1-r_L^i)\delta\alpha\Delta\theta+sr_L^i\left[\alpha\Delta\theta+\left(1-\alpha\right)\lambda_H^{-i}\theta_H\right].\end{aligned}$$

The first term is negative since $r_L^i\leq1-q$, so it suffices that

$$sr_L^i\Delta\gamma\leq(1-r_L^i)\delta\alpha\Delta\theta-sr_L^i\left[\alpha\Delta\theta+\left(1-\alpha\right)\lambda_H^{-i}\theta_H\right]. \tag{B.22}$$

This inequality is linear in r_L^i and holds for $r_L^i=0$. For $r_L^i=1$, it takes the form $\Delta\gamma\leq-\left[\alpha\Delta\theta+\left(1-\alpha\right)\lambda_H^{-i}\theta_H\right]$, which holds for all λ_H^i if $\Delta\gamma\leq-\left[\alpha\Delta\theta+\left(1-\alpha\right)\theta_H\right]$. This expression is greater than $-\max\{\Delta\theta,\theta_H\}$ whatever the sign of θ_L, hence the result ruling out any profitable individual deviation to $\lambda_L^i<1$. As to (B.17), it becomes

$$\begin{aligned}\left(U_{HL}^{i,}-U_{HH}^{i,}+m\right)/\delta &= -sr_L^i\left(\gamma_H-\gamma_L\right)+\left[c-\delta\alpha\theta_H\right]-s\alpha\theta_H-s\left(1-\alpha\right)r_L^i\lambda_H^{-i}\theta_H\\ &= -\left[(s+\delta)\alpha\theta_H-c\right]-sr_L^i\left[\Delta\gamma+\left(1-\alpha\right)\lambda_H^{-i}\theta_H\right].\end{aligned}$$

Since $-\Delta\gamma-\theta_H>0$, $\lambda_H^i=1$ is an equilibrium (implying $r_L^i=1$) if and only if

$$s\leq\frac{m/\delta+\delta\alpha\theta_H-c}{-\Delta\gamma-\theta_H}\equiv\underline{s}(1). \tag{B.23}$$

Similarly, $\lambda_H^i=0$ is an equilibrium (implying $r_L^i=1-q$) if and only if

$$s\geq\frac{m/\delta+\delta\alpha\theta_H-c}{\left(1-q\right)\left(-\Delta\gamma\right)-\alpha\theta_H}\equiv\bar{s}(0), \tag{B.24}$$

if $-\Delta\gamma>\alpha\theta_H/\left(1-q\right)$, otherwise, let $\bar{s}(0)\equiv+\infty$. Multiple equilibria occur for $\bar{s}(0)<\underline{s}(1)$, i.e. $q\left(-\Delta\gamma\right)<\left(1-\alpha\right)\theta_H$. The treatment of the mixed-strategy equilibrium is similar to that in Proposition 2. ■