Value of information under groupthink

Dissertation for MRes in Economics

Marcelo Woo
University of Nottingham

August 20, 2021

1 Introduction

Groupthink can be defined as a situation in which individuals in a group collectively engage in delusions and willful blindness to information despite it being available and valuable to improve decision-making. It is often suggested that groupthink has led to episodes of collective failures that sometimes have produced substantial losses, such as the Challenger shuttle disaster and the subprime financial crisis.

In a classic article, Benabou (2013) argues that groupthink may emerge as an equilibrium of a game where interacting individuals experience anticipatory utility over future outcomes, and where each individual may choose to selectively distort the signal he observes about the underlying state of the world. He characterizes the conditions under which a "delusion equilibrium" exists (which he calls groupthink), in which individuals "delude themselves" that everything is fine and go ahead with a doomed project, ignoring warning signals along the way. At the same time, "realist equilibria" are also possible, in which people correctly encode the information they receive and avoid engaging in low quality projects. What are the conditions that make delusion or realism more likely? Can something be done to avoid groupthink and its damaging consequences? This thesis aims to explore this question from the perspective of information design. We consider the problem of an information designer who can choose the structure of the signals that the agents –who may engage in delusion or not– observe, and determine how much information should be optimally provided such that it maximizes her ex-ante material payoffs. As such, this work combines the branches of literature on motivated beliefs and groupthink from behavioral economics with mechanism design.

Standard economic theory suggests that, in a principal-agent environment where the agents' interests are aligned with each other and with the principal's objective, more information is better: information has a decision-making value in that it improves the choices made by rational agents. However, if agents derive direct utility from their beliefs about the state of the world (e.g., the value of a project they are advancing) and are capable of self-deluding themselves to obtain optimistic beliefs (denying or distorting the information when encoding it into memory), this intuition may fail. Intuitively, if agents are sufficiently unaware of their tendency to distort information, the incentives for delusion may be strongest when signals are perfectly informative of the state of the world. Since they would naively believe the information encoded into memory without regard of the actual information, delusion might be more effective when signals are more informative. If that is the case, the information designer may face a trade-off in her choice of the optimal signal informativeness: more information improves decision-making in the case agents choose to treat it realistically, but at the same time it may increase the incentives for self-delusion if the utility directly derived from beliefs is sufficiently strong. Therefore, she might find that the optimal signals are less than perfectly informative, departing from the standard result.

Under a set of simplifying assumptions, we find that when self-delusion by agents is possible, a delusion equilibrium always exists, for any signal informativeness $\rho \in [1/2, 1]$; but a realism equilibrium exists only for some combinations of parameters and only for sufficiently high values of ρ. In terms of her ex-ante expected material payoffs, the information designer is always better off in a situation where agents treat information realistically (independently of whether this can be sustained as an equilibrium) than when all agents engage in denial of bad news. Conditional on realism being sustained as an equilibrium, the ex-ante material payoffs of realism are strictly increasing in ρ, giving the central planner incentives to raise the informativeness of the signals as much as possible.

However, since realism is not assured to be an equilibrium strategy of the agents in the organization, the optimal choice of ρ may be influenced by a trade-off between facilitating correct decision-making and mitigating the benefits of anticipation feelings that tempt agents to engage in delusion. If the downside value of the project is sufficiently negative, correct decision making matters more to the agents' payoffs, and providing more precise signals strengthens their incentives to engage in realism. Otherwise, anticipation feelings dominate, and making signals *less* precise restraints agents' temptation to deny bad news.

We find that there exist some combinations of parameters such that a realism equilibrium exists for sufficiently high ρ and the optimal choice by the central planner are perfectly informative signals. This

confirms the standard result of the literature and suggests that, at least for the assumptions made, anticipatory feelings are not strong enough to overcome the decision-value of the information.

Section 2 reviews literature. In section 3, we present the extension of Benabou (2013) model of groupthink. In section 4, we provide a benchmark and show that if agents cannot delude themselves then perfectly informative signals are always optimal. Section 5 analyzes the full model with self-delusion and determines the optimal informativeness. Section 6 discusses assumptions and manipulation of delusion equilibria through ρ. Section 7 concludes.

2 Literature review

This work relates mainly to two branches of literature: motivated beliefs and groupthink on the one hand, and information design, on the other.

Groupthink Groupthink is defined as "a pattern of thought characterized by self-deception, forced manufacture of consent, and conformity to group values and ethics". The term was coined in Janis (1979), a study of foreign policy decisions including the Cuban missile crisis and the Vietnam war, to refer to a number of symptoms common in group decisions that had ended disastrously. These symptoms include illusion of invulnerability, collective rationalization, direct pressure on dissenters, self-censorship, and illusion of unanimity.

While such errors in collective decision-making have been studied in a large literature in social and organizational psychology, it has received attention relatively recently in Economics, where two main approaches can be identified.

The first, rational approach has proposed a variety of mechanisms through which individually rational agents collectively end up taking wrong decisions at the group level. One source of group errors is signalling. When their "expertise" or quality of information is uncertain, agents whose private signal opposes the opinions already expressed by others may ignore it in order not to appear unqualified (Ottaviani and Sorensen (2001), Prat (2005)). Another source of group errors are bubbles, herding and cascading (Banerjee (1992), Bikhchandani, Hirshleifer and Welch (1992), Caplin and Leahy (1994), Chamley and Gale (1994)). In such models, each investor decides rationally to go against his own signal only when the herd is more likely to be correct. Harel et al study the speed of learning and the amount of information transmitted in a herding setting, where long-lived rational agents learn from repeatedly observing each others' actions. They show that after a consensus on an action is formed in initial periods, it is optimal for every agent to continue taking the consensus action, even when her private information indicates otherwise. A long time may pass until any of them breaks the wrong consensus (a situation they call *grouphink*), and this leads to long periods of loss of information aggregation and a slow speed of learning.

However, the rational approach lacks many of the psychological dimensions of groupthink that have been emphasized by Janis (1979) and the organizational psychology literature that followed. The second approach in Economics introduces wishful thinking and motivated reasoning to explain collective errors. This literature recognizes that information can affect utility not only indirectly, through improving a subsequent decision (strategic or instrumental value of information), but also directly, in the form of anticipation or realization (hedonic value) (Engelman et al (2019); Mijovic-Prelec et al (2010)). If individuals derive direct utility from their beliefs (e.g., about future outcomes, or about their type such as their health condition), then an incentive to manipulate beliefs by avoiding information ex-ante, or by distorting it or processing it asymmetrically ex-post, may emerge. There is growing experimental evidence that individuals may engage in such belief-influencing behaviour, even when incentivized to take accurate decisions (Eil and Rao (2010), Mobius et al (2010), Mijovic-Prelec and Prelec (2010), Mayraz (2011), Choi and Lou (2010)).

Benabou (2013) incorporates the possibility of agents engaging in wishful thinking that arises from anticipatory feelings as a building block to propose a model of groupthink[1]. Each agent faces a tradeoff

[1]Benabou (2013) also shows that the main insights do not depend on the assumptions about anticipatory utility and

between accepting the dismal news about the project's value and act accordingly, or maintaining hopeful beliefs by discounting or forgetting such news, at the risk of making overoptimistic decisions. He characterizes the conditions under a delusion equilibrium (groupthink) may arise as an equilibrium, noting that it may coexist with a realism equilibrium as well. The model assumes that signals perfectly reveal the state of the world, which we relax in order to analyze the optimal signal informativeness.

Information design / Bayesian persuasion The second branch of literature to which this work relates to is that of information design, a subfield of mechanism design that focuses on the optimal provision of information. In an information design problem, the payoffs of the players (agents) depend on their actions and also on the realization of a payoff-relevant state of the world. An information designer can choose and credibly commit to how to provide information about the state to the players. Information design studies how the information designer can influence, through her choice of the information provided, the optimal behavior of players to maximize her objective.

The seminal work by Gentzkow and Kamenika (2011), entitled "Bayesian persuasion", considers an environment with one player and an information designer with informational advantage about the state. They define "persuasion" as "influencing behavior via provision of information". A variety of alternative setups has flourished since then, with one or many players, and with or without prior information by the players (Bergemann and Morris (2013, 2016), Taneva (2015)), which could be either known or unknown to the information designer.

As the name of the paper suggests, the literature has focused on Bayesian agents, who understand how information is generated and react to it in a rational way. Our problem, however, involves agents that may optimally ignore or distort information, departing from the rational framework that the literature has been considering.

3 The model

To study the optimal informativeness under groupthink, we extend Benabou (2013) model by enriching its signal structure. Two versions of the extension are presented:

1. Case 1: Agents with perfect recall (no delusion possible)
2. Case 2: Imperfect recall (delusion strategy available)

Case 2 provides the full extension and is the main focus of analysis, while Case 1 is a simplified version used for benchmark purposes.

3.1 Setup

Here we present the full version of the model (case 2). First, the technology and preferences, which follow Benabou 2013, are described. Then the extensions are presented. Finally, a set of simplifying assumptions is introduced.

3.1.1 Technology

A group of N agents, $i \in \{1, 2, ..., N\}$ is engaged in a joint project or activity that generates spillovers. In $t = 1$, each agent chooses his level of effort $e^i = 0$ or 1 with cost ce^i, $c > 0$. In $t = 2$, i obtains expected utility

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

where $e^{-i} \equiv 1/(N-1) \sum_{j \neq i} e^j$ is the average effort of others and $(1 - \alpha) \in [0, 1 - 1/n]$ the degree of interdependence. The utility is additively separable in efforts e^i and e^{-i}, so there is no interdependence between effort decisions.

imperfect recall and provide an alternative, ex-ante information avoidance version that features Kreps-Porteus (1978) preferences for late resolution of uncertainty.

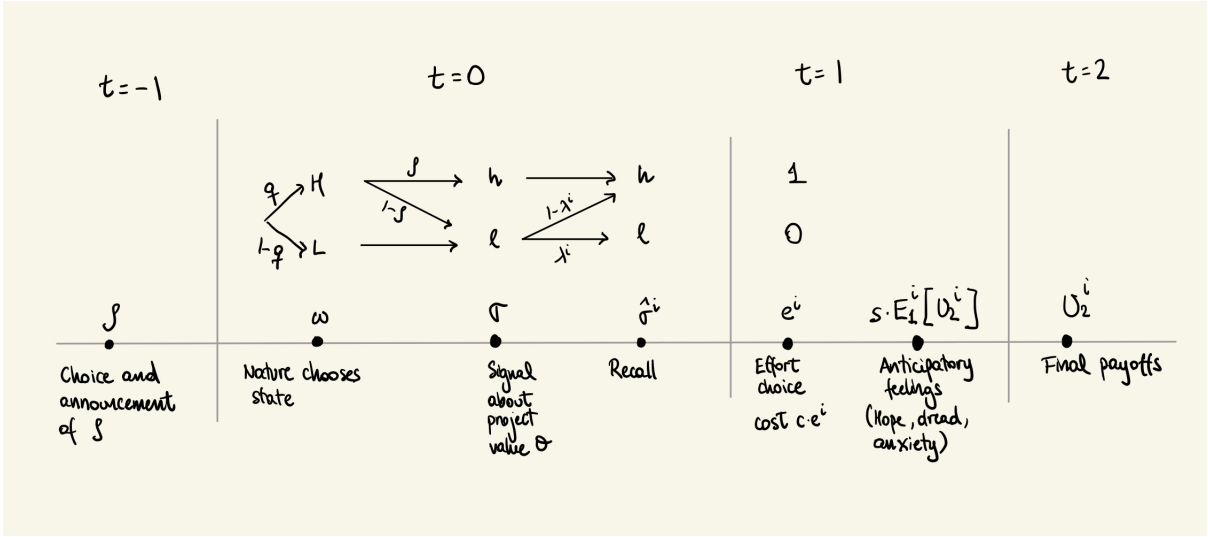


Figure 1: Timeline

The productivity of the project, θ, is uncertain. There are two states of the world, $\omega \in \{H, L\}$, with prior probabilities $(q, 1-q)$. The project's expected value is either $\theta = \theta_H$ in the high state (upside value) and $\theta = \theta_L$ in the low state (downside value) , where $\theta_H > 0$ and $\Delta\theta \equiv \theta_H - \theta_L > 0$. At $t = 0$, all agents observe an informative signal about the state; the precise structure of the signals is introduced in subsection 3.1.3. The value of inaction, $e^i = 0$, is normalized to zero. We focus on the case where $\theta_L < 0$: exerting effort to advance the project is harmful in state L so the ideal choice for each agent i (and the information designer) about others' effort is $e^j = 1$ in state H and $e^j = 0$ in state L.

3.1.2 Preferences

The date-1 agent's payoff includes not only the effort cost, $-ce^i$, but also an anticipatory utility that is psychologically experienced from imagining the future outcome U_2^i, which is realized in $t = 2$. Such anticipation feelings are given by $s \cdot \mathbb{E}_1^i \left[U_2^i\right]$, where $s \geq 0$ (for "savoring"). It also includes the standard discounting of future payoffs at a discount factor $\delta \in (0, 1]$. The date-1 agent i's choice is whether to exert effort or not to maximize his date-1 expected utility given the beliefs he has:

$$U_1^i = -ce^i + s\mathbb{E}_1^i \left[U_2^i\right] + \delta\mathbb{E}_1^i \left[U_2^i\right] \tag{2}$$

The beliefs he has at $t = 1$ depends on the signal he received at $t = 0$ and how objectively he (more precisely, his date-0 self) processed it in $t = 0$.

At $t = 0$, upon observing the signal $\sigma^i = H, L$, each agent chooses (consciously or not) how much attention to pay to the news or how to interpret it. He can either a) engage in *realism* –accept the news and encode $\sigma^i = \hat{\sigma}^i$ truthfully into memory (his date-1 information set); b) engage in *denial* –encoding $\hat{\sigma}^i \neq \sigma^i$– with a possible cost $m^i \geq 0$ that can be interpreted as a mental cost of distorting information; and c) when indifferent between those two strategies, use a mixed strategy. With these strategies available, date-0 agent i maximizes his date-0 expected utility –the discounted utility of all payoffs:

$$U_0^i = -m^i + \delta\mathbb{E}_0^i \left[-ce^i + s\mathbb{E}_1 \left[U_2^i\right]\right] + \delta^2\mathbb{E}_0^i \left[U_2^i\right] \tag{3}$$

where $\mathbb{E}_t [.]$ denotes expectations at $t = 0, 1$, and $m^i \geq 0$ is the cost if denial is chosen.

The inferences by the date-1 agent resulting from the date-0 choices of realism or denial are discussed in the next subsection, after introducing the signal structure.

3.1.3 Signal structure and informativeness

Signal structure The signal structure of the model is enriched by allowing signals to be imperfectly correlated with the state of the world. We consider a private signals setting where each agent i privately observes an independent realization of a state-dependent signal, $\sigma^i \in \{h, l\}$.

The signal has the following simple structure: in state L,

$$\sigma^i = l \text{ with probability } 1 \quad \forall i$$

whereas in state H,

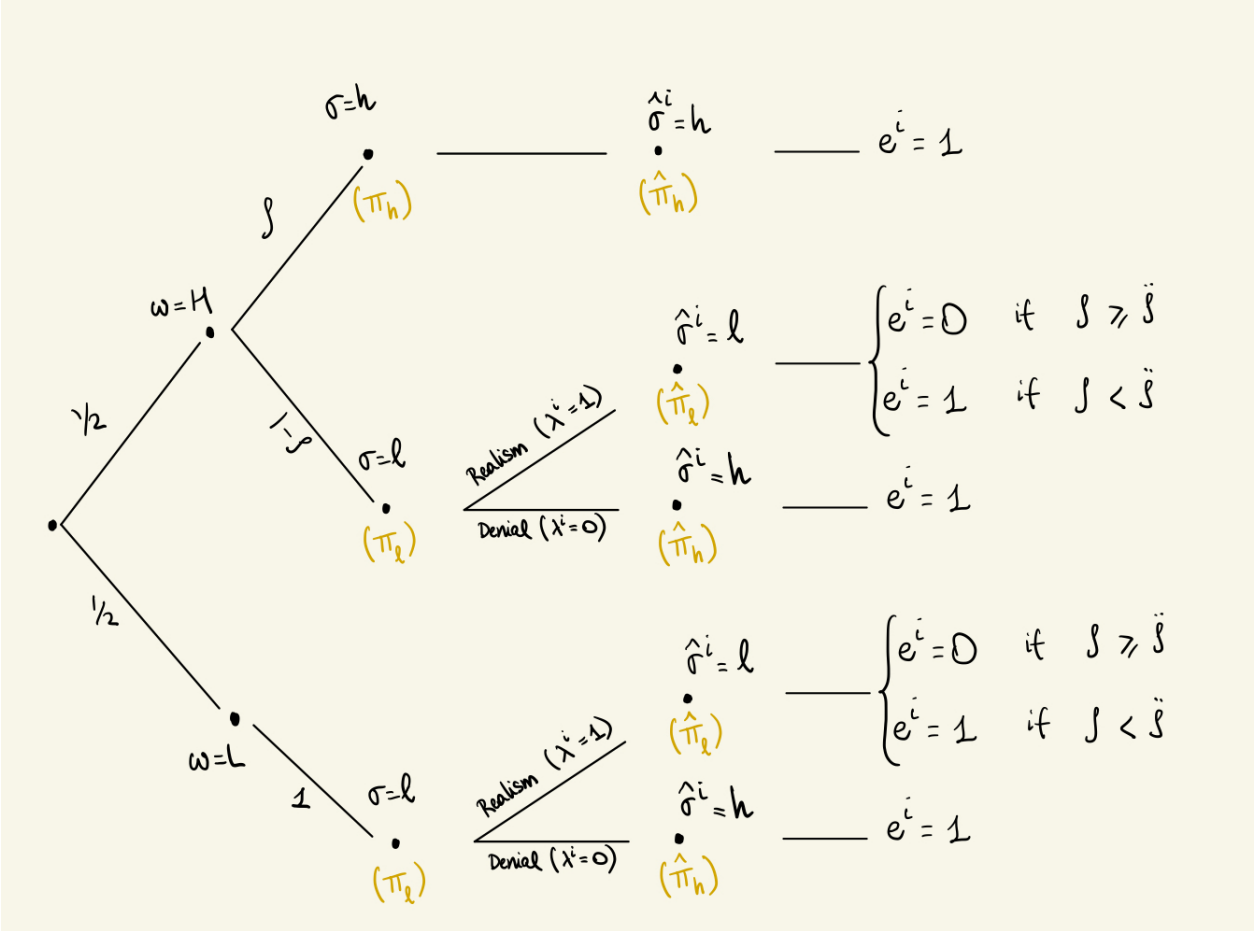
$$\sigma^i = \begin{cases} h & \text{with probability } \rho \\ l & \text{with probability } 1-\rho \end{cases} \quad \forall i$$

where $\rho \in [1/2, 1]$.

This simple structure ensures that if the observed news are good, $\sigma^i = h$, then the date-0 agent believes that the state is H for sure, while if the news are bad, the state may be H or L. Intuitively, less informative signals make bad news look not that bad, since there is a greater probability that the state is H.

This signal structure and the resulting inferences are shown in Figure 2.

Figure 2: Signal structure, posterior beliefs, and effort decisions



Information design In addition, we introduce an information design stage at $t = -1$. There is an information designer or principal ("she") of the organization whose aim is to maximize the ex-ante material payoffs –without regard to psychological anticipatory feelings– by choosing a signal informativeness, $\rho \in [1/2, 1]$. For example, we might think of a financial institution where its investors may ignore red flags (bad news) and engage in investments with negative expected value; the information designer wants to choose ρ such that it maximizes the expected value of the investment projects.

The information designer's payoff is defined as the average of the agents' material payoffs. Since we consider symmetric strategies by agents, the ex-ante material payoffs are equal across agents, and the information designer's problem yields the same solution as maximizing the ex-ante material payoff of an individual agent.

The choice of ρ by the information designer is credibly announced at $t = -1$ and it becomes common knowledge to all agents in the organization.

In sum, the timing of the game is summarized as follows:

- $t = -1$: The information designer chooses $\rho \in [1/2, 1]$ and credibly announces it to all agents.
- $t = 0$:
  - Nature chooses the state $\omega \in \{H, L\}$.
  - Given ω and ρ, each agent receives a signal $\sigma^i \in \{h, l\}$, realized independently across agents.
  - Given their signals σ^i, agents simultaneously choose their cognitive strategy $\lambda^i \in [0, 1]$, i.e., which signal $\hat{\sigma}^i$ to pass on to their date-1 selves.
- $t = 1$: Based on the recalled signal $\hat{\sigma}^i$, agent forms beliefs about the state of the world and chooses effort e^i. Agent experiences anticipatory feelings.
- $t = 2$: Final payoffs are realized.

Inferences Each agent i observes a realization of the signal σ^i in $t = 0$, and he recalls a signal $\hat{\sigma}^i$ in $t = 1$. The inferences made in each information set are as follows.

The posterior belief that the state is H upon observing a signal σ^i is denoted by $\pi_{\sigma^i} \equiv \Pr\left[\omega = H \mid \sigma^i\right]$. Upon receiving good news $\sigma^i = h$, agent i believes that the state is H for sure, since good news cannot be possibly produced in the state L:

$$\pi_h \equiv \Pr\left[\omega = H \mid \sigma^i = h\right] = 1 \tag{4}$$

The posterior belief upon receiving bad news is

$$\pi_l \equiv \Pr\left[\omega = H \mid \sigma^i = l\right] = \frac{1-\rho}{2-\rho} \tag{5}$$

The posterior belief that the state is H upon *recalling* a (potentially distorted) signal $\hat{\sigma}^i$ is denoted by $\hat{\pi}_{\hat{\sigma}^i} \equiv \Pr\left[\omega = H \mid \hat{\sigma}^i\right]$. If good news are recalled, the posterior belief is

$$\begin{aligned} \hat{\pi}_h &\equiv \Pr\left[\omega = H \mid \hat{\sigma}^i = h\right] \\ &= \frac{\left[\rho + \chi^i (1-\rho)\left(1-\lambda^i\right)\right] \cdot \frac{1}{2}}{\left[\rho + \chi^i (1-\rho)(1-\lambda^i)\right] \cdot \frac{1}{2} + \chi^i (1-\lambda^i) \cdot \frac{1}{2}} \\ &= \frac{\rho + \chi^i (1-\rho)\left(1-\lambda^i\right)}{\rho + \chi^i (2-\rho)(1-\lambda^i)} \end{aligned} \tag{6}$$

where $\chi^i \in [0, 1]$ is the cognitive sophistication of agent i (the degree to which the date-1 agent is aware of the possibility that date-0 agent might distort the signals), which ranges from 1 (full Bayesian) to 0 (full naivete). If $\chi^i = 0$, then the posterior belief simplifies to:

$$\hat{\pi}_h = 1 \tag{7}$$

Upon recalling bad news, $\hat{\sigma}^i = l$, the belief that the state is H is

$$\begin{aligned}\hat{\pi}_l(\rho) &\equiv \Pr\left[\omega = H \mid \hat{\sigma}^i = l\right] \\ &= \frac{\left[(1-\rho)\lambda^i\right]^{\frac{1}{2}}}{\left[(1-\rho)\lambda^i\right]^{\frac{1}{2}} + \left[1\cdot\lambda^i\right]^{\frac{1}{2}}} \\ &= \frac{1-\rho}{2-\rho}\end{aligned} \tag{8}$$

The inferences have the following properties:

1. $\hat{\pi}_l(\rho) = \pi_l(\rho) \quad \forall\rho$.
2. $\pi_l(\rho)$ and $\hat{\pi}_l(\rho)$ strictly decreasing in ρ: $\partial\pi_l/\partial\rho = \partial\hat{\pi}_l/\partial\rho < 0$.
3. $\pi_l(\rho)$ bounded below by $\pi_l(1) = 0$, and above by $\pi_l(1/2) = 1/3$.
4. Under naivete, $\pi_h = \hat{\pi}_h = 1$ for all ρ.

3.1.4 Assumptions

We make the following assumptions:

Assumption 1. *(Equiprobable states)* $q = 1/2$.

Assumption 2. *(Fully naive agents)* $\chi^i = 0$ *for all* i.

Assumption 3. *(Simplified utility functions)* $\alpha = 1/2$ *and* $s = \delta = 1$

Assumption 4. *(Groupthink)*

$$\theta_L < 0 < \theta_H$$

Assumption 5. *(Effort under full certainty of high value project)*

$$\frac{1}{2}\theta_H > c + m^i$$

The purpose of assumptions 1,2, and 3 is to simplify the analysis. In particular, assumption 2 has the effect of making incentives to deny stronger, making realism equilibrium (analyzed in section 5.1) harder to sustain[2]. Assumption 4 is made to focus the analysis on cases where collective delusions are harmful (groupthink) rather than beneficial (group morale); see Benabou 2013. Assumption 5 guarantees that if an agent is sure that the state is H (even through denial of bad news) then it is optimal for him to exert effort.

Under those assumptions, the utility functions simplify to

- Date-2 utility:

$$U_2^i = \theta\left[\frac{1}{2}e^i + \frac{1}{2}e^{-i}\right] \tag{9}$$

- Date-1 utility:

$$\begin{aligned}U_1^i &= -ce^i + 2\mathbb{E}_1^i\left[U_2^i\right] \\ &= -ce^i + \mathbb{E}_1^i\left[\theta e^i\right] + \mathbb{E}_1^i\left[\theta e^{-i}\right]\end{aligned} \tag{10}$$

[2]For a discussion of the implications of relaxing these assumptions, see section 6

- Date-0 utility:

$$\begin{aligned} U_0^i &= -m^i + \mathbb{E}_0^i\left[-ce^i + \mathbb{E}_1^i\left[U_2^i\right]\right] + \mathbb{E}_0^i\left[U_2^i\right] \\ &= -m^i + \mathbb{E}_0^i\left[-ce^i\right] + \mathbb{E}_1^i\left[U_2^i\right] + \mathbb{E}_0^i\left[U_2^i\right] \end{aligned} \tag{11}$$

where $\mathbb{E}_0\left[\mathbb{E}_1\left(.\right)\right] = \mathbb{E}_1\left(.\right)$ since the date-1 beliefs are perfectly anticipated by a date-0 agent. Given that U_2^i is linearly separable in efforts, for any $t = 0, 1$ the expectation of the date-2 payoff $\mathbb{E}_t\left[U_2^i\right]$ is a weighted sum of the expected contributions from i and from other agents:

$$\mathbb{E}_t^i\left[U_2^i\right] = \frac{1}{2}\mathbb{E}_t^i\left[\theta e^i\right] + \frac{1}{2}\mathbb{E}_t^i\left[\theta e^{-i}\right] \tag{12}$$

Then the date-0 utility (11) can be re-expressed as:

$$U_0^i = -m^i + \mathbb{E}_0^i\left[-ce^i\right] + \underbrace{\left[\frac{1}{2}\mathbb{E}_1^i\left[\theta e^i \mid \hat{\pi}^i\right] + \frac{1}{2}\mathbb{E}_1^i\left[\theta e^{-i} \mid \hat{\pi}^i\right]\right]}_{\text{Anticipatory feelings}} + \underbrace{\left[\frac{1}{2}\mathbb{E}_0^i\left[\theta e^i \mid \pi^i\right] + \frac{1}{2}\mathbb{E}_0^i\left[\theta e^{-i} \mid \pi^i\right]\right]}_{\text{Discounted material utility}} \tag{13}$$

3.2 Informativeness and date-1 effort decisions

In $t = 1$, each agent i maximizes U_1^i by choosing his effort level, given his posterior beliefs upon recalling news $\hat{\sigma}^i$. If he chooses effort $e^i = 1$, he expects to obtain $-c + \mathbb{E}_1^i\left[\theta\right] + \mathbb{E}_1^i\left[\theta e^{-i}\right]$, whereas if $e^i = 0$, he gets $\mathbb{E}_1^i\left[\theta e^{-i}\right]$. The net incentive for effort is then:

$$U_1^i\left(1\right) - U_1^i\left(0\right) = -c + \mathbb{E}_1^i\left[\theta\right]$$

Hence, i chooses effort whenever

$$\mathbb{E}_1^i\left[\theta\right] = \hat{\pi}^i\theta_H + \left(1 - \hat{\pi}^i\right)\theta_L > c \tag{14}$$

Therefore, the effort is determined only by the beliefs about θ that agent i carries. This depends on which signal $\hat{\sigma}^i$ is recalled in date 1.

On the one hand, the posterior belief upon recalling good news ($\hat{\sigma}^i = h$) is $\hat{\pi}_h = 1$ for all ρ. Since $\theta_H > c$, agents always work when they recall good news, for all $\rho \in [1/2, 1]$. On the other hand, the belief upon recalling bad news, $\hat{\pi}_l$, depends on ρ and is bounded above by $1/3$. Condition (14) implies that for agent i to exert effort upon recalling bad news the belief $\hat{\pi}_l$ should be sufficiently high, and equivalently, ρ sufficiently low:

$$\hat{\pi}_l \geq \frac{c - \theta_L}{\theta_H - \theta_L} \equiv \ddot{\pi} \tag{15}$$

$$\rho \leq \frac{\left(\theta_H - c\right) - \left(c - \theta_L\right)}{\theta_H - c} = 1 - \frac{c - \theta_L}{\theta_H - c} \equiv \ddot{\rho} \tag{16}$$

where the cutoff $\ddot{\pi}$ is defined as the belief π that satisfies the indifference condition $c = \mathbb{E}_1^i\left[\theta \mid \ddot{\pi}\right] = \ddot{\pi}\theta_H + \left(1 - \ddot{\pi}\right)\theta_L$, and the cutoff $\ddot{\rho}$ is the ρ such that $\pi_l\left(\ddot{\rho}\right) = \ddot{\pi}$.

Therefore, upon recalling bad news, agents exert effort whenever $\rho \leq \ddot{\rho}$, and do not work otherwise. Intuitively, noisier signals make it more likely that, given bad news, the state is H, and this makes it more likely to exert effort. Since optimal effort decisions depend on ρ, the analysis of equilibria will be separated in two cases, $\rho \geq \ddot{\rho}$ and $\rho < \ddot{\rho}$.

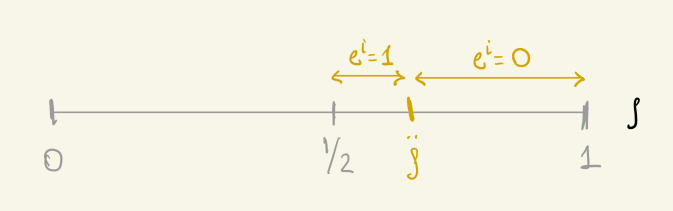
Given assumptions (4) and (5), the threshold $\ddot{\rho}$ has the following properties.

- $\ddot{\rho} < 1$ always

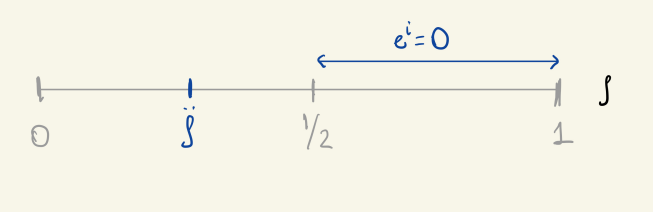
- $\ddot{\rho} < 1/2 \iff c > \mathbb{E}_1^i [\theta \mid \hat{\pi}_l = 1/3] \iff \theta_L < -\frac{1}{2}\theta_H + \frac{3}{2}c.$

For the derivations of these conditions, see Appendix 7.1. These relationships are illustrated in Figure 3.

Figure 3: Informativeness and effort



$\theta_L > -\theta_H + 3c$



$\theta_L < -\theta_H + 3c$

For θ_L sufficiently small in absolute value, the belief about the value of the project θ can be higher than the cost of effort for some ρ. Since lower ρ raises the belief $\hat{\pi}_l$ that the state is H, the agent works for all ρ sufficiently small. However, if θ_L is sufficiently large in absolute value, the belief about θ is guaranteed to be lower than c for all $\rho \in [1/2, 1]$; so agents do not work for any $\rho \in [1/2, 1]$.

4 Benchmark model with no self-delusion

We now turn to the analysis of the optimal signal informativeness. In this section, we begin by considering the standard case where agents have perfect recall, so they cannot delude themselves (Case 1 of the model). In the next section, we consider the model with the possibility of delusion, our main focus of analysis.

Under perfect recall, all agents in the organization perfectly remember the date-0 news. Formally, this imposes the restriction $\hat{\sigma}^i = \sigma^i \ \ \forall i$. When $\sigma^i = h$, then $\hat{\pi}^i = 1$ and agent i works. When $\sigma^i = l$, then $e^i = 0$ for $\rho \geq \ddot{\rho}$ and $e^i = 1$ otherwise.

Consider the ex-ante material utility of the information designer. For $\rho \geq \ddot{\rho}$, agents only work when the signal is good. The state is expected to be H with prior probability $1/2$. In that state, a fraction ρ of agents observes a good signal and works, obtaining $\rho(\theta_H - c)$ from them; and a fraction $1 - \rho$ gets the bad signal and does not work, obtaining 0 from them. With probability $1/2$ the state is L, where all agents get a bad signal for sure; then nobody works, and the material payoff is 0. For $\rho < \ddot{\rho}$, agents always work, for both signals; so in every state $\omega = H, L$ the information designer expects to receive $(\theta_\omega - c)$ since all agents always work. Hence, the ex-ante material payoff is given by

$$U^{Mat}_{Bench}(\rho) = \begin{cases} \frac{1}{2}\rho(\theta_H - c) & \text{for } \rho \geq \ddot{\rho} \\ \frac{1}{2}(\theta_H + \theta_L) - c & \text{for } \rho < \ddot{\rho} \end{cases}$$

Proposition 6. *If agents have perfect recall ($\hat{\sigma}^i = \sigma^i \ \ \forall i$), then the information designer's ex-ante material payoffs are maximized at $\rho^*_{Bench} = 1$.*

Proof. See section 7.1. □

For $\rho \geq \ddot{\rho}$, the ex-ante material utility of the information designer is strictly increasing in ρ: more information is better. In addition, it can be shown that for any $\rho \geq \ddot{\rho}$, the utility is higher than for any $\rho' < \ddot{\rho}$. Therefore, the utility for the information designer is maximized at $\rho^*_{Bench} = 1$.

If agents cannot delude themselves, perfect informativeness is optimal. This is what standard economic theory would suggest: information has a decision-value as it improves decision-making. Intuitively, if agents treat information realistically (and work only when they believe the state is H, when it is appropriate do to so), then the information designer does not want agents to miss valuable opportunities when the state is good and the projects have positive expected value.

5 Model with self-delusion

Now we turn to analyze the optimal signal informativeness in the case where agents may engage in delusion. In subsection 5.1, we determine the equilibria of this game and show that a) when delusion is possible, for some ρ a delusion equilibrium may emerge, and b) when it exists, a realism equilbrium requires sufficiently high ρ. Importantly, we highlight a trade-off between correct effort decisions and higher temptations for delusion from anticipatory feelings that is involved in the choice of ρ. In subsection 5.2, the information designer's optimal choice when delusion by agents is possible is analyzed.

5.1 Equilibria

The equilibrium concept we use is that of Benabou (2013): an **intra-personal equilibrium** for individual i is a Perfect Bayesian equilibrium in the game between his date-0 and date-1 selves. This equilibrium concept treats the individual i at date-0 and at date-1 as if they were two different players in a self-signalling game.

We will focus on equilibria with pure and symmetric strategies.

To find equilibria, we look at the incentives of date-0 agents to realistically pass bad news to their date-1 selves or to delude themselves and carry overoptimistic beliefs. The agent maximizes his date-0 utility

$$U_0^i = -m^i + \mathbb{E}_0^i\left[-ce^i\right] + \underbrace{\left[\frac{1}{2}\mathbb{E}_1^i\left[\theta e^i \mid \hat{\pi}^i\right] + \frac{1}{2}\mathbb{E}_1^i\left[\theta e^{-i} \mid \hat{\pi}^i\right]\right]}_{\text{Anticipatory feelings}} + \underbrace{\left[\frac{1}{2}\mathbb{E}_0^i\left[\theta e^i \mid \pi^i\right] + \frac{1}{2}\mathbb{E}_0^i\left[\theta e^{-i} \mid \pi^i\right]\right]}_{\text{Discounted Material utility}}$$

by choosing the probability of engaging in Realism, $\lambda^i \equiv \Pr\left[\hat{\sigma}^i = l \mid \sigma = l\right]$ –i.e., correctly encoding the signal as l when it is l.

The optimal date-1 effort decisions depend on whether $\rho \geq \ddot{\rho}$, so the analysis of equilibria is divided in two cases.

5.1.1 Range $\rho \geq \ddot{\rho}$

When $\rho \geq \ddot{\rho}$, agents make effort only when they recall good news. Since $\ddot{\rho} < 1$, the range $[\ddot{\rho}, 1]$ is always non-empty.

Delusion equilibrium Consider first a delusion equilibrium, one in which all agents deny bad news ($\lambda^{-i} = 0$).

Since $\rho \geq \ddot{\rho}$, if agent i chooses realism, he would not work in date-1, while all other agents work. Hence, his utility from realism is:

$$\begin{aligned} U^i_{0,R} = & \underbrace{\frac{1}{2}\mathbb{E}^i_1\left[\theta e^{-i} \mid \hat{\pi}_l\right]}_{\text{Anticipatory feelings}(\hat{\pi}_l)} + \underbrace{\frac{1}{2}\mathbb{E}^i_0\left[\theta e^{-i} \mid \pi_l\right]}_{\text{Standard discounting}} \\ & = \frac{1}{2}\mathbb{E}^i_1\left[\theta \mid \hat{\pi}_l\right] + \frac{1}{2}\mathbb{E}^i_0\left[\theta \mid \pi_l\right] \\ & = \left(\frac{1-\rho}{2-\rho}\right)\theta_H + \left(\frac{1}{2-\rho}\right)\theta_L = \mathbb{E}^i_0\left[\theta \mid \pi_l\right] \end{aligned} \tag{17}$$

[3]. Since in a delusion equilibrium everybody else has optimistic date-1 beliefs $\hat{\pi}_h$ –either from having observed a high signal, or from a low signal being distorted–, the average effort by other agents is $e^{-i} = 1$. Therefore, agent i expects a full contribution of others to the project, so $1/2 \cdot \mathbb{E}_t\left[\theta e^{-i}\right] = 1/2 \cdot \mathbb{E}_t\left[\theta\right]$.

If i chooses to deny the news, then he carries optimistic date-1 beliefs, works, and pays the mental and effort costs:

$$\begin{aligned} U^i_{0,D} = & -m^i - c + \underbrace{\frac{1}{2}\mathbb{E}^i_1\left[\theta \mid \hat{\pi}_h\right] + \frac{1}{2}\mathbb{E}^i_1\left[\theta e^{-i} \mid \hat{\pi}_h\right]}_{\text{Anticipatory feelings } (\hat{\pi}_h)} + \underbrace{\frac{1}{2}\mathbb{E}^i_0\left[\theta \mid \pi_l\right] + \frac{1}{2}\mathbb{E}^i_0\left[\theta e^{-i} \mid \pi_l\right]}_{\text{Standard discounting}} \\ & = -m^i - c + \frac{1}{2}\theta_H + \frac{1}{2}\theta_H + \frac{1}{2}\mathbb{E}^i_0\left[\theta \mid \pi_l\right] + \frac{1}{2}\mathbb{E}^i_0\left[\theta \mid \pi_l\right] \\ & = -m^i - c + \theta_H + \left[\frac{1-\rho}{2-\rho}\theta_H + \frac{1}{2-\rho}\theta_L\right] \end{aligned} \tag{18}$$

The net incentive for delusion is given by

$$\begin{aligned} \Psi\left(\lambda^{-i} = 0\right) = & -m^i - c + \underbrace{\frac{1}{2}\mathbb{E}^i_1\left[\theta \mid \hat{\pi}_h\right] + \frac{1}{2}\mathbb{E}^i_1\left[\theta e^{-i} \mid \hat{\pi}_h\right] - \frac{1}{2}\mathbb{E}^i_1\left[\theta e^{-i} \mid \hat{\pi}_l\right]}_{\text{Gain in anticipatory feelings}} + \underbrace{\frac{1}{2}\mathbb{E}^i_0\left[\theta \mid \pi_l\right]}_{\text{Gain/Loss in standard}} \\ & = \theta_H - m^i - c \end{aligned} \tag{19}$$

When others are deluding themselves, the informativeness of the signal has no effect in the incentive for delusion, because of a couple of reasons. First, the optimistic belief $\hat{\pi}_h$ does not depend on ρ (a consequence of naivete). Second, in a delusion equilibrium where $e^{-i} = 1$, the last two terms are equal and end up cancelling out. This is a particular result that depends on the fact that $\pi_l = \hat{\pi}_l$ or on $\alpha = 1/2$ and $s = \delta$, so that the discounted expected contributions of agent i and of other agents $-i$ are equal. As a result, ρ does not affect the incentives for delusion in a delusion equilibrium when $\rho \geq \ddot{\rho}$.[4]

Delusion is an equilibrium if $\Psi\left(\lambda^{-i} = 0\right) \geq 0$. This is always the case, given that $\frac{1}{2}\theta_H - m^i - c > 0$ by assumption (5).

Lemma 7. *A delusion equilibrium exists for all $\rho \in [\ddot{\rho}, 1]$.*

Realism equilibrium Now consider a realism equilibrium: one in which everybody selects $\lambda^j = 1$. In this equilibrium, agents who receive bad signals ($\sigma^i = l$) recall them correctly (realist equilibrium) and then do not work (since $\rho \geq \ddot{\rho}$). Hence, the only ones who work are a proportion ρ who get a good signal in the H state.

[3]Detailed derivations of expressions that are not part of propositions and lemmas, like this one, are available for reference in section 7.2

[4]We provide a discussion of the effect of relaxing some of the assumptions and the possibility of affecting Ψ through manipulation of ρ in section 6.

Consider agent i's incentives to deviate. If he chooses realism, he encodes $\hat{\sigma}^i = l$ and does not exert effort ($\rho \geq \ddot{\rho}$). Therefore, all his utility comes from the expectations of others' contributions.

$$\begin{aligned} U^i_{0,R} = & \underbrace{\frac{1}{2}\mathbb{E}^i_1\left[\theta e^{-i} \mid \hat{\pi}_l\right]}_{\text{Anticipatory feelings}(\hat{\pi}_l)} + \underbrace{\frac{1}{2}\mathbb{E}^i_0\left[\theta e^{-i} \mid \pi_l\right]}_{\text{Standard discounting}} \\ & = \frac{1}{2}\hat{\pi}_l \rho \theta_H + \frac{1}{2}\pi_l \rho \theta_H \\ & = \left(\frac{1-\rho}{2-\rho}\right)\rho\theta_H \end{aligned} \tag{20}$$

The expectation $\mathbb{E}_t\left[\theta e^{-i} \mid \hat{\pi}_l\right]$ is computed in a similar way as before. For the posterior belief $\hat{\pi}_l$, the state is expected to be H with probability $\hat{\pi}_l$, where a proportion ρ of other agents gets a good signal and then works while a proportion $1-\rho$ obtains a bad signal, among which a fraction $1-\lambda^{-i}$ denies it and then works. In the state L, everybody gets bad news but a proportion $1-\lambda^{-i}$ engages in delusion and still works. In a realism equilibrium, $\lambda^{-i} = 1$, so there are no deniers among other agents, and the only ones who are expected to work are a proportion ρ in the H state. Hence, $\mathbb{E}_t\left[\theta e^{-i} \mid \hat{\pi}_l\right] = \hat{\pi}_l \rho \theta_H$.

If instead he chooses to delude himself, then he encodes $\hat{\sigma}^i = h$ and then engages in work. The utility he gets is

$$U^i_{0,D} = -m^i - c + \underbrace{\frac{1}{2}\mathbb{E}^i_1\left[\theta \mid \hat{\pi}_h\right] + \frac{1}{2}\mathbb{E}^i_1\left[\theta e^{-i} \mid \hat{\pi}_h\right]}_{\text{Anticipatory feelings } (\hat{\pi}_h)} + \underbrace{\frac{1}{2}\mathbb{E}^i_0\left[\theta \mid \pi_l\right] + \frac{1}{2}\mathbb{E}^i_0\left[\theta e^{-i} \mid \pi_l\right]}_{\text{Standard discounting}} \tag{21}$$

The net incentive for denial is $\Psi = U_{0,D} - U_{0,R}$:

$$\Psi\left(\lambda^{-i}=1\right) = -m^i - c + \underbrace{\frac{1}{2}\mathbb{E}^i_1\left[\theta \mid \hat{\pi}_h\right] + \frac{1}{2}\mathbb{E}^i_1\left[\theta e^{-i} \mid \hat{\pi}_h\right] - \frac{1}{2}\mathbb{E}^i_1\left[\theta e^{-i} \mid \hat{\pi}_l\right]}_{\text{Gain in anticipatory feelings}} + \underbrace{\frac{1}{2}\mathbb{E}^i_0\left[\theta \mid \pi_l\right]}_{\text{Gain/Loss in standard}} \tag{22}$$

$$\begin{aligned} & = -m^i - c + \frac{1}{2}\theta_H + \frac{1}{2}\rho\theta_H - \frac{1}{2}\hat{\pi}_l\rho\theta_H + \frac{1}{2}\left[\pi_l\theta_H + (1-\pi_l)\,\theta_L\right] \\ & = \left(\frac{1}{2}\theta_H - m^i - c\right) + \left(\frac{1}{2-\rho}\right)\left[\frac{1}{2}\theta_H + \frac{1}{2}\theta_L\right] \end{aligned} \tag{23}$$

Expression (22) shows the trade-offs involved in delusion. On the one hand, Delusion gives a boost in anticipatory feelings, and this boost comes from two sources: i engages in extra effort due to overoptimistic beliefs, adding a positive anticipatory utility of $1/2 \cdot \mathbb{E}^i_1\left[\theta \mid \hat{\pi}_h\right]$; in addition, others' effort is fixed but, by improving beliefs from $\hat{\pi}_l$ to $\hat{\pi}_h$, a gain of $1/2\left[\mathbb{E}^i_1\left[\theta e^{-i} \mid \hat{\pi}_h\right] - \mathbb{E}^i_1\left[\theta e^{-i} \mid \hat{\pi}_l\right]\right]$ is obtained. But on the other hand, there is a decision cost from i mistakenly exerting effort based on too high beliefs. Delusion also entails incurring the mental and the effort costs.

In expression (23), assumption 5 implies that the boost in anticipatory feelings about i's own contribution ($\frac{1}{2}\mathbb{E}^i_1\left[\theta \mid \hat{\pi}_h\right] = \frac{1}{2}\theta_H$) is more than enough to compensate for the mental cost of delusion m^i and of the induced effort c; and none of these depends on ρ. Then, for realism to be an equilibrium ($\Psi \leq 0$), the remaining terms must be negative enough: the losses from i's mistaken effort decision $\left(\frac{1}{2}\mathbb{E}^i_0\left[\theta \mid \pi_l\right]\right)$ must *at least* be stronger than the gains in anticipatory feelings about others' contributions $\left(\frac{1}{2}\mathbb{E}^i_1\left[\theta e^{-i} \mid \hat{\pi}_h\right] - \frac{1}{2}\mathbb{E}^i_1\left[\theta e^{-i} \mid \hat{\pi}_l\right]\right)$. As the second term in (23) shows, this depends on whether $\theta_L + \theta_H < 0$. Whenever $\theta_L < -\theta_H$ (i.e., whenever the project has a very negative downside value), the loss from i's mistake dominates over the anticipatory feelings from others' fixed contributions. In such case, the net incentive for denial *decreases* with ρ: for agent i's incentives, the gains in decision-making are more important than improvements in anticipatory feelings, so more informative signals make realism equilibrium more likely. However, if $\theta_L > -\theta_H$, the anticipatory feelings from others' contributions

dominate over the loss from own mistakes, and the incentive for delusion of agent i *increases* with ρ: providing *less* information makes the realism equilibrium more likely. This trade-off between better anticipatory feelings and making right decisions shapes how ρ changes the incentives for denial:

$$\frac{\partial \Psi}{\partial \rho} = \frac{1}{2}\frac{(\theta_H + \theta_L)}{(2-\rho)^2} \begin{cases} > 0 & \text{if } \theta_L > -\theta_H \\ < 0 & \text{if } \theta_L < -\theta_H \end{cases} \tag{24}$$

From the discussion above, it is clear that for realism to be a best response ($\Psi \leq 0$), $\theta_L < -\theta_H$ is not sufficient: θ_L should be even more negative to compensate any excedent from the first term. Given such a sufficiently negative θ_L, if there exists a $\rho < 1$ such that a realism equilibrium exists, then by $\partial\Psi/\partial\rho < 0$ realism is an equilibrium for any $\rho' > \rho$. So the equilibrium condition for ρ is defined by a threshold $\bar{\rho}_R$. These intuitions are formalized in the following lemmas.

Lemma 8. *For $\rho \geq \ddot{\rho}$, a realism equilibrium exists if and only if:*

$$\rho \geq 2 + \frac{\frac{1}{2}(\theta_H + \theta_L)}{\frac{1}{2}\theta_H - m^i - c} \equiv \bar{\rho}_R\left(\theta_H, \theta_L, m^i, c\right) \tag{25}$$

The following lemma characterizes this threshold $\bar{\rho}_R$, as a function of θ_L.

Lemma 9. *The threshold $\bar{\rho}_R$ has the following properties:*

1. $\bar{\rho}_R < 1 \iff \theta_L < -\theta_H - 2\left(\frac{1}{2}\theta_H - m^i - c\right)$.
2. $\bar{\rho}_R < 1/2 \iff \theta_L < -\theta_H - 3\left(\frac{1}{2}\theta_H - m^i - c\right)$.
3. $\bar{\rho}_R > \ddot{\rho} \iff \theta_L > -\theta_H \left[2(\theta_H - c) - \left(2m^i + c\right)\right] / \left[2m^i + c\right]$
   with $-\theta_H \left[2(\theta_H - c) - \left(2m^i + c\right)\right] / \left[2m^i + c\right] < -\theta_H - 3\left(\frac{1}{2}\theta_H - m^i - c\right)$
4. *If* $\bar{\rho}_R > 1/2 \implies \bar{\rho}_R > \ddot{\rho}$.

Therefore, if

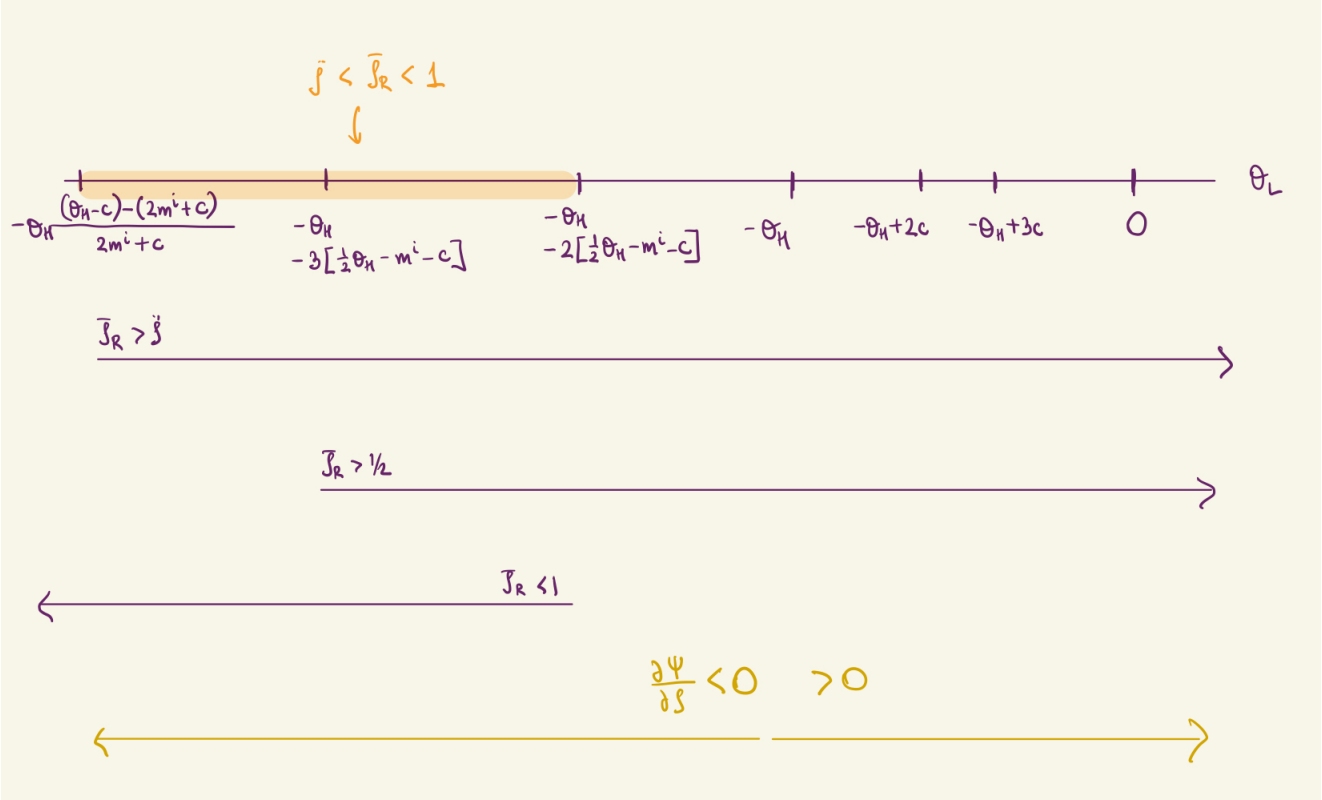
$$\theta_L \in \left(-\theta_H \frac{2(\theta_H - c) - \left(2m^i + c\right)}{2m^i + c}, -\theta_H - 2\left(\frac{1}{2}\theta_H - m^i - c\right)\right)$$

$\implies \ddot{\rho} < \bar{\rho}_R < 1 \implies$ *a non-empty range $\rho \in [\bar{\rho}_R, 1]$ that supports a realism equilibrium is guaranteed to exist.*

Proof. See section 7.1. □

The lemma is depicted in the Figure below.

Figure 4: θ_L and realism equilibrium ($\rho \geq \ddot{\rho}$)



5.1.2 Range $\rho < \ddot{\rho}$

When $\rho < \ddot{\rho}$, agents always work (upon recalling both good or bad news).

Delusion equilibrium Suppose other agents are engaging in delusion, and consider the best response by agent i. Since $\rho < \ddot{\rho}$, if i chooses realism, his utility now also includes his own contribution to the project and the cost c, despite recalling bad news. Since this is a delusion equilibrium, all other agents are working so $\mathbb{E}\left[\theta e^{-i}\right] = \mathbb{E}\left[\theta\right]$.

$$\begin{aligned} U^i_{0,R} &= -c + \underbrace{\frac{1}{2}\mathbb{E}^i_1\left[\theta \mid \hat{\pi}_l\right] + \frac{1}{2}\mathbb{E}^i_1\left[\theta e^{-i} \mid \hat{\pi}_l\right]}_{\text{Anticipatory feelings}(\hat{\pi}_l)} + \underbrace{\frac{1}{2}\mathbb{E}^i_0\left[\theta \mid \pi_l\right] + \frac{1}{2}\mathbb{E}^i_0\left[\theta e^{-i} \mid \pi_l\right]}_{\text{Standard discounting}} \\ &= -c + \frac{1}{2}\mathbb{E}^i_1\left[\theta \mid \hat{\pi}_l\right] + \frac{1}{2}\mathbb{E}^i_1\left[\theta \mid \hat{\pi}_l\right] + \frac{1}{2}\mathbb{E}^i_0\left[\theta \mid \pi_l\right] + \frac{1}{2}\mathbb{E}^i_0\left[\theta \mid \pi_l\right] \\ &= -c + 2\left[\left(\frac{1-\rho}{2-\rho}\right)\theta_H + \left(\frac{1}{2-\rho}\right)\theta_L\right] \end{aligned} \tag{26}$$

If i engages in delusion, his expected payoff remains the same as when $\rho \geq \ddot{\rho}$, since he does not

reach information set $\hat{\sigma}^i = l$ (where the condition $\rho < \ddot{\rho}$ makes a difference).

$$U^i_{0,D} = -m^i - c + \underbrace{\frac{1}{2}\mathbb{E}^i_1\left[\theta \mid \hat{\pi}_h\right] + \frac{1}{2}\mathbb{E}^i_1\left[\theta e^{-i} \mid \hat{\pi}_h\right]}_{\text{Anticipatory feelings } (\hat{\pi}_h)} + \underbrace{\frac{1}{2}\mathbb{E}^i_0\left[\theta \mid \pi_l\right] + \frac{1}{2}\mathbb{E}^i_0\left[\theta e^{-i} \mid \pi_l\right]}_{\text{Standard discounting}} \tag{27}$$

$$= -m^i - c + \frac{1}{2}\theta_H + \frac{1}{2}\theta_H + \frac{1}{2}\mathbb{E}^i_0\left[\theta \mid \pi_l\right] + \frac{1}{2}\mathbb{E}^i_0\left[\theta \mid \pi_l\right]$$

$$= -m^i - c + \theta_H + \left[\frac{1-\rho}{2-\rho}\theta_H + \frac{1}{2-\rho}\theta_L\right]$$

The net incentive for delusion is then:

$$\Psi\left(\lambda^{-i} = 0\right) = -m^i + \underbrace{\frac{1}{2}\left(\mathbb{E}^i_1\left[\theta \mid \hat{\pi}_h\right] - \mathbb{E}^i_1\left[\theta \mid \hat{\pi}_l\right]\right) + \frac{1}{2}\left(\mathbb{E}^i_1\left[\theta e^{-i} \mid \hat{\pi}_h\right] - \mathbb{E}^i_1\left[\theta e^{-i} \mid \hat{\pi}_l\right]\right)}_{\text{Gain in anticipatory feelings}}$$

$$= \frac{1}{2-\rho}\left(\theta_H - \theta_L\right) - m^i \tag{28}$$

Since agent i works anyways, there are no gains in decision making by engaging in realism. The only source of incentives for delusion come from anticipatory feelings.

Delusion is a best response (and hence an equilibrium) if $\Psi \geq 0$. If $m^i = 0$, then delusion is always a best response, for any $\rho < \ddot{\rho}$, since $\theta_H > \theta_L$. If $m^i > 0$, then delusion is a best response if and only if ρ is sufficiently high:

$$\rho \geq 2 - \frac{(\theta_H + \theta_L)}{m^i} \equiv \bar{\rho}_B \tag{29}$$

Assumptions 4 and 5 imply that $\bar{\rho}_B < 1/2$. A denial equilibrium always exists for all the range $\rho < \ddot{\rho}$.

Lemma 10. *For $\rho < \ddot{\rho}$, a delusion equilibrium always exists, for all ρ.*
More specifically,

- *Delusion is an equilibrium if there are no mental costs, $m^i = 0$; or, when $m^i > 0$, if*

$$\rho \geq 2 - \frac{(\theta_H - \theta_L)}{m^i} \equiv \bar{\rho}_B \tag{30}$$

- $\bar{\rho}_B < 1/2$

Proof. See section 7.1. □

Realism equilibrium In a realism equilibrium, agents who receive bad signals encode them correctly, however given $\rho \leq \ddot{\rho}$ they do work. Then everybody exerts effort, regardless of which beliefs they have in date 1.

Since all agents in the organization work, the only incentive for delusion comes from anticipatory feelings.

If i engages in realism, he obtains:

$$U^i_{0,R} = -c + \underbrace{\frac{1}{2}\mathbb{E}^i_1\left[\theta \mid \hat{\pi}_l\right] + \frac{1}{2}\mathbb{E}^i_1\left[\theta e^{-i} \mid \hat{\pi}_l\right]}_{\text{Anticipatory feelings}(\hat{\pi}_l)} + \underbrace{\frac{1}{2}\mathbb{E}^i_0\left[\theta \mid \pi_l\right] + \frac{1}{2}\mathbb{E}^i_0\left[\theta e^{-i} \mid \pi_l\right]}_{\text{Standard discounting}} \tag{31}$$

$$= -c + \mathbb{E}_1\left[\theta \mid \hat{\pi}_l\right] + \mathbb{E}_0\left[\theta \mid \pi_l\right]$$

$$= -c + 2\left[\left(\frac{1-\rho}{2-\rho}\right)\theta_H + \left(\frac{1}{2-\rho}\right)\theta_L\right]$$

The utility of delusion is:

$$
\begin{aligned}
U^i_{0,D} &= -m^i - c + \underbrace{\frac{1}{2}\mathbb{E}^i_1\left[\theta \mid \hat{\pi}_h\right] + \frac{1}{2}\mathbb{E}^i_1\left[\theta e^{-i} \mid \hat{\pi}_h\right]}_{\text{Anticipatory feelings } (\hat{\pi}_h)} + \underbrace{\frac{1}{2}\mathbb{E}^i_0\left[\theta \mid \pi_l\right] + \frac{1}{2}\mathbb{E}^i_0\left[\theta e^{-i} \mid \pi_l\right]}_{\text{Standard discounting}} \qquad (32)\\
&= -m^i - c + \mathbb{E}_1\left[\theta \mid \hat{\pi}_h\right] + \mathbb{E}_0\left[\theta \mid \pi_l\right]\\
&= -m^i - c + \theta_H + \left[\frac{1-\rho}{2-\rho}\theta_H + \frac{1}{2-\rho}\theta_L\right]
\end{aligned}
$$

The net incentive for delusion

$$
\begin{aligned}
\Psi\left(\lambda^{-i}=1\right) &= -m^i + \underbrace{\frac{1}{2}\left(\mathbb{E}^i_1\left[\theta \mid \hat{\pi}_h\right] - \mathbb{E}^i_1\left[\theta \mid \hat{\pi}_l\right]\right) + \frac{1}{2}\left(\mathbb{E}^i_1\left[\theta e^{-i} \mid \hat{\pi}_h\right] - \mathbb{E}^i_1\left[\theta e^{-i} \mid \hat{\pi}_l\right]\right)}_{\text{Gain in anticipatory feelings}}\\
&= \frac{1}{2-\rho}\left(\theta_H - \theta_L\right) - m^i
\end{aligned}
$$

Realism is a best response to realism if $\Psi\left(\lambda^{-i}=1\right) \leq 0$. However, this is never the case. If $m^i = 0$, then $\Psi > 0$. If $m^i > 0$, $\Psi \leq 0$ requires

$$\rho \leq \bar{\rho}_B \equiv 2 - \frac{(\theta_H - \theta_L)}{m^i} \qquad (33)$$

But assumptions 4 and 5 imply that $\bar{\rho}_B < 1/2$. Realism cannot be an equilibrium for any $\rho < \ddot{\rho}$.

Intuitively, when $\rho < \ddot{\rho}$, all agents work no matter the signals, so the errors in decision making are the same across strategies. However, delusion entails a gain in anticipatory feelings with respect to realism. The only thing that could possibly counteract this incentive is a sufficiently high mental cost. However, the mental cost has an upper bound, $m^i < 1/2 \cdot \theta_H - c$. In addition, θ_L has a bound as well, since $\theta_L < 0$. These restrictions imply that under no parameter combination $\bar{\rho}_B \geq 1/2$. Therefore a realism equilibrium never exists.

Lemma 11. *Realism is never an equilibrium for $\rho < \ddot{\rho}$.*

5.2 Optimal signal informativeness

Now the problem of the information designer in $t = -1$ is analyzed in this section.

We have shown that if $\rho \geq \ddot{\rho}$, for some combinations of parameters and some values of ρ, there might exist multiple equilibria (both realism and delusion), which raises the question of equilibrium selection. In such cases, as is usual in the literature of information design, we will assume that the information designer can select the equilibrium most favorable to her among those attainable by her chosen signal structure. This amounts to assuming that the realism equilibrium prevails (since, as we will show below in proposition 12, realism is preferred by the information designer), so the analysis will focus on the maximization of ex-ante material payoffs rather than equilibrium selection.[5]

5.2.1 Ex-ante material payoffs of collective realism and collective delusion

We first show that, independently of whether it can be sustained as equilibrium or not, a situation of collective realism (where $\lambda^i = 1 \ \ \forall i$) delivers higher ex-ante material payoffs than a situation of collective delusion (where $\lambda^i = 0 \ \ \forall i$).

[5] For a discussion about the delusion equilibrium, see section 6.

Proposition 12. *For any given ρ, the ex-ante material payoff of a central planner is higher in a situation of collective realism than under collective delusion. In particular,*

- *For $\rho > \ddot{\rho}$, collective realism delivers strictly higher expected material payoffs:*

$$U_{-1,R}^{Mat}(\rho) = \frac{1}{2}\rho(\theta_H - c) > \frac{1}{2}(\theta_H + \theta_L) - c = U_{-1,D}^{Mat}(\rho) \tag{34}$$

- *For $\rho \leq \ddot{\rho}$, collective realism delivers exactly the same expected material payoffs as collective denial:*

$$U_{-1,R}^{Mat}(\rho) = \frac{1}{2}(\theta_H + \theta_L) - c = U_{-1,D}^{Mat}(\rho)$$

Proof. See section 7.1. □

The result for $\rho < \ddot{\rho}$ implies that even if (by relaxing some assumptions) a realism equilibrium existed in the range $\rho < \ddot{\rho}$, there would be no strict gains for the information designer relative to the delusion equilibrium. However, overall, the proposition shows that an organization designer is *at least* indifferent by choosing realism, and sometimes strictly better. Collective realism is always (weakly) better, for all ρ. The ex-ante material payoffs of collective denial, on the other hand, are independent of ρ.

5.2.2 Optimal choice by the information designer

For $\rho < \ddot{\rho}$, there is never a realism equilibrium. For $\rho \geq \ddot{\rho}$, a realism equilibrium may exist. Since for $\rho \geq \ddot{\rho}$ the ex-ante payoff of a realism equilibrium is strictly increasing in ρ, the optimal choice for the CP is the maximum ρ such that realism is an equilibrium. By Lemmas (8) and (9), when a realism equilibrium exists, the optimal choice is $\rho^* = 1$. Otherwise, delusion is the only equilibrium, with payoffs that are constant in ρ.

Proposition 13. *If $\theta_L \leq -\theta_H - 2\left(\frac{1}{2}\theta_H - m^i - c\right)$, then the ex-ante material utility is maximized at $\rho^* = 1$ in a realism equilibrium, with a payoff of $\frac{1}{2}(\theta_H - c)$. Otherwise, delusion is the only equilibrium and the information designer is indifferent between any $\rho \in [1/2, 1]$, obtaining a payoff of $\frac{1}{2}(\theta_H + \theta_L) - c$.*

Therefore, we find that there exist at least some combinations of parameters such that a realism equilibrium exists and the optimal choice by the information designer is to provide perfectly informative signals. More specifically, when the downside value is very negative (e.g., disastrous events, "black swans"), then it is in the best interest of the information designer to provide perfect information. This confirms the standard result of the literature, and suggests that, in our specification, anticipatory feelings are not strong enough to overcome the decision-value of the information.

However, it remains to study whether this result, which may follow from the particular assumptions made, is general. In particular, assumption 5 seems to be crucial in driving this result. As discussed in section 5.1, it ensures that any realism equilibrium require $\theta_L < -\theta_H$ (see expression 23). But if this is the case, in the trade-off in the incentives for delusion the losses from own mistakes dominate over the gains in anticipatory feelings, and $\partial\Psi/\partial\rho < 0$, implying that the condition for realism equilibrium takes the form $\rho \geq \bar{\rho}_R$. However, if the assumption were relaxed so that $\frac{1}{2}\theta_H - m^i - c < 0$ (for example, if the mental cost m^i were large enough), then the condition $\Psi \leq 0$ would require a less negative $\theta_L < -\theta_H + k$, for some $k > 0$. Then for a non-empty range $\theta_L \in (-\theta_H, -\theta_H + k)$, anticipatory feelings would dominate in the trade-off and $\partial\Psi/\partial\rho > 0$, reverting the inequality of the equilibrium condition to $\rho \leq \bar{\rho}'_R$ (ρ sufficiently *low*), where $\bar{\rho}'_R < 1$ (by Lemma 9).[6] If this were the case, the

[6] In fact, the direction of the inequality in this new realism equilibrium condition would be precisely the same of the realism equilibrium condition (33) for $\rho < \ddot{\rho}$, where *only* anticipatory feelings matter.

information designer would want to introduce noise in the signals to mitigate the gains in anticipatory feelings so that incentives for delusion are not too strong, and ensure that realism (which offers strictly higher payoffs, as in Proposition 12) is sustained as an equilibrium. Since ex-ante payoffs are maximized at the maximum ρ such that realism is an equilibrium, $\rho^* = \bar{\rho}'_R < 1$. Therefore, it may be possible that for some combinations of parameter values the optimal choice is less than perfect informativeness.

6 Discussion

6.1 Delusion equilibria

Robustness The delusion equilibrium appears to be very robust to ρ in this specification: it always exists, for all ρ. Assumption 5 is likely to be important for this result, since it implies an upper bound to the mental cost, which is one (and sometimes, the only) deterrent from delusion, making the incentive for denial harder to revert. As discussed below, α and s may also have an important role.

Manipulation of through ρ When multiple equilibria exist, we assumed that realism prevails. What happens if instead delusion prevails? In that case, the information designer would want to make sure that there is no delusion equilibrium. Because $\alpha = 1/2$, there is no possible manipulation through ρ. But for a generic α, a possibility of manipulation may arise. The incentive for delusion in a delusion equilibrium (19) is

$$\begin{aligned}\Psi\left(\lambda^{-i}=0\right) &= -m^i - c + \underbrace{\alpha\mathbb{E}_1^i\left[\theta \mid \hat{\pi}_h\right] + (1-\alpha)\,\mathbb{E}_1^i\left[\theta e^{-i} \mid \hat{\pi}_h\right] - (1-\alpha)\,\mathbb{E}_1^i\left[\theta e^{-i} \mid \hat{\pi}_l\right]}_{\text{Gain in anticipatory feelings}} + \underbrace{\alpha\mathbb{E}_0^i\left[\theta \mid \pi_l\right]}_{\text{Gain/Loss in standard}} \\ &= -m^i - c + \alpha\theta_H + (1-\alpha)\,\theta_H - (1-\alpha)\,\mathbb{E}_1^i\left[\theta \mid \hat{\pi}_l\right] + \alpha\mathbb{E}_0^i\left[\theta \mid \pi_l\right] \\ &= -m^i - c + \theta_H + (2\alpha-1)\left[\frac{1-\rho}{2-\rho}\theta_H + \frac{1}{2-\rho}\theta_L\right]\end{aligned}$$

The expression now depends on ρ, so the information designer may influence the existence of a delusion equilibrium through ρ:

$$\frac{\partial\Psi}{\partial\rho} = (1-2\alpha)\,\frac{1}{(2-\rho)^2}\left[\theta_H - \theta_L\right]\begin{cases} >0 & \text{if } \alpha < 1/2 \\ <0 & \text{if } \alpha > 1/2 \\ =0 & \text{if } \alpha = 1/2 \end{cases}$$

In a delusion equilibrium, the effect of ρ on the incentive for delusion depends on whether $\alpha > 1/2$. When $\alpha > 1/2$, own's effort has a higher contribution to the payoff; therefore own's effort mistakes become more important, so increasing ρ lowers the incentives for delusion by allowing more accurate decision-making if realism is chosen. In this case, the information designer would prefer $\rho_D^* = 1$ (assuming that realism equilibrium is not available). When $\alpha < 1/2$, others' contribution become more important, and so does the associated anticipatory utility; in this case, increasing ρ raises the temptation for delusion. The information designer would thus prefer $\rho_D^* = 1/2$ (if realism equilibrium not in play). Depending on α, then, the information designer could manipulate ρ and in some cases she might be able to break the delusion equilibrium. Similar results could be obtained by varying s, the strength of anticipatory feelings.

6.2 Naivete

The naivete assumption has two effects. On the one hand, it makes beliefs upon recalling good news $\hat{\pi}_h = 1$ (and hence, the associated anticipatory feelings) maximal; this strengthens the temptations for self-delusion and makes realism harder to sustain. On the other hand, it greatly simplifies the expression of $\hat{\pi}_h$ to a constant, but removes a potential dependency on ρ.

Perhaps counterintuitively, relaxing this assumption and allowing agents to be cognitively more sophisticated may be harmful for the information designer. Consider first the extreme opposite, fully Bayesian agents ($\chi^i = 1$). In that case, the posterior beliefs upon recalling good news

$$\hat{\pi}_h = \frac{\rho + (1-\rho)\left(1-\lambda^i\right)}{\rho + (2-\rho)\left(1-\lambda^i\right)} = \begin{cases} 1 & \text{if } \lambda^i = 1 \\ \frac{1}{2} & \text{if } \lambda^i = 0 \end{cases}$$

In pure strategies, the posterior belief $\hat{\pi}_h$ does not depend on ρ either, so the results are not expected to change much from full naivete. However, for partial cognitive sophistications $\chi^i \in (0,1)$, the posterior belief does change with ρ, since

$$\hat{\pi}_h = \frac{\rho + \chi^i(1-\rho)\left(1-\lambda^i\right)}{\rho + \chi^i(2-\rho)\left(1-\lambda^i\right)} = \begin{cases} 1 & \text{if } \lambda^i = 1 \\ \frac{\rho+\chi^i(1-\rho)}{\rho+\chi^i(2-\rho)} & \text{if } \lambda^i = 0 \end{cases}$$

and, when $\lambda^i = 0$,

$$\frac{\partial \hat{\pi}_h}{\partial \rho} = \frac{\left(1-\chi^i\right)\chi^i}{\left[\rho\left(1-\chi^i\right) + 2\chi^i\right]^2} \begin{cases} > 0 & \text{if } \chi^i \in (0,1) \\ = 0 & \text{if } \chi^i \in \{0,1\} \end{cases}$$

For partially sophisticated agents, the posterior belief upon delusion increases with the informativeness. This introduces a complication for the information designer, since making signals more informative now strengthens the incentives for delusion. Moreover, this effect is likely to be independent of other parameter values. Thus, the introduction of (partial) cognitive sophistication may drive ρ^* downwards, moving it even farther away from 1. As the optimal ρ^* is reduced, so does the maximized ex-ante payoff.

7 Conclusion

This work aims to determine the optimal signal informativeness in an organization where agents are capable of self-delusion and may fall into a groupthink equilibrium.

The standard result from decision theory is that more information is better. However, when agents can delude themselves, giving more precise information may also affect the temptations for delusion, introducing a trade-off for the principal. Under a set of simplifying assumptions, we found that there exist at least some combinations of parameters such that a realism equilibrium exists and the optimal choice by the central planner are perfectly informative signals. This confirms the standard result of the literature, and suggests that anticipatory feelings are not strong enough to overcome the decision-value of the information in this specification. However, it remains to study whether, under more general assumptions, there are cases in which anticipatory feelings can be strong enough to override the value of information.

Appendix: Proofs

7.1 Propositions and Lemmas

Derivation and characterization of $\ddot{\rho}$

Derivation of $\ddot{\rho}$:

- $\ddot{\pi}$ is the belief that satisfies the indifference condition:

$$\begin{aligned} \ddot{\pi}\theta_H + (1-\ddot{\pi})\,\theta_L &= c \\ \ddot{\pi}\Delta\theta + \theta_L &= c \\ \ddot{\pi} &= \frac{c-\theta_L}{\Delta\theta} \end{aligned}$$

$$\ddot{\pi}\left(\theta_H, \theta_L, c\right) = \frac{c-\theta_L}{\theta_H - \theta_L}$$

- This cutoff for $\hat{\pi}_l$ in turn determines a cutoff value for ρ. Values of $\hat{\pi}_l$ such that agent works are:

$$\begin{aligned} \hat{\pi}_l &\geq \ddot{\pi} \\ \hat{\pi}_l &\geq \frac{c-\theta_L}{\theta_H-\theta_L} \\ \frac{1-\rho}{2-\rho} &\geq \frac{c-\theta_L}{\theta_H-\theta_L} && \text{(From (8))} \\ (1-\rho)\,(\theta_H-\theta_L) &\geq (2-\rho)\,(c-\theta_L) && \text{(Assumption (4))} \\ (\theta_H-\theta_L) - 2\,(c-\theta_L) &\geq -\rho\,(c-\theta_L) + \rho\,(\theta_H-\theta_L) \\ \theta_H+\theta_L-2c &\geq \rho\,(\theta_H-c) \\ \frac{\theta_H+\theta_L-2c}{\theta_H-c} &\geq \rho && \text{(Assu (5))} \\ \frac{(\theta_H-c)-(c-\theta_L)}{\theta_H-c} &\geq \rho \end{aligned}$$

Characterization of $\ddot{\rho}$:

- $\ddot{\rho} < 1$
  - Always true by assumptions (5) and (4):

$$\begin{aligned} \ddot{\rho} &< 1 \\ \frac{\theta_H+\theta_L-2c}{\theta_H-c} &< 1 \\ \theta_H+\theta_L-2c &< \theta_H-c \\ \theta_L &< c \end{aligned}$$

- Condition for $\ddot{\rho} \geq 1/2$:

$$\begin{aligned}
\ddot{\rho} &\geq \frac{1}{2} \\
\frac{(\theta_H - c) - (c - \theta_L)}{\theta_H - c} &\geq \frac{1}{2} \\
(\theta_H - c) - (c - \theta_L) &\geq \frac{1}{2}(\theta_H - c) \qquad \text{(Assu. (5))} \\
\frac{1}{2}(\theta_H - c) - (c - \theta_L) &\geq 0 \\
\frac{1}{2}\theta_H - \frac{3}{2}c + \theta_L &\geq 0 \\
\theta_L &\geq -\left(\frac{1}{2}\theta_H - \frac{3}{2}c\right) \\
c &\leq \frac{1}{3}\theta_H + \frac{2}{3}\theta_L
\end{aligned}$$

Proof of Proposition 6

Ex-ante material utility (benchmark case):

- For $\rho \geq \ddot{\rho}$:

$$U^{Mat}_{Bench}(\rho) = \frac{1}{2}\left[\rho(\theta_H - c) + (1-\rho)\cdot 0\right] + \frac{1}{2}\cdot 0 = \frac{1}{2}\rho(\theta_H - c)$$

- For $\rho < \ddot{\rho}$:

$$\begin{aligned}
U^{Mat}_{Bench}(\rho) &= \frac{1}{2}\left[\rho(\theta_H - c) + (1-\rho)\cdot(\theta_H - c)\right] + \frac{1}{2}\cdot(\theta_L - c) \\
&= \frac{1}{2}(\theta_H + \theta_L) - c
\end{aligned}$$

- If $\rho \geq \ddot{\rho}$, then $U^{Mat}_{Bench}(\rho) \geq U^{Mat}_{Bench}(\rho')$ for any $\rho' < \ddot{\rho}$:

$$\begin{aligned}
\frac{1}{2}\rho(\theta_H - c) &> \frac{1}{2}(\theta_H + \theta_L) - c \\
\rho(\theta_H - c) &> \theta_H + \theta_L - 2c \\
\rho &> \frac{\theta_H - c + \theta_L - c}{\theta_H - c} \equiv \ddot{\rho}
\end{aligned}$$

  True by hypothesis $\rho \geq \ddot{\rho}$.

- If $\rho \geq \ddot{\rho}$, then $U^{Mat}_{Bench}(\rho)$ is strictly increasing in ρ:

$$\frac{\partial}{\partial \rho} U^{Mat}_{Bench}(\rho) = \frac{1}{2}(\theta_H - c) > 0$$

- Therefore, the maximum is at $\rho^*_{Bench} = 1$.

Proof of Lemma 8

- Expression of incentive for denial:

$$\begin{aligned}
\Psi\left(\lambda^{-i}=1\right) &= -m^i - c + \frac{1}{2}\mathbb{E}_1^i\left[\theta \mid \hat{\pi}_h\right] + \frac{1}{2}\mathbb{E}_1^i\left[\theta e^{-i} \mid \hat{\pi}_h\right] - \frac{1}{2}\mathbb{E}_1^i\left[\theta e^{-i} \mid \hat{\pi}_l\right] + \frac{1}{2}\mathbb{E}_0^i\left[\theta \mid \pi_l\right] \\
&= -m^i - c + \frac{1}{2}\theta_H + \frac{1}{2}\rho\theta_H - \frac{1}{2}\hat{\pi}_l\rho\theta_H + \frac{1}{2}\left[\pi_l\theta_H + (1-\pi_l)\,\theta_L\right] \\
&= -m^i - c + \frac{1}{2}\theta_H + \frac{1}{2}\rho\left(1-\hat{\pi}_l\right)\theta_H + \frac{1}{2}\left[\pi_l\theta_H + (1-\pi_l)\,\theta_L\right] \\
&= -m^i - c + \frac{1}{2}\rho\left(1-\hat{\pi}_l\right)\theta_H + \frac{1}{2}\left[(1+\pi_l)\,\theta_H + (1-\pi_l)\,\theta_L\right]
\end{aligned}$$

$$\begin{aligned}
&= -m^i - c + \frac{1}{2}\frac{\rho}{2-\rho}\theta_H + \frac{1}{2}\left[\left(\frac{2-\rho+1-\rho}{2-\rho}\right)\theta_H + \left(\frac{1}{2-\rho}\right)\theta_L\right] \\
&= -m^i - c + \frac{1}{2}\frac{\rho}{2-\rho}\theta_H + \frac{1}{2}\left[\left(\frac{3-2\rho}{2-\rho}\right)\theta_H + \left(\frac{1}{2-\rho}\right)\theta_L\right] \\
&= -m^i - c + \frac{1}{2}\left[\left(\frac{\rho}{2-\rho} + \frac{3-2\rho}{2-\rho}\right)\theta_H + \left(\frac{1}{2-\rho}\right)\theta_L\right] \\
&= -m^i - c + \frac{1}{2}\left[\left(\frac{3-\rho}{2-\rho}\right)\theta_H + \left(\frac{1}{2-\rho}\right)\theta_L\right] \\
&= -m^i - c + \frac{1}{2}\left[\left(\frac{2-\rho+1}{2-\rho}\right)\theta_H + \left(\frac{1}{2-\rho}\right)\theta_L\right] \\
&= -m^i - c + \frac{1}{2}\left[\left(1+\frac{1}{2-\rho}\right)\theta_H + \left(\frac{1}{2-\rho}\right)\theta_L\right] \\
&= -m^i - c + \frac{1}{2}\left(\frac{1}{2-\rho}\right)\left[\theta_H + \theta_L\right] + \frac{1}{2}\theta_H \\
&= \left(\frac{1}{2}\theta_H - m^i - c\right) + \frac{1}{2}\left(\frac{1}{2-\rho}\right)\left[\theta_H + \theta_L\right]
\end{aligned}$$

- Realism equilibrium ($\Psi \leq 0$) if and only if:

$$\begin{aligned}
\left(\frac{1}{2}\theta_H - m^i - c\right) + \frac{1}{2}\left(\frac{1}{2-\rho}\right)\left[\theta_H + \theta_L\right] &\leq 0 \\
\frac{1}{2}\theta_H - m^i - c &\leq -\frac{1}{2}\left(\frac{1}{2-\rho}\right)\left[\theta_H + \theta_L\right] \\
\left[\frac{1}{2}\theta_H - m^i - c\right](2-\rho) &\leq -\frac{1}{2}\left[\theta_H + \theta_L\right] \\
(2-\rho) &\leq \frac{-\frac{1}{2}\left[\theta_H + \theta_L\right]}{\left[\frac{1}{2}\theta_H - m^i - c\right]} \qquad \text{(Assu. (5))} \\
2 + \frac{\frac{1}{2}\left(\theta_H + \theta_L\right)}{\frac{1}{2}\theta_H - m^i - c} &\leq \rho
\end{aligned}$$

- Derivative of Ψ with respect to ρ:

$$\begin{aligned}
\frac{\partial\Psi}{\partial\rho} &= \frac{\partial}{\partial\rho}\left[\frac{1}{2}\left(\frac{1}{2-\rho}\right)\left[\theta_H + \theta_L\right]\right] \\
&= -\frac{1}{2}\left(\theta_H + \theta_L\right)\left(2-\rho\right)^{-2}(-1)
\end{aligned}$$

Positive if and only if:

$$\begin{aligned}\frac{1}{2}\left(\theta_H+\theta_L\right)\left(2-\rho\right)^{-2} &\geq 0\\ \left(\theta_H+\theta_L\right) &\geq 0\\ \theta_H+\theta_L &\geq 0\\ \theta_L &\geq -\theta_H\end{aligned}$$

Proof of Lemma 9

- Condition for $\bar{\rho}_R < 1$

$$\begin{aligned}2+\frac{\frac{1}{2}\left(\theta_H+\theta_L\right)}{\frac{1}{2}\theta_H-m^i-c} &< 1\\ \frac{\frac{1}{2}\left(\theta_H+\theta_L\right)}{\frac{1}{2}\theta_H-m^i-c} &< -1\\ \frac{1}{2}\left(\theta_H+\theta_L\right) &< -\left[\frac{1}{2}\theta_H-m^i-c\right] && \text{(Assu. (5))}\\ \theta_L &< -\theta_H-2\left[\frac{1}{2}\theta_H-m^i-c\right] < 0\end{aligned}$$

which is consistent with assumptions (4) and (5).

- Condition for $\bar{\rho}_R > 1/2$:

$$\begin{aligned}2+\frac{\frac{1}{2}\left(\theta_H+\theta_L\right)}{\frac{1}{2}\theta_H-m^i-c} &> \frac{1}{2}\\ \frac{\frac{1}{2}\left(\theta_H+\theta_L\right)}{\frac{1}{2}\theta_H-m^i-c} &> -\frac{3}{2}\\ \frac{1}{2}\left(\theta_H+\theta_L\right) &> -\frac{3}{2}\left[\frac{1}{2}\theta_H-m^i-c\right] && \text{(Assu. (5))}\\ \theta_L &> -\theta_H-3\left[\frac{1}{2}\theta_H-m^i-c\right]\end{aligned}$$

  - Therefore, for $1/2 < \bar{\rho}_R < 1$ it is necessary that θ_L lies in the interval:

$$-3\left[\frac{1}{2}\theta_H-\left(m^i+c\right)\right]-\theta_H < \theta_L < -2\left[\frac{1}{2}\theta_H-\left(m^i+c\right)\right]-\theta_H$$

    which again is consistent with A4 and A5.
  - This interval is non-empty:

$$\begin{aligned}&\left\{-2\left[\frac{1}{2}\theta_H-\left(m^i+c\right)\right]-\theta_H\right\}-\left\{-3\left[\frac{1}{2}\theta_H-\left(m^i+c\right)\right]-\theta_H\right\}\\ =&\frac{1}{2}\theta_H-\left(m^i+c\right) > 0 && \text{(Assu (5))}\end{aligned}$$

- Conditions for $\bar{\rho}_R \geq \ddot{\rho}$

– $\bar{\rho}_R \geq \ddot{\rho}$ if and only if

$$\begin{aligned}
\bar{\rho}_R &\geq \ddot{\rho} \\
2 + \frac{\frac{1}{2}(\theta_H + \theta_L)}{\left(\frac{1}{2}\theta_H - m^i - c\right)} &\geq 1 - \frac{(c - \theta_L)}{\theta_H - c} \\
\frac{\frac{1}{2}(\theta_H + \theta_L)}{\left(\frac{1}{2}\theta_H - m^i - c\right)} &\geq -1 - \frac{(c - \theta_L)}{\theta_H - c} \\
\frac{(\theta_H + \theta_L)}{2\left(\frac{1}{2}\theta_H - m^i - c\right)} &\geq \frac{-\theta_H + \theta_L}{\theta_H - c} \\
(\theta_H + \theta_L)(\theta_H - c) &\geq (-\theta_H + \theta_L) 2\left(\frac{1}{2}\theta_H - m^i - c\right) \qquad \text{(Assu (5))} \\
(\theta_H - c)\theta_H + (\theta_H - c)\theta_L &\geq -2\theta_H\left(\frac{1}{2}\theta_H - m^i - c\right) + 2\theta_L\left(\frac{1}{2}\theta_H - m^i - c\right) \\
\left[(\theta_H - c) - 2\left(\frac{1}{2}\theta_H - m^i - c\right)\right]\theta_L &\geq -\theta_H\left[2\left(\frac{1}{2}\theta_H - m^i - c\right) + (\theta_H - c)\right] \\
\left[2m^i + c\right]\theta_L &\geq -\theta_H\left[2\theta_H - 3c - 2m^i\right] \\
\left[2m^i + c\right]\theta_L &\geq -\theta_H\left[2(\theta_H - c) - \left(2m^i + c\right)\right] \\
\theta_L &\geq -\theta_H \frac{2(\theta_H - c) - \left(2m^i + c\right)}{2m^i + c}
\end{aligned}$$

– The ratio in the RHS:
  * $\left[2(\theta_H - c) - \left(2m^i + c\right)\right] / \left[2m^i + c\right] > 0$
    · The numerator is positive: $2(\theta_H - c) - \left(2m^i + c\right) = 2\left(\frac{1}{2}\theta_H - m^i - c\right) + (\theta_H - c) > 0$
    · The denominator is positive
  * $\left[2(\theta_H - c) - \left(2m^i + c\right)\right] / \left[2m^i + c\right] > 1$

$$\begin{aligned}
\frac{2(\theta_H - c) - \left(2m^i + c\right)}{2m^i + c} &> 1 \\
2(\theta_H - c) - \left(2m^i + c\right) &> 2m^i + c \\
2(\theta_H - c) &> 2\left(2m^i + c\right) \\
\theta_H - 2\left(m^i + c\right) &> 0 \\
2\left[\frac{1}{2}\theta_H - \left(m^i + c\right)\right] &> 0 \qquad \text{(Assu (5))}
\end{aligned}$$

– So $\bar{\rho}_R \geq \ddot{\rho} \iff \theta_L \geq -\theta_H\left[2(\theta_H - c) - \left(2m^i + c\right)\right] / \left[2m^i + c\right]$. Is this condition consistent with the conditions for $1/2 < \bar{\rho}_R < 1$?

* For $\bar{\rho}_R > 1/2$ we need $\theta_L > -\theta_H - 3\left[\frac{1}{2}\theta_H - \left(m^i + c\right)\right]$. Which lower bound is higher?

$$-3\left[\frac{1}{2}\theta_H - \left(m^i + c\right)\right] - \theta_H >^{?} -\theta_H \frac{2\left(\theta_H - c\right) - \left(2m^i + c\right)}{2m^i + c}$$

$$-3\left[\frac{1}{2}\theta_H - \left(m^i + c\right)\right] - \theta_H > -\theta_H \frac{2\left(\frac{1}{2}\theta_H - m^i - c\right) + \left(\theta_H - c\right)}{2m^i + c}$$

$$-3\left[\frac{1}{2}\theta_H - \left(m^i + c\right)\right]\left(2m^i + c\right) - \theta_H\left(2m^i + c\right) > -2\theta_H\left[\frac{1}{2}\theta_H - m^i - c\right] - \theta_H\left(\theta_H - c\right.$$

$$\left[\frac{1}{2}\theta_H - \left(m^i + c\right)\right]\left[2\theta_H - 3\left(2m^i + c\right)\right] + \theta_H\left[\left(\theta_H - c\right) - \left(2m^i + c\right)\right] > 0$$

$$\left[\frac{1}{2}\theta_H - \left(m^i + c\right)\right]\left[2\theta_H - 3\left(2m^i + c\right)\right] + \theta_H\left[\theta_H - 2\left(m^i + c\right)\right] > 0$$

$$\left[\frac{1}{2}\theta_H - \left(m^i + c\right)\right]\left[2\theta_H - 3\left(2m^i + c\right)\right] + \theta_H 2\left[\frac{1}{2}\theta_H - \left(m^i + c\right)\right] > 0$$

$$\left[4\theta_H - 3\left(2m^i + c\right)\right]\left[\frac{1}{2}\theta_H - \left(m^i + c\right)\right] > 0$$

$$\left[4\theta_H - 6m^i - 3c\right]\left[\frac{1}{2}\theta_H - \left(m^i + c\right)\right] > 0$$

$$\underbrace{\left[6\left(\frac{1}{2}\theta_H - m^i - c\right) + \theta_H + 3c\right]}_{>0}\underbrace{\left[\frac{1}{2}\theta_H - \left(m^i + c\right)\right]}_{>0} > 0$$

True by assumption 5. Therefore, $\underline{\text{whenever } \bar{\rho}_R > 1/2 \Longrightarrow \bar{\rho}_R > \ddot{\rho}}$.

* For $\bar{\rho}_R < 1$, we need $\theta_L < -\theta_H - 2\left[\frac{1}{2}\theta_H - \left(m^i + c\right)\right]$. Is this consistent with $\bar{\rho}_R > \ddot{\rho}$? We need the upper bound to be above the lower bound.

$$-2\left[\frac{1}{2}\theta_H - \left(m^i + c\right)\right] - \theta_H \geq^{?} -\theta_H \frac{2\left(\theta_H - c\right) - \left(2m^i + c\right)}{2m^i + c}$$

$$-2\left[\frac{1}{2}\theta_H - \left(m^i + c\right)\right] - \theta_H \geq^{?} -\theta_H \frac{2\left(\frac{1}{2}\theta_H - m^i - c\right) + \left(\theta_H - c\right)}{2m^i + c}$$

$$-2\left[\frac{1}{2}\theta_H - \left(m^i + c\right)\right]\left(2m^i + c\right) - \theta_H\left(2m^i + c\right) \geq^{?} -2\theta_H\left[\frac{1}{2}\theta_H - \left(m^i + c\right)\right] - \theta_H\left(\theta_H\right.$$

$$\left[2\theta_H - 2\left(2m^i + c\right)\right]\left[\frac{1}{2}\theta_H - \left(m^i + c\right)\right] + \theta_H\left[\left(\theta_H - c\right) - \left(2m^i + c\right)\right] \geq^{?} 0$$

$$\left[2\theta_H - 2\left(2m^i + c\right)\right]\left[\frac{1}{2}\theta_H - \left(m^i + c\right)\right] + \theta_H\left[\theta_H - \left(2m^i + 2c\right)\right] \geq^{?} 0$$

$$\left[2\theta_H - 2\left(2m^i + c\right)\right]\left[\frac{1}{2}\theta_H - \left(m^i + c\right)\right] + 2\theta_H\left[\frac{1}{2}\theta_H - \left(m^i + c\right)\right] \geq^{?} 0$$

$$\left[4\theta_H - 2\left(2m^i + c\right)\right]\left[\frac{1}{2}\theta_H - \left(m^i + c\right)\right] \geq^{?} 0$$

$$\underbrace{\left[4\left(\frac{1}{2}\theta_H - \left(m^i - c\right)\right) + 2\theta_H + 2c\right]}_{>0}\underbrace{\left[\frac{1}{2}\theta_H - \left(m^i + c\right)\right]}_{>0} > 0$$

Always true. So $\bar{\rho}_R < 1$ is always consistent $\bar{\rho}_R > \ddot{\rho}$.

Proof of Lemma 10

Delusion equilibrium condition:

- Denial is preferrable if $\Psi \geq 0$:

$$\frac{1}{2-\rho}(\theta_H - \theta_L) - m^i \geq 0$$
$$\frac{1}{2-\rho}(\theta_H - \theta_L) \geq m^i$$

- If $m^i = 0$, then $\Psi \geq 0$ always, for all ρ, since $\theta_H - \theta_L > 0$.
- If $m^i > 0$, then

$$\frac{1}{2-\rho}(\theta_H - \theta_L) - m^i \geq 0$$
$$\frac{1}{2-\rho}(\theta_H - \theta_L) \geq m^i$$
$$\frac{(\theta_H - \theta_L)}{m^i} \geq 2 - \rho$$
$$\rho \geq 2 - \frac{(\theta_H - \theta_L)}{m^i}$$

Characterization of $\bar{\rho}_B$:

- $\bar{\rho}_B < \ddot{\rho}$ if and only if:

$$2 - \frac{\theta_H - \theta_L}{m^i} < \frac{(\theta_H - c) - (c - \theta_L)}{\theta_H - c}$$
$$\frac{2m^i - \theta_H + \theta_L}{m^i} < \frac{(\theta_H + \theta_L - 2c)}{\theta_H - c}$$
$$(\theta_H - c)\left(2m^i - \theta_H + \theta_L\right) < (\theta_H + \theta_L - 2c)\, m^i \qquad \text{(Assu (5))}$$
$$\theta_H\left(2m^i - \theta_H + \theta_L\right) - c\left(2m^i - \theta_H + \theta_L\right) < (\theta_H + \theta_L - 2c)\, m^i$$
$$2\theta_H m^i - \theta_H^2 + \theta_H\theta_L - 2m^i c + \theta_H c - \theta_L c < \theta_H m^i + \theta_L m^i - 2cm^i$$
$$-\theta_H^2 + (\theta_H - \theta_L)\, m^i + \theta_H\theta_L + (\theta_H - \theta_L)\, c < 0$$
$$\theta_H\left(-\theta_H + \theta_L\right) + (\theta_H - \theta_L)\, m^i + (\theta_H - \theta_L)\, c < 0$$
$$\theta_H\left(-\theta_H + \theta_L\right) + (\theta_H - \theta_L)\left(m^i + c\right) < 0$$
$$0 < \theta_H\left(\theta_H - \theta_L\right) - (\theta_H - \theta_L)\left(m^i + c\right)$$
$$0 < (\theta_H - \theta_L)\left(\theta_H - m^i + c\right) \qquad \text{(Assu (5) and (4))}$$

  True by assumptions.

- $\bar{\rho}_B \geq 1/2$ if and only if:

$$\bar{\rho}_B = 2 - \frac{\theta_H - \theta_L}{m^i} \geq 1/2$$
$$\frac{3}{2} \geq \frac{\theta_H - \theta_L}{m^i}$$
$$m^i \geq \frac{2}{3}(\theta_H - \theta_L)$$

Equivalently,

$$\begin{aligned} m^i &\geq \frac{2}{3}(\theta_H - \theta_L) \\ \theta_L &\geq \theta_H - \frac{3}{2}m^i \\ &= 2\left[\frac{1}{2}\theta_H - m^i\right] + \frac{1}{2}m^i > 0 \qquad \text{(Assu. 5)} \end{aligned}$$

but this never true since $\theta_L < 0$.

- Check whether $\bar{\rho}_B \geq 1/2$ is consistent with A5
  - Condition for $\bar{\rho}_B \geq 1/2$:
$$m^i \geq 2/3 \cdot (\theta_H - \theta_L)$$
  - A5 requires:
$$m^i \leq \frac{1}{2}\theta_H - c$$
  - Both conditions together imply:
$$\begin{aligned} \frac{2}{3}(\theta_H - \theta_L) &\leq m^i \leq \frac{1}{2}\theta_H - c \\ \frac{2}{3}(\theta_H - \theta_L) &\leq \frac{1}{2}\theta_H - c \\ 2(\theta_H - \theta_L) &\leq \frac{3}{2}\theta_H - 3c \\ \frac{1}{2}\theta_H - 2\theta_L &\leq -3c \\ \theta_H + 6c &\leq 4\theta_L \\ \frac{1}{4}\underbrace{(\theta_H + 6c)}_{>0} &\leq \theta_L \end{aligned}$$
    But this contradicts assumption 4.
  - Therefore, A5 implies that $m^i < 2/3 \cdot (\theta_H - \theta_L)$, so $\bar{\rho}_B < 1/2$.

Proof of Proposition 12

- In the range $\rho > \ddot{\rho}$:
  - Agents do not work under beliefs $\hat{\pi}_l$, only work under $\hat{\pi}_h$.
  - Ex-ante material utility of collective realism. With probability $1/2$ the state is H. In that case, a fraction ρ of agents gets the good signal and works, obtaining $\rho(\theta_H - c)$ from them; and a fraction $1-\rho$ gets the bad signal and do not work (realism with $\rho > \ddot{\rho}$), obtaining 0 from them. With probability $1/2$ the state is L, all agents get a bad signal, nobody works, and the material payoff is 0.
$$\begin{aligned} U_{-1}^{Material}(\rho) &= \frac{1}{2}\left[\rho(\theta_H - c) + (1-\rho)(0)\right] + \frac{1}{2}\left[(0)\right] \\ &= \frac{1}{2}\rho(\theta_H - c) \end{aligned}$$
    It is strictly increasing in ρ.

- Ex-ante material utility of collective delusion. Regardless the signal, everybody is carrying high date-1 beliefs (either by having observed a good signal or by distorting up a bad signal) and therefore working. The informativeness of the signal plays no role for the expected material payoffs (no matter how informative are the signals, everybody is totally ignoring them). The project materializes fully by all N agents, and then the expected material payoff is:

$$U_{-1,D}^{Material} = \frac{1}{2}\theta_H + \frac{1}{2}\theta_L - c$$

  It is independent of ρ (because bad signals are always ignored by everyone).

- Conditions for higher payoff under collective realism. It is always the case that collective realism offers a higher ex-ante material payoff:

$$\begin{aligned} \frac{1}{2}\rho\left(\theta_H - c\right) &\geq \frac{1}{2}\theta_H + \frac{1}{2}\theta_L - c \\ \rho\left(\theta_H - c\right) &\geq \theta_H + \theta_L - 2c \\ \rho &\geq \frac{\theta_H + \theta_L - 2c}{\theta_H - c} \qquad \text{(Since 5)} \\ \rho &\geq \frac{\left(\theta_H - c\right) - \left(c - \theta_L\right)}{\theta_H - c} \equiv \ddot{\rho} \end{aligned}$$

  The latter inequality is true by hypothesis $\rho \geq \ddot{\rho}$.

- Therefore a realism delivers higher payoffs, for all $\rho \geq \ddot{\rho}$.

- In the range $\rho < \ddot{\rho}$:

  - Agents work under both beliefs, $\hat{\pi}_l$ and $\hat{\pi}_h$.

  - Ex-ante material utility of collective realism. All agents work regardless their date-0 signals σ^i and regardless of their strategy λ^i, so the project is fully materialized by N agents regardless of the distribution of signals induced by ρ. The ex-ante material payoff is then given by the full expected value of the project minus the effort cost:

$$\begin{aligned} U_{-1}^{Material}\left(\rho\right) &= ET\left(1/2\right) - c \\ &= \frac{1}{2}\left(\theta_H + \theta_L\right) - c \end{aligned}$$

    It is independent of ρ (because agents work independently of which signal is encoded into memory).

  - Ex-ante material utility of collective denial. The same logic applies: everybody is working no matter the signals and beliefs. Again, ρ has no influence on the material payoffs of collective delusion.

$$\begin{aligned} U_{-1}^{Material}\left(\rho\right) &= ET\left(1/2\right) - c \\ &= \frac{1}{2}\left(\theta_H + \theta_L\right) - c \end{aligned}$$

  - Then collective realism provides the same ex-ante material payoffs as collective denial, for all $\rho < \ddot{\rho}$.

7.2 Derivation of other expressions

(17): Utility of realism in delusion equilibrium ($\rho \geq \ddot{\rho}$)

Date-0 utility of Realism:

$$\begin{aligned}
U_0^i &= -m^i + \mathbb{E}_0^i\left[-ce^i\right] + \left[\frac{1}{2}\mathbb{E}_1^i\left[\theta e^i \mid \hat{\pi}^i\right] + \frac{1}{2}\mathbb{E}_1^i\left[\theta e^{-i} \mid \hat{\pi}^i\right]\right] + \left[\frac{1}{2}\mathbb{E}_0^i\left[\theta e^i \mid \pi^i\right] + \frac{1}{2}\mathbb{E}_0^i\left[\theta e^{-i} \mid \pi^i\right]\right] \\
&= 0 + \mathbb{E}_0^i\left[-c \cdot 0\right] + \left[\frac{1}{2}\mathbb{E}_1^i\left[\theta \cdot 0 \mid \hat{\pi}^i\right] + \frac{1}{2}\mathbb{E}_1^i\left[\theta e^{-i} \mid \hat{\pi}^i\right]\right] + \left[\frac{1}{2}\mathbb{E}_0^i\left[\theta \cdot 0 \mid \pi^i\right] + \frac{1}{2}\mathbb{E}_0^i\left[\theta e^{-i} \mid \pi^i\right]\right] \\
&= \frac{1}{2}\mathbb{E}_1^i\left[\theta e^{-i} \mid \hat{\pi}^i\right] + \frac{1}{2}\mathbb{E}_0^i\left[\theta e^{-i} \mid \pi^i\right]
\end{aligned}$$

The date-t expectation $\mathbb{E}_t^i\left[\theta e^{-i} \mid \pi_t\right]$:

$$\begin{aligned}
\mathbb{E}_t^i\left[\theta e^{-i} \mid \pi_t\right] &= \pi_t\left[\rho \cdot \theta_H + (1-\rho) \cdot \left(1-\lambda^{-i}\right) \cdot \theta_H\right] + (1-\pi_t)\left[\left(1-\lambda^{-i}\right) \cdot \theta_L\right] \\
&= \pi_t\left[\rho\theta_H + (1-\rho)\left(1-\lambda^{-i}\right)\theta_H\right] + (1-\pi_t)\left[\left(1-\lambda^{-i}\right)\theta_L\right] \\
&= \pi_t\left[\theta_H\right] + (1-\pi_t)\left[\theta_L\right] && \text{(Denial eqm: } \lambda^{-i} = 0\text{)} \\
&= \pi_t\theta_H + (1-\pi_t)\,\theta_L \\
&= \mathbb{E}_t\left[\theta \mid \pi_t\right]
\end{aligned}$$

Hence the utility:

$$\begin{aligned}
&= \frac{1}{2}\mathbb{E}_1^i\left[\theta e^{-i} \mid \hat{\pi}^i\right] + \frac{1}{2}\mathbb{E}_0^i\left[\theta e^{-i} \mid \pi^i\right] \\
&= \frac{1}{2}\mathbb{E}_1^i\left[\theta \mid \hat{\pi}_l\right] + \frac{1}{2}\mathbb{E}_0^i\left[\theta \mid \pi_l\right] \\
&= \mathbb{E}_1^i\left[\theta \mid \pi_l\right] && \text{(Since } \pi_l = \hat{\pi}_l\text{)} \\
&= \pi_l\theta_H + (1-\pi_l)\,\theta_L \\
&= \left(\frac{1-\rho}{2-\rho}\right)\theta_H + \left(\frac{1}{2-\rho}\right)\theta_L
\end{aligned}$$

(18): Utility of delusion in delusion equilibrium ($\rho \geq \ddot{\rho}$)

$$\begin{aligned}
U_{0,D}^i &= -m^i + \mathbb{E}_0^i\left[-ce^i\right] + \frac{1}{2}\mathbb{E}_1^i\left[\theta e^i \mid \hat{\pi}^i\right] + \frac{1}{2}\mathbb{E}_1^i\left[\theta e^{-i} \mid \hat{\pi}^i\right] + \frac{1}{2}\mathbb{E}_0^i\left[\theta e^i \mid \pi^i\right] + \frac{1}{2}\mathbb{E}_0^i\left[\theta e^{-i} \mid \pi^i\right] \\
&= -m^i - c + \frac{1}{2}\mathbb{E}_1^i\left[\theta \mid \hat{\pi}_h\right] + \frac{1}{2}\mathbb{E}_1^i\left[\theta e^{-i} \mid \hat{\pi}_h\right] + \frac{1}{2}\mathbb{E}_0^i\left[\theta \mid \pi_l\right] + \frac{1}{2}\mathbb{E}_0^i\left[\theta e^{-i} \mid \pi_l\right] \\
&= -m^i - c + \frac{1}{2}\theta_H + \frac{1}{2}\mathbb{E}_1^i\left[\theta e^{-i} \mid \hat{\pi}_h\right] + \frac{1}{2}\left(\pi_l\theta_H + (1-\pi_l)\,\theta_L\right) + \frac{1}{2}\mathbb{E}_0^i\left[\theta e^{-i} \mid \pi_l\right] \\
&= -m^i - c + \frac{1}{2}\theta_H + \frac{1}{2}\mathbb{E}_1^i\left[\theta e^{-i} \mid \hat{\pi}_h\right] + \frac{1}{2}\left(\pi_l\theta_H + (1-\pi_l)\,\theta_L\right) + \frac{1}{2}\mathbb{E}_0^i\left[\theta e^{-i} \mid \pi_l\right]
\end{aligned}$$

The expectation $\mathbb{E}_t^i\left[\theta e^{-i} \mid \pi\right]$

$$\begin{aligned}
\mathbb{E}_t^i\left[\theta e^{-i} \mid \pi_t\right] &= \pi_t\left[\rho \cdot \theta_H + (1-\rho) \cdot \left(1-\lambda^{-i}\right) \cdot \theta_H\right] + (1-\pi_t)\left[\left(1-\lambda^{-i}\right) \cdot \theta_L\right] \\
&= \pi_t\left[\rho\theta_H + (1-\rho)\left(1-\lambda^{-i}\right)\theta_H\right] + (1-\pi_t)\left[\left(1-\lambda^{-i}\right)\theta_L\right] \\
&= \pi_t\left[\theta_H\right] + (1-\pi_t)\left[\theta_L\right] && \text{(Denial eqm: } \lambda^{-i} = 0\text{)} \\
&= \pi_t\theta_H + (1-\pi_t)\,\theta_L \\
&= \mathbb{E}_t\left[\theta \mid \pi_t\right]
\end{aligned}$$

Therefore the date-0 utility:

$$\begin{aligned}
U^i_{0,D} &= -m^i - c + \frac{1}{2}\theta_H + \frac{1}{2}\mathbb{E}^i_1\left[\theta \mid \hat{\pi}_h\right] + \frac{1}{2}\mathbb{E}^i_0\left[\theta \mid \pi_l\right] + \frac{1}{2}\mathbb{E}^i_0\left[\theta \mid \pi_l\right] \\
&= -m^i - c + \frac{1}{2}\theta_H + \frac{1}{2}\theta_H + \frac{1}{2}\mathbb{E}^i_0\left[\theta \mid \pi_l\right] + \frac{1}{2}\mathbb{E}^i_0\left[\theta \mid \pi_l\right] \\
&= -m^i - c + \theta_H + \mathbb{E}^i_0\left[\theta \mid \pi_l\right] \\
&= -m^i - c + \theta_H + \left[\frac{1-\rho}{2-\rho}\theta_H + \frac{1}{2-\rho}\theta_L\right] \\
&= -m^i - c + \frac{1}{2}\theta_H + \frac{1}{2}\rho\theta_H + \frac{1}{2}\left(\pi_l\theta_H + (1-\pi_l)\,\theta_L\right) + \frac{1}{2}\pi_l\rho\theta_H \\
&= -m^i - c + \frac{1}{2}\left(1+\rho\right)\theta_H + \frac{1}{2}\left(\pi_l\left(1+\rho\right)\theta_H + (1-\pi_l)\,\theta_L\right) \\
&= -m^i - c + \frac{1}{2}\left(1+\rho\right)\theta_H + \frac{1}{2}\left(\frac{1-\rho}{2-\rho}\left(1+\rho\right)\theta_H + \frac{1}{2-\rho}\theta_L\right)
\end{aligned}$$

(19): Net incentive for delusion in delusion equilibrium ($\rho \geq \ddot{\rho}$)

$$\begin{aligned}
\Psi\left(\lambda^{-i}=0\right) &= -m^i - c + \frac{1}{2}\mathbb{E}^i_1\left[\theta \mid \hat{\pi}_h\right] + \frac{1}{2}\mathbb{E}^i_1\left[\theta e^{-i} \mid \hat{\pi}_h\right] - \frac{1}{2}\mathbb{E}^i_1\left[\theta e^{-i} \mid \hat{\pi}_l\right] + \frac{1}{2}\mathbb{E}^i_0\left[\theta \mid \pi_l\right] \\
&= -m^i - c + \frac{1}{2}\theta_H + \frac{1}{2}\mathbb{E}^i_1\left[\theta \mid \hat{\pi}_h\right] - \frac{1}{2}\mathbb{E}^i_1\left[\theta \mid \hat{\pi}_l\right] + \frac{1}{2}\mathbb{E}^i_0\left[\theta \mid \pi_l\right] \\
&= -m^i - c + \frac{1}{2}\theta_H + \frac{1}{2}\theta_H - \frac{1}{2}\mathbb{E}^i_1\left[\theta \mid \hat{\pi}_l\right] + \frac{1}{2}\mathbb{E}^i_0\left[\theta \mid \pi_l\right] \\
&= -m^i - c + \theta_H \qquad \text{(Since } \hat{\pi}_l = \pi_l\text{)}
\end{aligned}$$

(20): Utility of realism in realism equilibrium ($\rho \geq \ddot{\rho}$)

Date-0 utility of Realism:

$$\begin{aligned}
U^i_0 &= -m^i + \mathbb{E}^i_0\left[-ce^i\right] + \left[\frac{1}{2}\mathbb{E}^i_1\left[\theta e^i \mid \hat{\pi}^i\right] + \frac{1}{2}\mathbb{E}^i_1\left[\theta e^{-i} \mid \hat{\pi}^i\right]\right] + \left[\frac{1}{2}\mathbb{E}^i_0\left[\theta e^i \mid \pi^i\right] + \frac{1}{2}\mathbb{E}^i_0\left[\theta e^{-i} \mid \pi^i\right]\right] \\
&= 0 + \mathbb{E}^i_0\left[-c\cdot 0\right] + \left[\frac{1}{2}\mathbb{E}^i_1\left[\theta\cdot 0 \mid \hat{\pi}^i\right] + \frac{1}{2}\mathbb{E}^i_1\left[\theta e^{-i} \mid \hat{\pi}^i\right]\right] + \left[\frac{1}{2}\mathbb{E}^i_0\left[\theta\cdot 0 \mid \pi^i\right] + \frac{1}{2}\mathbb{E}^i_0\left[\theta e^{-i} \mid \pi^i\right]\right] \\
&= \frac{1}{2}\mathbb{E}^i_1\left[\theta e^{-i} \mid \hat{\pi}^i\right] + \frac{1}{2}\mathbb{E}^i_0\left[\theta e^{-i} \mid \pi^i\right]
\end{aligned}$$

The date-1 expectation $\mathbb{E}^i_1\left[\theta e^{-i} \mid \hat{\pi}^i\right]$:

$$\begin{aligned}
\mathbb{E}^i_1\left[\theta e^{-i} \mid \hat{\pi}^i\right] &= \hat{\pi}^i\left[\rho\cdot\theta_H\cdot 1 + (1-\rho)\cdot\left(1-\lambda^{-i}\right)\cdot\theta_H\cdot 1\right] + \left(1-\hat{\pi}^i\right)\left[\left(1-\lambda^{-i}\right)\cdot\theta_L\cdot 1\right] \\
&= \hat{\pi}^i\left[\rho\theta_H + (1-\rho)\left(1-\lambda^{-i}\right)\theta_H\right] + \left(1-\hat{\pi}^i\right)\left[\left(1-\lambda^{-i}\right)\theta_L\right] \\
&= \hat{\pi}^i\left[\rho\theta_H + 0\right] + \left(1-\hat{\pi}^i\right)[0] \qquad \text{(Realism eqm: } \lambda^{-i}=1\text{)} \\
&= \hat{\pi}^i\rho\theta_H
\end{aligned}$$

In a realism equilibrium, $\hat{\sigma}^i = l$, so $\hat{\pi}^i = \hat{\pi}_l$:

$$\mathbb{E}^i_1\left[\theta e^{-i} \mid \hat{\pi}_l\right] = \hat{\pi}_l\rho\theta_H = \left(\frac{1-\rho}{2-\rho}\right)\rho\theta_H$$

(23): Net incentive in a realism equilibrium

$$\begin{aligned}
\Psi\left(\lambda^{-i}=1\right) &= -m^i - c + \frac{1}{2}\mathbb{E}_1^i\left[\theta \mid \hat{\pi}_h\right] + \frac{1}{2}\mathbb{E}_1^i\left[\theta e^{-i} \mid \hat{\pi}_h\right] - \frac{1}{2}\mathbb{E}_1^i\left[\theta e^{-i} \mid \hat{\pi}_l\right] + \frac{1}{2}\mathbb{E}_0^i\left[\theta \mid \pi_l\right] \\
&= -m^i - c + \frac{1}{2}\theta_H + \frac{1}{2}\rho\theta_H - \frac{1}{2}\hat{\pi}_l\rho\theta_H + \frac{1}{2}\left[\pi_l\theta_H + (1-\pi_l)\,\theta_L\right] \\
&= -m^i - c + \frac{1}{2}\theta_H + \frac{1}{2}\rho\left(1-\hat{\pi}_l\right)\theta_H + \frac{1}{2}\left[\pi_l\theta_H + (1-\pi_l)\,\theta_L\right] \\
&= -m^i - c + \frac{1}{2}\rho\left(1-\hat{\pi}_l\right)\theta_H + \frac{1}{2}\left[(1+\pi_l)\,\theta_H + (1-\pi_l)\,\theta_L\right]
\end{aligned}$$

$$\begin{aligned}
&= -m^i - c + \frac{1}{2}\frac{\rho}{2-\rho}\theta_H + \frac{1}{2}\left[\left(\frac{2-\rho+1-\rho}{2-\rho}\right)\theta_H + \left(\frac{1}{2-\rho}\right)\theta_L\right] \\
&= -m^i - c + \frac{1}{2}\frac{\rho}{2-\rho}\theta_H + \frac{1}{2}\left[\left(\frac{3-2\rho}{2-\rho}\right)\theta_H + \left(\frac{1}{2-\rho}\right)\theta_L\right] \\
&= -m^i - c + \frac{1}{2}\left[\left(\frac{\rho}{2-\rho} + \frac{3-2\rho}{2-\rho}\right)\theta_H + \left(\frac{1}{2-\rho}\right)\theta_L\right] \\
&= -m^i - c + \frac{1}{2}\left[\left(\frac{3-\rho}{2-\rho}\right)\theta_H + \left(\frac{1}{2-\rho}\right)\theta_L\right] \\
&= -m^i - c + \frac{1}{2}\left[\left(\frac{2-\rho+1}{2-\rho}\right)\theta_H + \left(\frac{1}{2-\rho}\right)\theta_L\right] \\
&= -m^i - c + \frac{1}{2}\left[\left(1+\frac{1}{2-\rho}\right)\theta_H + \left(\frac{1}{2-\rho}\right)\theta_L\right] \\
&= -m^i - c + \frac{1}{2}\left(\frac{1}{2-\rho}\right)\left[\theta_H + \theta_L\right] + \frac{1}{2}\theta_H \\
&= \left(\frac{1}{2}\theta_H - m^i - c\right) + \frac{1}{2}\left(\frac{1}{2-\rho}\right)\left[\theta_H + \theta_L\right]
\end{aligned}$$

(26): Utility of realism in delusion equilibrium, $\rho < \ddot{\rho}$

Date-0 utility of Realism:

$$\begin{aligned}
U_0^i &= -m^i + \mathbb{E}_0^i\left[-ce^i\right] + \left[\frac{1}{2}\mathbb{E}_1^i\left[\theta e^i \mid \hat{\pi}^i\right] + \frac{1}{2}\mathbb{E}_1^i\left[\theta e^{-i} \mid \hat{\pi}^i\right]\right] + \left[\frac{1}{2}\mathbb{E}_0^i\left[\theta e^i \mid \pi^i\right] + \frac{1}{2}\mathbb{E}_0^i\left[\theta e^{-i} \mid \pi^i\right]\right] \\
&= 0 + -c + \left[\frac{1}{2}\mathbb{E}_1^i\left[\theta \mid \hat{\pi}_l\right] + \frac{1}{2}\mathbb{E}_1^i\left[\theta e^{-i} \mid \hat{\pi}_l\right]\right] + \left[\frac{1}{2}\mathbb{E}_0^i\left[\theta \mid \pi_l\right] + \frac{1}{2}\mathbb{E}_0^i\left[\theta e^{-i} \mid \pi_l\right]\right] \\
&= -c + \frac{1}{2}\mathbb{E}_1^i\left[\theta \mid \hat{\pi}_l\right] + \frac{1}{2}\mathbb{E}_1^i\left[\theta e^{-i} \mid \hat{\pi}_l\right] + \frac{1}{2}\mathbb{E}_0^i\left[\theta \mid \pi_l\right] + \frac{1}{2}\mathbb{E}_0^i\left[\theta e^{-i} \mid \pi_l\right]
\end{aligned}$$

The date-t expectation $\mathbb{E}_t^i\left[\theta e^{-i} \mid \pi_t\right]$:

$$\begin{aligned}
\mathbb{E}_t^i\left[\theta e^{-i} \mid \pi_t\right] &= \pi_t\left[\rho\cdot\theta_H + (1-\rho)\cdot\left(1-\lambda^{-i}\right)\cdot\theta_H\right] + (1-\pi_t)\left[\left(1-\lambda^{-i}\right)\cdot\theta_L\right] \\
&= \pi_t\left[\rho\theta_H + (1-\rho)\left(1-\lambda^{-i}\right)\theta_H\right] + (1-\pi_t)\left[\left(1-\lambda^{-i}\right)\theta_L\right] \\
&= \pi_t\left[\theta_H\right] + (1-\pi_t)\left[\theta_L\right] \qquad \text{(Denial eqm: } \lambda^{-i}=0\text{)} \\
&= \pi_t\theta_H + (1-\pi_t)\,\theta_L \\
&= \mathbb{E}_t\left[\theta \mid \pi_t\right]
\end{aligned}$$

Hence the utility:

$$
\begin{aligned}
&= -c + \frac{1}{2}\mathbb{E}_1^i\left[\theta \mid \hat{\pi}_l\right] + \frac{1}{2}\mathbb{E}_1^i\left[\theta e^{-i} \mid \hat{\pi}_l\right] + \frac{1}{2}\mathbb{E}_0^i\left[\theta \mid \pi_l\right] + \frac{1}{2}\mathbb{E}_0^i\left[\theta e^{-i} \mid \pi_l\right] \\
&= -c + \frac{1}{2}\mathbb{E}_1^i\left[\theta \mid \hat{\pi}_l\right] + \frac{1}{2}\mathbb{E}_1^i\left[\theta \mid \hat{\pi}_l\right] + \frac{1}{2}\mathbb{E}_0^i\left[\theta \mid \pi_l\right] + \frac{1}{2}\mathbb{E}_0^i\left[\theta \mid \pi_l\right] \\
&= -c + \mathbb{E}_1^i\left[\theta \mid \hat{\pi}_l\right] + \mathbb{E}_0^i\left[\theta \mid \pi_l\right] \\
&= -c + 2\mathbb{E}_1^i\left[\theta \mid \pi_l\right] && \text{(Since } \pi_l = \hat{\pi}_l\text{)} \\
&= -c + 2\left[\pi_l\theta_H + (1-\pi_l)\,\theta_L\right] \\
&= -c + 2\left[\left(\frac{1-\rho}{2-\rho}\right)\theta_H + \left(\frac{1}{2-\rho}\right)\theta_L\right]
\end{aligned}
$$

(28) : Net incentive for delusion, $\rho < \ddot{\rho}$.

$$
\begin{aligned}
\Psi\left(\lambda^{-i} = 0\right) &= -m^i + \frac{1}{2}\left(\mathbb{E}_1^i\left[\theta \mid \hat{\pi}_h\right] - \mathbb{E}_1^i\left[\theta \mid \hat{\pi}_l\right]\right) + \frac{1}{2}\left(\mathbb{E}_1^i\left[\theta e^{-i} \mid \hat{\pi}_h\right] - \mathbb{E}_1^i\left[\theta e^{-i} \mid \hat{\pi}_l\right]\right) \\
&= -m^i + \frac{1}{2}\left(\mathbb{E}_1^i\left[\theta \mid \hat{\pi}_h\right] - \mathbb{E}_1^i\left[\theta \mid \hat{\pi}_l\right]\right) + \frac{1}{2}\left(\mathbb{E}_1^i\left[\theta \mid \hat{\pi}_h\right] - \mathbb{E}_1^i\left[\theta \mid \hat{\pi}_l\right]\right) \\
&\qquad\qquad (e^{-i} = 1 \text{ in delusion eqm}) \\
&= -m^i + \left(\mathbb{E}_1^i\left[\theta \mid \hat{\pi}_h\right] - \mathbb{E}_1^i\left[\theta \mid \hat{\pi}_l\right]\right) \\
&= -m^i + \theta_H - \left(\frac{1-\rho}{2-\rho}\theta_H + \frac{1}{2-\rho}\theta_L\right) \\
&= -m^i + \theta_H - \frac{1-\rho}{2-\rho}\theta_H - \frac{1}{2-\rho}\theta_L \\
&= -m^i + \frac{1}{2-\rho}\theta_H - \frac{1}{2-\rho}\theta_L \\
&= -m^i + \frac{1}{2-\rho}\left(\theta_H - \theta_L\right)
\end{aligned}
$$

References

[1] Itai Arieli and Yakov Babichenko. Private Bayesian persuasion. *Journal of Economic Theory*, 182:185–217, 2019.

[2] Ritwik Banerjee, Nabanita Datta Gupta, and Marie Claire Villeval. Feedback spillovers across tasks, self-confidence and competitiveness. *Games and Economic Behavior*, 123:127–170, 2020.

[3] Roland Bénabou. Groupthink: Collective delusions in organizations and markets. *Review of Economic Studies*, 80(2):429–462, 2013.

[4] Roland Bénabou. The economics of motivated beliefs. *Revue d'Economie Politique*, 125(5):665–685, 2015.

[5] Roland Bénabou and Jean Tirole. Self-confidence and personal motivation. *Quarterly Journal of Economics*, 117(3):871–915, 2002.

[6] Roland Bénabou and Jean Tirole. Mindful economics: The production, consumption, and value of beliefs. *Journal of Economic Perspectives*, 30(3):141–164, 2016.

[7] Dirk Bergemann and Stephen Morris. Information design. *Journal of Economic Perspectives*, 2019.

[8] Markus K. Brunnermeier and Jonathan A. Parker. Optimal expectations. *American Economic Review*, 95(4):1092–1118, 2005.

[9] Kfir Eliaz and Ran Spiegler. Can anticipatory feelings explain anomalous choices of information sources? *Games and Economic Behavior*, 56(1):87–104, 2006.

[10] Matthew Gentzkow and Emir Kamenica. Costly persuasion. *American Economic Review*, 104(5):457–462, 2014.

[11] Matthew Gentzkow and Emir Kamenica. A rothschild-stiglitz approach to Bayesian persuasion. *American Economic Review*, 106(5):597–601, 2016.

[12] Matthew Gentzkow and Emir Kamenica. Competition in persuasion. *Review of Economic Studies*, 84(1):300–322, 2017.

[13] Russell Golman, David Hagmann, and George Loewenstein. Information Avoidance. *Journal of Economic Perspectives*, 2017.

[14] Daniel Gottlieb. Imperfect memory and choice under risk. *Games and Economic Behavior*, 85(1):127–158, 2014.

[15] Jonas Hedlund. Bayesian persuasion by a privately informed sender. *Journal of Economic Theory*, 167:229–268, 2017.

[16] David Huffman, Collin Raymond, and Julia Shvets. Persistent overconfidence and biased memory: evidence from managers. *Unpublished manuscript*, pages 1–97, 2019.

[17] Emir Kamenica and Matthew Gentzkow. Bayesian persuasion. *American Economic Review*, 101(6):2590–2615, 2011.

[18] Botond Köszegi. Ego utility, overconfidence, and task choice. *Journal of the European Economic Association*, 4(4):673–707, 2006.

[19] Botond Köszegi. Utility from anticipation and personal equilibrium. *Economic Theory*, 44(3):415–444, 2017.

[20] M M Markus, Muriel Niederle, and Tanya S Rosenblat. Managing Self-Confidence : Theory and Experimental Evidence. 2011.

[21] Markus Mobius, Muriel Niederle, Paul Niehaus, and Tanya Rosenblat. Managing Self-Confidence. 2014.

[22] Eduardo Perez-Richet. Interim Bayesian persuasion: First steps. *American Economic Review*, 104(5):469–474, 2014.

[23] Joshua Tasoff and Kristof Madarasz. A Model of Attention and Anticipation. *SSRN Electronic Journal*, pages 1–47, 2015.

[24] Yun Wang. Bayesian Persuasion with Multiple Receivers. *SSRN Electronic Journal*, pages 1–30, 2015.