Identities, values and persistent wrong beliefs

Yves Le Yaouanq *

This version : April 30, 2013.

Abstract

This paper studies the conditions under which individuals can sustain a distorted view of scientific phenomena. The baseline model features a social or individual decision for which both objective knowledge and idiosyncratic preferences are at stake. In a political context, the agents form motivated beliefs over their environment in order to protect their *values*, or their view of the ideal society. It is shown that the possibilities of persisting in a wrong belief across time crucially depend on the shape of social interactions : uncertainty surrounding the others' *values* or close-knit social networks limit the *ex post* updating, allowing some social groups to maintain a distorted cognition and potentially giving rise to multiple equilibria. Links with the empirical literature on collective beliefs are discussed. A second part explores the individual incentives to update one's own beliefs after some decision has been made (or some opinion expressed). The main result is that, if the individual's idiosyncratic preference (one's *identity*) is fragile and subject to imperfect recall, the *ex post* objective information might be selectively denied in order to maintain a positive self-view.

Keywords: identities, values, collective beliefs, collective denial, cognitive dissonance, political polarization, risk perception, cultural theory, ex post rationalization

*Le Yaouanq: Toulouse School of Economics. TSE, Université de Toulouse, Manufacture des Tabacs, 21 allée de Brienne, Fr - 31000 Toulouse. E-mail: yves.le-yaouanq@polytechnique.org. I am grateful to Christian Gollier, Eric Mengus, Sébastien Pouget, Jean Tirole and Karine Van der Straeten for their very helpful comments.

"*Views about climate change continue to be sharply divided along party lines. A substantial majority of Democrats (79%) say there is solid evidence that the average temperature on earth has been increasing over the past few decades, and 53% think the earth is warming mostly because of human activity. Among Republicans, only 38% agree the earth is warming and just 16% say warming is caused by humans.*" (Pew Research Center, *Wide Partisan Divide Over Global Warming*, October 27, 2010. Online publication)

1 Introduction

A common observation in social sciences is that laypeoples' beliefs regarding scientific phenomena do not always reflect the evaluation made by professional scientists. Striking examples have been reported, mostly in the field of risk assessment (e.g. nuclear power, GMOs, hazardous waste sites, climate change), but also for many topics including both scientific knowledge and personal values (e.g. struggle between neo-creationists and evolutionists, debate over handgun possession in the US). The explanation traditionally relies on the lack of scientific literacy in the population. However, recent works by psychologists have rejected this explanation, and highlighted the role of *values* or *identities* in the formation of beliefs. For instance, citizens' beliefs regarding the veracity and the origin of the climate change tend to polarize according to the individuals' intrinsic predisposition towards government intervention in the economy. In another context, the explanation based on the lack of numeracy fails to explain why countries with similar education levels present very different patterns of cognition regarding the origin of the species.

This paper builds a theory of motivated cognition under the influence of *values* or *identity*. It proposes a model of why people might form an opinion that differs from the scientific messages, and that confirms their preexisting worldview ; of why the mistrust towards the insights made by science may considerably vary from one social group to another ; of how these motivated beliefs can be sustained accross time, as a function of the shape of social interactions ; and of what this pattern of social cognition implies for effective communication. In the baseline model, an individual or a group has to make a binary decision in a context including both an unknown environment variable and some idiosyncratic preference. The distribution of *tastes* in the society is heterogeneous : some types have dominant strategies (always prefer some decision over the other), while some other are responsive to the payoff-relevant information.

In the political model of section 4, the individuals have heterogeneous prior preferences regarding one decision, and distort their cognition due to the existence of anticipatory feelings

at an interim stage. Since some objective information deflects the decision in one direction, those who *a priori* do not share this view have an incentive to deny the truth for the purpose of *identity-protection*. Contrary to the existing literature on anticipatory feelings, the agents do not care about being well informed *per se*, but only about the personal and social rewards derived from the expression of their opinion. In a benchmark where the distribution of values is known by all the agents, and under some additional assumptions, the pattern of social cognition is unique and is such that the agents who have the strongest identity selectively deny the information.

However, in such a framework, the social cognition would be immediately restored by the expression of opinions, leading all the *biased* agents to recognize their mistake. That is why I extend the benchmark model to a multiperiod setting, in order to study the possibility of sustaining a wrong belief over time in spite of the social interactions. Several forces are shown to help some social groups to maintain their distorted view of reality. First, if the assumption of common knowledge of the values is relaxed, it is always possible for a *biased* agent to rationalize the expression of opinions by attributing the large support for a theory as a proof that a large part of the society has itself *ideologically* motivated beliefs. In this case, the treatments of scientific information by the members of a social group are shown to be strategic complements : the more they deny the truth, the easier it is for their peers to sustain the associated cognitive dissonance. Applications include, for instance, the rise of neo-creationnism in several countries, or the perception of dangers linked to radioactivity in France in the XXth century. Second, the shape of the social network plays an important role : the more the individuals interact with people sharing their *values*, the less they receive contradictory information, and the easier it is to rationalize their cognition. These findings are linked to experimental evidence in social psychology.

Finally, section 6 studies more precisely the individual attitudes towards some *ex post* information received after a decision has been made (or, equivalently, after an opinion has been expressed), and provides an alternative explanation for the correspondence between *values* and beliefs. It explores, in an individual choice setting, the possibility of maintaining a distorted cognition if some explicit and perfect feedback over the environment is received. In this context, the idiosyncratic preference is called an *identity* and represents a personal taste regarding some decision (e.g., smoking, eating GMOs, mountaineering) or some opinion (e.g., for or against regulatory measures). There is no *ex ante* incentive to distort ones cognition, and the heterogeneity in the decision stems both from different preferences and from different levels of understanding of the scientific messages. After the choice has been made, the agent has imperfect recall about the reasons that drove his past choice. When some perfect feedback

signal about the environment is received, and if the agent cares about appearing smart, there exists an incentive to selectively deny this feedback in order to maintain the illusion that the decision was the right one, perhaps at the cost of wrong decision-making in the future. In other words, rationalizing one's behavior leads to deny the news that cast doubt on the relevance of past choices and positions. This result also sheds light on a broad pattern of individual perceptions regarding risky activities, from the refusal to admit the dangers of cigarette to the deliberate exposure to some risks by teenagers.

2 Self deception and cultural cognition : insights from economics, sociology and psychology

2.1 Related evidence

The question of what determines collective beliefs is central in social sciences. Considerable evidence in various fields shows that the laypeople's perceptions of scientific phenomena often do not correspond to the experts' opinion, and that this discrepancy does not vanish with communication of reliable objective information, nor with more understanding of the facts. A striking and important example is the opinion towards climate change. Although some famous skeptical experts bring contradiction in the debate, the scientific community is close to a full consensus about the reality of the climate change and its anthropogenic origin[1]. On the contrary, the public opinion remains much more divided : according to a Gallup survey, only 49% of the Americans believe in the causal relationship between human activities and the climate. The traditional explanation of the lack of understanding of the scientific content is not helpful to account for these puzzles, since countries with very similar education levels can experience very different patterns of cognition (see figure 2.1). Moreover, inside one country, the political affiliation (Democrats vs Republicans) has proved to be a much better determinant of one's opinion than the education level. Among the Republicans, the denial of the scientific consensus is so strong that only 38% of them agree that the average temperature has been growing compared with the pre-industrial level. The same polarization occurs for the opinion towards evolutionist theories : 60% of Republicans believe that God created the human being as it is 10 000 years ago, against only 38% for Democrats.

[1]Several polls or surveys mention a proportion of opponents of the ACC theory (anthropogenic climate change) between 1% and 5%, see for instance Farnsworth and Lichter (2012) and Anderegg et al. (2010).

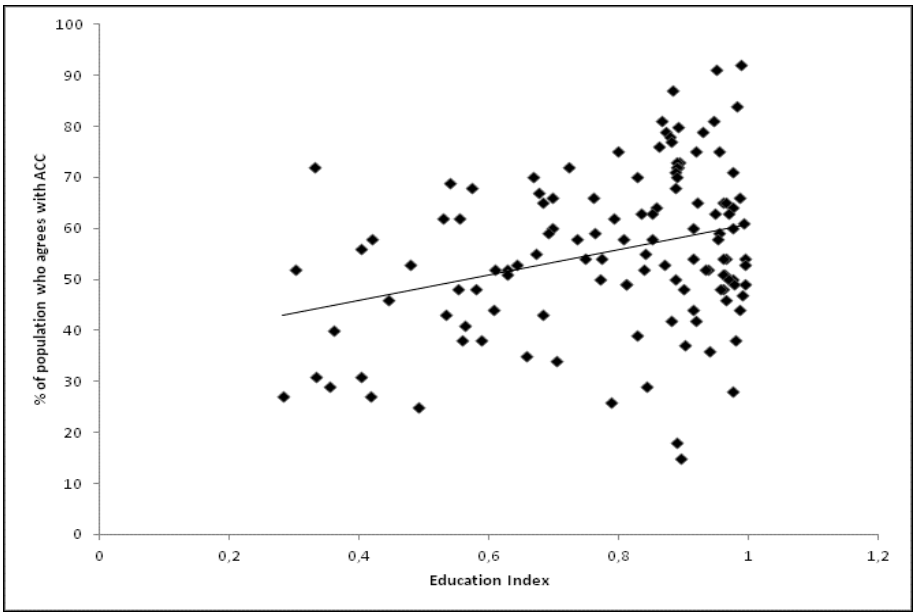


Figure 1: Education Index / Prevalence of the ACC opinion

The implication of political values or worldviews in the cognitive process is particularly stringent in the field of risk perception. Laypeople's perceptions of dangers and expert's evaluations can be very different, with the former relying more on affective concerns to build their opinion. For instance, Kahan et al. (2010) shows that people's beliefs regarding the objective benefits and costs of the vaccination of school girls against Human Papillomavirus (a sexually transmitted disease) are influenced by their view of the ideal social structure : those who tend to favor a hierarchist and individualist society tend to see lower benefits and greater risks than the "communitarian" and "egalitarian" types. This explanation relying on *values* has been first proposed by Douglas and Wildavsky (1982) and Douglas (1994) under the name Cultural Theory, formalizing the idea that the construction of a risk is the result of the expression of various preexisting conceptions of the society. The influence of values in the cognitive process has been validated in experiments for beliefs regarding climate change (Kahan et al. (2012)), nanotechnologies, gun possession (Kahan (2012)) or nuclear power (Plous (1991)). A striking result of this literature is that the polarization of beliefs does not vanish where more objective information is provided, even for people demonstrating a high level of scientific literacy and numeracy. At the individual level, a highly documented fact is the overconfidence manifested by most people engaged in risky activities : among marihuana smokers (Peretti-Watel (2001)), mountaineers (Loewenstein (1999)), or homosexuals after the appearance of AIDES (Pollak (1988)), many individuals disregard self-protection measures on behalf of their "identity", and in spite of considerable evidence. Contrary to the predictions of standard economic theory, the cognition process seems to be self-interested and biased for

signalling motives.

Moreover, the importance of individual misperceptions for appropriate policy-making has also been called into question, by Portney (1992) among others. Such patterns of individual behaviors and beliefs have indeed been shown to exert a great influence on the regulatory outcomes in practice. A recent striking example, in France, is the case of electromagnetic field produced by relay antennas. In spite of numerous scientific studies that found no evidence of an adverse effect of antennas on human health, an important citizen movement came into being in order to call for regulatory measures. The government eventually decided to impose a minimal safety distance between the antennas and the hospitals, the schools and the nurseries, while at the same time insisting on the harmlessness of the electromagnetic fields. In the US, Breyer (1995) and Viscusi and Hamilton (1999) document that the protection expenditures undertaken by the American Congress faithfully reflect public fears, even when the latter contradict existing scientific evidence. Both argue that the citizens' misperceptions of the risks take root in the cognitive limitations (e.g., availability bias, difficulties to manage probabilities) that hinder their decisionmaking, and, consequently, that efficient risk regulation should simply ignore this question. Conversely, Salanié and Treich (2009) show that even a rational regulator concerned with efficiency should take public perceptions into account, since the effect of the regulatory measures ultimately depends on the citizens' reactions. Hence, it is crucial to understand why individual behaviors depart from the paradigm of a rational decisionmaker provided by the neoclassic economic theory.

2.2 Related theories

This paper is linked to three streams of scientific literature. The first and recent one incorporates identity concerns in economic models. Starting with Akerlof and Kranton (2000), this literature typically assumes that the agents are not only driven by their personal preferences, but also by their ideal of "how one should behave" according to social considerations. These ideas have been mostly formalized by adding directly in the utility function a cost of making a decision that does not fit the expected decision of an (exogeneously defined) category. This specification has been applied in various contexts, as for ethnic disparities in educational success (Austen-Smith and Fryer (2005)) or individual risk and time preferences (Benjamin et al. (2010)). In this vein, the work closest to this paper is Bénabou and Tirole (2011), who provide a model in which "identity-driven" decisionmakers signal their social preferences (i.e., their altruism, "what kind of a person they are") to their future selves by investing in their social relationships. Their model also sheds light on intriguishing facts, such as ostracism of

"sinners" or "do-gooders", by explaining how these behaviors can arise from the willingness to ex post rationalize one's past actions. This paper shares with this literature the emphasis on *identities* and *values* to explain intriguishing social phenomena. However, this literature generally starts from the fact that people signal their type by engaging in some costly behavior, while the emphasis here is put on the manipulation of beliefs.

The second body of literature relates to the intrapersonal manipulation of information. Psychologists have documented the tendency of the human beings to deny "uncomfortable information". A lot of work has been devoted to the instrumental value of the cognition, with the underlying idea that pushing up one's self-esteem or personal motivation helps to overcome motivation problems (see for instance Tetlock et al. (2000)). Economic applications of this idea range from the avoidance of information at the individual level (Carrillo and Mariotti (2000), Bénabou and Tirole (2002)) to the explanation of the presence of different patterns of redistribution between countries (Bénabou and Tirole (2006)). While keeping the possibility for beliefs to have some instrumental value, the present work is mainly concerned with *affective* reasons that lead individuals to come round to wrong beliefs. This "wishful thinking" phenomenon, arising for instance due to the presence of anticipatory feelings regarding future prospects, has been used in a variety of setting : portfolio choices (Brunnermeier and Parker (2005), Brunnermeier et al. (2007), Gollier (2007)), contagious denial of reality (Bénabou (2009)), political agency issues (Levy (2012)). In this work, the emphasis is put on the factors that help the agents to sustain a distorted view of reality. In that sense, it is closely linked to the concept of *cognitive dissonance*, first proposed by Festinger (1957) to describe the situation where incompatible sources of cognition (here, the *affective* motives and the *objective* information) create some disutility for the agent and lead him to deny some truth.

Last, this work is linked to several papers dealing with the polarization of beliefs in the society. Glaeser and Sunstein (2013) propose a theory of why presenting the same information to a group of people can increase the polarization of opinions. They propose two mechanisms : *Asymmetric Bayesianism*, which includes an assessment of the credibility of the source by the individuals, and *memory boomerang*, which states that some signal might recall the history of past information that led the agents to form their opinion. These explanations are different, and hence complementary to ours : they consider the political orientations as *prior beliefs*, while they are here modelled as *idiosyncratic tastes*, and they exclude the *motivated* side of the cognition process. One crucial aspect of the present work is that the divergence of opinions does not stem from heterogeneous past beliefs nor from differing past information : instead, the situation that we consider is one in which agents all have the same prior opinion, and receive the same signal (for the first time). We believe that this feature, together with

the emphasis put on affective concerns, makes the model appealing for situations involving both complex science (origin of the species, climate change, nanotechnologies) and deep affective implications. It is also consistent with experimental findings suggesting that motivated reasoning is the result of a self-deception mechanism : for instance, Ditto et al. (1998) show that the selective skepticism regarding feedbacks about one's opinions is attenuated by the presence of a cognitive load, giving credence to the idea that a cognitive effort is incurred by the agents in the process of rationalizing preference-inconsistent information.

Other explanations for political polarizations, quite distinct from the present arguments, include the presence of information cascades, the structure of the media market (Mullainathan and Shleifer (2005)) or the fact that the heterogeneity in initial prior beliefs does not necesarily monotonically vanish to zero after a sequence of observations (Dixit and Weibull (2007)).

3 Baseline model

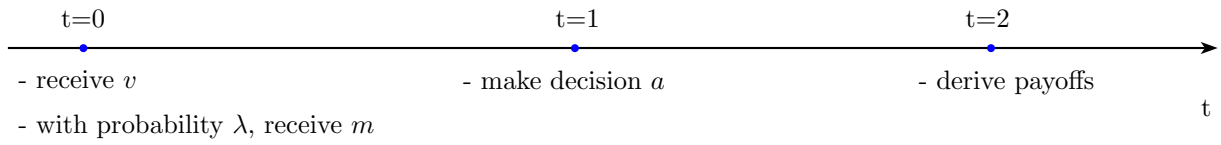
This section presents the baseline model of the paper, which will be enriched in the following sections.

Baseline payoff structure I consider a three-period economy where a decision $a = 1$ or $a = -1$ has to be made regarding some specific issue which involves both objective knowledge and individual characteristics. The state of the world is uncertain and is represented by the random variable X, which takes values $x > 0$ or $-x < 0$ with equal probabilities $\frac{1}{2}$ each. Each agent is endowed with an idiosyncratic exogeneous preference for the decision, denoted v, which will subsequently be interpreted either as a "value" or an "identity". If $v > 0$, the individual has a preference for decision $a = 1$, and if $v < 0$ he has a preference for decision $a = -1$, $|v|$ being a measure of the strength of this preference. More precisely, if decision a is made and if the state of the world is X, a v-type agent ultimately receives a baseline payoff equal to $\pi(v, X, a) = Xa + va$.

Timeline and information structure The basic timeline is as follows. At $t = 0$, with some probability $\lambda < 1$ the agent receives some informative signal about the state of the world, $m = x$ or $m = -x$. The quality of this signal is measured by the parameter $\pi > \frac{1}{2}$, such that $\mathbb{P}(X = x|m = x) = \mathbb{P}(X = -x|m = -x) = \pi$. π will be interpreted as either the quality of the common public signal, or as the agent's skill (i.e., the ability to properly understand the message). In the latter case, a fraction π of the population receives the good signal, while a fraction $1-\pi$ receives the wrong piece of information. I assume $\pi > \frac{1}{2}$. With probability $1-\lambda$,

no signal is received by the agent, or, equivalently, all the existing information is insignificant.

At $t = 1$, the choice is made between actions $a = -1$ and $a = 1$ (either an individual or a political decision). At $t = 2$, the state of the world is revealed and the payoffs are derived.



Distorted cognition : the demand side In addition to the baseline payoff $Xa + va$, whose terms reflect the *specific* payoff and the *idiosyncratic* preference, the agent is also endowed with some beliefs-contingent preferences related to his personal characteristic, which lead them to bias his cognition :

- **Identity-protective concern** : in the political model of section 4, the decision is made through a vote. I assume that some scientific information comes prior to the vote, after which the agents form expectations about the future political decision and derive some anticipatory utility (or disutility) from their beliefs. For instance, suppose that the agent is endowed with a high preference $v > 0$ for $a = 1$, and receives the signal $m = -x$. This signal is bad news, in the sense that it increases the likelihood that $a = -1$ will be chosen. Hence, the agent has an incentive to discard this piece of evidence in order to form rosier beliefs about the future decision.

- **Self-esteem concern** : in the individual version of section 6, I assume that some *posterior* information is received after some decision has been made. The driving force of the manipulation of beliefs is the desire to maintain a positive self-assessment. A large literature in psychology reports that the individuals derive some utility of sustaining a positive view of themselves. This assumption can be grounded on several motives : apart from a hedonic consumption value, a positive evaluation of one's ability helps to overcome motivation problems (as in Carrillo and Mariotti (2000) or Bénabou and Tirole (2011)), or improves one's future prospects. In this work, this taste is modelled by the presence of an additional term in the utilty function at $t = 2$, $\kappa\hat{\pi}$, where $\hat{\pi}$ is the ex post probability that the agent attaches to having received the correct information and κ is the magnitude of the self-esteem concern.

Distorted cognition : the supply side This paragraph describes the modelling of the cognitive process by which the agents can manipulate their own cognition according to the

reasons discussed above, either before or after the decision is made. In both cases, the agent is likely to receive some evidence about the state of the world in the form of a signal $m \in \{-x, x, \varnothing\}$ ($m = \varnothing$ stands for the no signal case). Several modelling strategies are possible, which are all equivalent provided that they allow for a differential awareness of news $m = -x$ and $m = x$. I assume that the agent can only influence the probability of transmitting some evidence to his future incarnations. The natural rate of recall of a message $m \in \{-x, x\}$ is equal to 1, but the agent has the possibility to manipulate this probability and to choose a lower value $\sigma(m) < 1$. Hence, the signal transmitted to the future incarnation is either $\hat{m} = m$ or $\hat{m} = \varnothing$ according to the strategy $\sigma(m)$. The cost of this manipulation will be either exogeneous (related to a psychic cost of cognitive dissonance, or a cost of trying to avoid this signal if it is repeatedly provided) or endogeneous in the applications. If no signal was received, however, the agent has no choice to make and transmits $\hat{m} = \varnothing$ with probability 1.

4 Cultural values and social cognition

This part applies the baseline model exposed in section 3 to explore the phenomena of social cognition regarding scientific phenomena.

4.1 Interpretation of the model

Political decision In this case, the choice $a = -1$ or $a = 1$ refers to a political decision : for instance, tackling the climate change by mitigating the CO_2 emissions, allowing for cultivation of GMOs or deciding that some theory will be taught at school. The society is composed of a continuum of individuals of measure one, who all derive the same *specific* payoff Xa, but who have different idiosyncratic preferences. The parameter v represents a personal preconception of the ideal society, independently of the specific issue at stake here : relying on their prior information, some agents think that the decision $a = 1$ makes a step forward a "better" society, while some others think that decision $a = -1$ is preferable. After the aggregate decision a_{agg} is made, a v-type individual receives a payoff equal to $\pi(v, X, a_{agg}) = Xa_{agg} + va_{agg}$.

The individuals are indexed by their *value* v, which is distributed on $\mathbb{R}$ according to the atomless continuous distribution function $f(v)$ and the respective cumulative distribution function $F(v)$, which are common knowledge. For simplicity, I assume that the support of f is an interval.

Voting process At date $t = 1$, all the agents express their preferred opinion a. I assume that the probability of choosing a decision is increasing with the number of citizens who express this preference. This "smooting" assumption (which is made for instance in Grossman and Helpman (1996)) can be interpreted as the fact that the legislative power is allocated according to the proportional representation, and that the probability of implementing a political decision increases with the number of representatives who share this view. Formally, if N_1 is the fraction of the population expressing a preference for $a = 1$, the probability of implementing $a = 1$ is equal to $\phi(N_1)$, where ϕ is a continuously differentiable function such that $\phi' > 0$.

The agents vote non-strategically according to their true preference. Thus, if $|v| \geq x$, the individual is not responsive to the information and will always vote for his *a priori* preferred decision, while if $|v| < x$, a contradictory piece of evidence leads the individual to change his mind.

Self-deception As in the baseline model, the agents receive a public signal m at date $t = 0$[2]. The driving force for self-deception is that the agents form expectations about their future prospects, with respect both to the relevance of the political decision and to the extent to which this choice is compatible with their values[3]. The baseline model is enriched here by an interim stage $t = 0.b$ between dates 0 and 1, in which the agents receives a flow of anticipatory utility equal to $s_1\mathbb{E}\mathbb{E}(Xa_{agg}|\hat{m}) + s_2\mathbb{E}\mathbb{E}(va_{agg}|\hat{m})$, where $\hat{m}$ is the signal remembered from the first period ($\hat{m} = m$ or $\hat{m} = \varnothing$). The first expectation is taken with respect to the possible values of X, and the second with respect to the possible values of a_{agg} once the preferences have been formulated in the society. s_1 and s_2 can be thought of as being equal if the anticipatory utility relates to the total payoff $Xa_{agg} + va_{agg}$. However, different values of s_1 and s_2 will also be allowed for in applications, typically with s_2 being larger than s_1 in order to reflect the fact that the payoffs Xa and va are derived at different periods (va is received immediately after the decision is made, while Xa is received after some duration), or that the individuals have a greater tendency to anticipate the effect of the decisions on their *values* than the relevance of the decision with respect to the particular situation.

Information processing The manipulation of one's memory occurs between periods 0 and $0.b$. If a signal is received, the agents have the possibility to manipulate the probability of recalling it : agent v's strategy is a pair $(\sigma_+(v), \sigma_-(v))$, which denote the respective probabilities

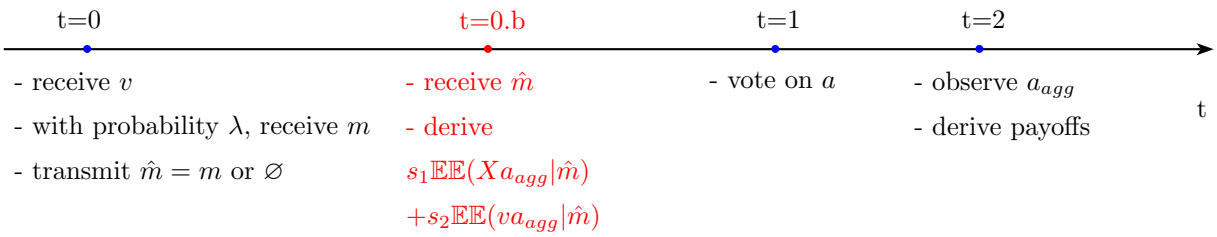
[2]This signal is the same for all players, so that the results have nothing to do with information cascades.

[3]Manipulating one's beliefs in order to improve one's anticipation of the state of the society is also a driving force in Bénabou (2008) and Levy (2012)

of recalling $m = x$ and $m = -x$. In this baseline model, the only force that counterbalances the distortion of reality is the presence of an exogeneous cost of cognitive dissonance $c > 0$ which is incurred by the agent as soon as the chosen rate of recall is lower than 1.

It is noteworthy that in existing models involving anticipatory utility, the agents typically trade off the pleasure to form rosy beliefs about the future against the adverse consequences for decision making that it implies. Here, on the contrary, each player individually has no role on the election outcome. Thus, the citizens do not take into account any benefit of being well informed for "proper voting" at the last stage. This assumption reflects the fact that the most salient consequence of one's belief regarding political decisions is the social and personal gratifications resulting from the expression of opinions.

The (modified) timing of the game is now as follows :



An equilibrium of the game is a set of acceptance strategies $(\sigma_+(v), \sigma_-(v))$ for all values of v, where $\sigma_+(v)$ (resp. $\sigma_-(v)$) denote the probability of transmitting the signal $m = x$ (resp. $-x$) to one's future incarnation.

4.2 Patterns of social cognition

This section derives the general pattern of social cognition that results from the game defined above.

Voting behavior In this model, a v-type agent receiving at date $0.b$ the information $\hat{m}$ votes for $a = 1$ if and only if $v + \mathbb{E}(X|\hat{m}, \sigma_+, \sigma_-) > 0$. Thus, if $\hat{m} = x$, the expected value of the state of the world is $\mathbb{E}(X|\hat{m} = x) = (2\pi - 1)x$, and the agent votes for $a = 1$ if and only if $v + (2\pi - 1)x > 0$ [4]. Similarly, if $\hat{m} = -x$, the agent votes for $a = 1$ if and only if $v - (2\pi - 1)x > 0$. In the following, deny $N_1^R(m)$ (R stands for *realism*) the fraction of

[4]The behavior of the agents that are indifferent between $a = -1$ and $a = 1$ does not matter here, since they have a marginal impact on the outcome of the vote. I assume that these agents vote for $a = -1$.

agents who would vote for decision $a = 1$ after signal m if no cognitive manipulation occurred : $N_1^R(-x) = \int_{(2\pi-1)x}^{+\infty} f(v)dv$, $N_1^R(\varnothing) = \int_0^{+\infty} f(v)dv$ and $N_1^R(x) = \int_{-(2\pi-1)x}^{+\infty} f(v)dv$.

For expositional simplicity, suppose here that $\pi = 1$, i.e. the signal is perfectly informative. Consider the case where the agent recalls the signal $\hat{m} = \varnothing$. Bayesian inference leads the agent to update the probability attached to both states of the world according to his personal cognitive strategies σ_+ and σ_-. By Bayes' rule, the posterior probability attached to the state $X = x$ is given by :

$$\mathbb{P}(X = x|\hat{m} = \varnothing) = \frac{\frac{1-\lambda}{2} + \frac{\lambda}{2}(1 - \sigma_+(v))}{1 - \lambda + \frac{\lambda}{2}(1 - \sigma_+(v)) + \frac{\lambda}{2}(1 - \sigma_-(v))}$$

The posterior probability attached to the state $X = -x$ is given by :

$$\mathbb{P}(X = -x|\hat{m} = \varnothing) = \frac{\frac{1-\lambda}{2} + \frac{\lambda}{2}(1 - \sigma_-(v))}{1 - \lambda + \frac{\lambda}{2}(1 - \sigma_+(v)) + \frac{\lambda}{2}(1 - \sigma_-(v))}$$

Thus, the posterior expectation of X is :

$$\begin{aligned}\mathbb{E}(X|\hat{m} = \varnothing) &= \frac{\frac{\lambda}{2}(\sigma_-(v) - \sigma_+(v))}{1 - \lambda + \frac{\lambda}{2}(1 - \sigma_+(v)) + \frac{\lambda}{2}(1 - \sigma_-(v))}x \\ &\equiv p(v) \text{ (def.)}\end{aligned}$$

If $\hat{m} = \varnothing$ (no signal is recalled), the agent votes for $a = 1$ if and only if $v + p(v) > 0$.

Incentives to deny Define $N_1(m)$ the proportion of votes for $a = 1$ if all the agents have received the signal m.

Suppose, that at date 1 the agent knows with certainty that the message $m = \varnothing$ was sent. In this case, no agent recalls a signal other than $\hat{m} = \varnothing$ and the agents who vote for $a = 1$ are those endowed with value $v > x$ and those with value $-x \leq v \leq x$ such that $v + p(v) > 0$:

$$N_1(\varnothing) = \int_x^{\infty} f(v)dv + \int_{-x}^{x} 1_{v+p(v)\geq 0} f(v)dv \tag{1}$$

Suppose now that the agent knows with certainty that $m = -x$ was sent. In this case, the agents who vote for $a = 1$ are those endowed with extreme-right value and those with intermediate value who do not recall the signal and are such that $v + p(v) > 0$:

$$N_1(-x) = \int_x^{\infty} f(v)dv + \int_{-x}^{x} (1 - \sigma_-(v))1_{v+p(v)\geq 0} f(v)dv \tag{2}$$

Similarly,

$$N_1(x) = \int_x^{\infty} f(v)dv + \int_{-x}^{x} (1 - (1 - \sigma_+(v)))1_{v+p(v)\geq 0} f(v)dv \tag{3}$$

Equations (1), (2) and (3) show that $N_1(-x) \leq N_1(\varnothing) \leq N_1(x)$. This shows that an agent endowed with $v > 0$ (who prefers to think that most citizens will vote for $a = 1$) tends to deny the signal $m = -x$ and to remember the signal $m = x$. The contrary holds for an agent with $v < 0$, who tends to deny $m = x$ and to remember the signal $m = -x$. In other words, people are inclined to deny the news that contradict their worldview. The appendix provides a complete proof of the following proposition.

Proposition 1. *In equlibrium, the cognition pattern is characterized by some thresholds* $(v_1 < v_2 < v_3 < v_4)$ *such that :*

(a) $\sigma_-(v) = 1$ *for* $v \leq v_3$, $\sigma_-(v) = 0$ *for* $v \geq v_4$ *and* σ_- *is affine on* (v_3, v_4)

(b) $\sigma_+(v) = 0$ *for* $v \leq v_1$, $\sigma_+(v) = 1$ *for* $v \geq v_2$ *and* σ_+ *is affine on* (v_1, v_2)

Moreover, if either

Assumption 1. *a)* $\max s_1, s_2 < \frac{c}{2x}$

or

b) $s_1 = 0$,

the equilibrium is unique and such that $(v_1 < v_2 < 0 < v_3 < v_4)$.

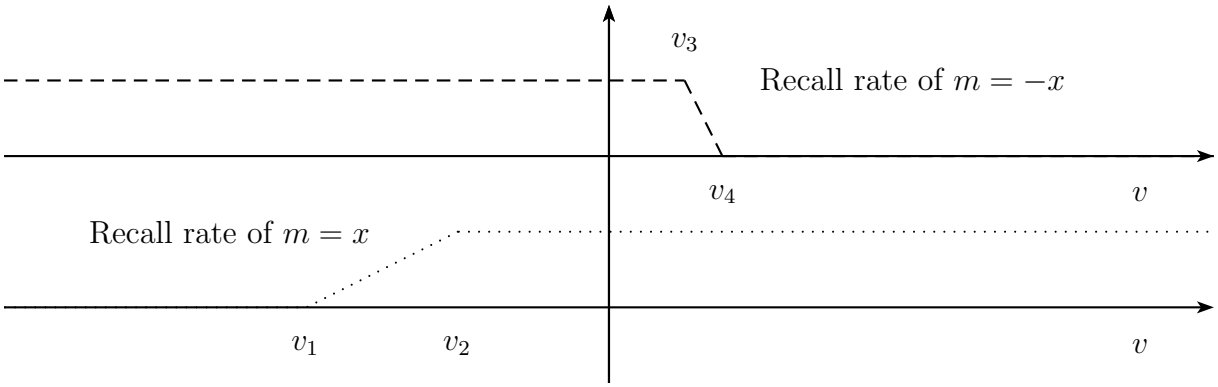


Figure 2: Equilibrium recall rates

This proposition shows that the more ideologically-oriented types bias their memory in favor of the evidence that conforts their worldview. Consequently, no consensus is obtained in general at the voting stage. This result is consistent with the common empirical finding that the social opinions tend to polarize, whatever the strength of the evidence presented to them. This very robust pattern has been noticed for instance for climate change (Kahan et al. (2012)), nanotechnologies (Kahan (2012)), nuclear power (Plous (1991)) or mobile antennas

(Borraz (2008)). A key ingredient of the model is that the distribution of opinions according to the underlying values arises even if the information transmitted to the agents is perfect, and even if they have a complete understanding of the arguments presented to them.

In the general case, the equilibrium is not unique since the voter's cognitive strategies are *complements* with respect to the specific payoff Xa : indeed, when more citizens deny some news, it becomes more likely that the wrong decision will be made, and the incentive to discard the payoff-relevant signal is, in turn, higher. This "treadmill effect" is reminiscent of the *Mutually Assured Delusion (MAD)* principle highlighted in Bénabou (2009) : when some agents ignore the reality (here, for ideological purpose) and vote against the evidence, this reinforces the others' incentive to engage in wishful thinking. It is hence, perfectly possible that in equilibrium even the neutral types ($v = 0$) discard some information, if the society is sufficiently *ideologically* biased towards the inverse decision.

In contrast, the voters' behaviors are *strategic substitutes* with respect to the *value-related* payoff va : more denial of the news $m = -x$ by the right-biased agents actually makes this prospect more desirable for a right-biased agent, lowering the incentive to engage in denial. Assumption 1 limits the extent of the strategic complementarity, and results in the uniqueness of the equilibrium. Notice that, under this assumption, the neutral agents (endowed with $v = 0$) do not distort their cognition.

4.3 Comparative statics

To derive comparative statics results, I will assume that assumption 1 holds, in order to ignore the issue of contagious wishful thinking and to concentrate on the expression of *values*. More specifically, I will restrict attention to equilibria that belong to one of the two following classes :

- Under assumption 1a), **Type A equilibrium** (*Denial by the extremely biased agents*) : it can be shown that $v_3 > x$ as soon as assumption 1a) holds. Hence, the voting pattern is identical to the case where no manipulation of information would occur. The *moderate* agents (such that $|v| < x$) vote according to the information received, while the *extremist* agents ($|v| > x$) follow their prior. Among the latter group, some chose to deny the uncomfortable signals while some other (those with $x < v < v_3$) keep their memory unbiased but vote for the same decision ($a = 1$) in all cases. This situation occurs when the cost of cognitive dissonance is high with respect to the issue at stake, and when only a small group with a strong ideology denies the reality. Potential applications include sectarist movements or conspiracy theories.

- Under assumption 1b) **Type B equilibrium** (*Denial by moderately biased agents*) : I assume that the parameters of the model are such that $v_3 < v_4 < x$ (see the Appendix). Hence, some agents switch their vote due to their denial (those with $v_3 < v < x$ who forget the news $m = -x$). Hence, the agents who vote for $a = 1$ after the signal $m = -x$ are those such that $v_3 < v \leq v_4$ (with some positive probability) and those such that $v_4 < v$ (with probability 1). This situation is likely to occur when the issue has a strong meaning in terms of identity (so that s_1 is small with respect to s_2), leading a large fraction of the population to engage in wishful thinking. Examples such as death penalty or environmental regulation are in this scope.

Notice that, in both equilibria, if no signal is provided ($m = \varnothing$) the society perfectly polarizes and $N_1(\varnothing) = \int_0^{+\infty} f(v)dv$.

Consider now a **type A** equilibrium, where $s_1 = s_2$ to simplify (but without loss of generality). An individual endowed with value $v > 0$ has the following net incentive to deny the news $m = -x$ (see the appendix) :

$$s_1 \mathbb{P}_{\varnothing}(2\pi - 1)x(2\phi(N_1(-x)) - 1) + 2s_2 v(\phi(N_1(\varnothing)) - \phi(N_1(-x))) \tag{4}$$

The expression outside the brackets, $\mathbb{P}_{\varnothing}$, represents the ex post probability that the agent attaches to the state $m = \varnothing$ (no scientific discourse is relevant) after recalling $\hat{m} = \varnothing$ and processing Bayesian updating. If the agent is fully naive, this probability is equal to 1, whereas for a Bayesian agent it is lower than 1, but positive.

The expression in brackets represents the net gain from distorting one's belief towards the state $m = \varnothing$ after having received the evidence $m = -x$. The first term represents the anticipation of the relevance of the political decision : if $2\phi(N_1(-x)) - 1 > 0$, i.e. if the vote for $a = 1$ remains majoritarian even after seeing some evidence that $X = -x$, this gain is positive since it attenuates the prospect of wrong decision making.

The second term in the brackets is the gain in anticipated beliefs from thinking that more people will vote for $a = 1$, which is beneficial for an individual with positive value $v > 0$. Obviously, this incentive is increasing with the attachment to the value measured by the magnitude of v.

Consider now a change in the quality of the public information, namely an increase in π. Equation (4) makes clear that two effects are at play. Observe first that $\phi(N_1(-x))$ decreases if π increases, since some moderate people endowed with a positive value $v > 0$ switch from vote $a = 1$ to vote $a = -1$ as the evidence that $X = -x$ becomes more significant. Hence, the information $X = -x$ becomes *more frightening* for an individual with positive v, and the

value-incentive $2v(\phi(N_1(\varnothing)) - \phi(N_1(-x)))$ also becomes higher. Thus, the *value-protecting* effect is such that more information always worsens the pattern of social cognition.

The other effect, through the first term of equation (4) is related to the anticipation of the relevance of the political decision, and makes clear how the *MAD* principle works. Suppose, indeed, that $\phi(N_1(-x)) > \frac{1}{2}$, i.e. that the society is right-biased and chooses by a majority to vote for $a = 1$ in spite of some evidence that $X = -x$. In this case, a more significant signal actually makes it more likely that the society will make the wrong decision, and the incentive to deny this gloomy prospect is higher. In this case, the *value-protecting* effect and the *anticipation* effect reinforce each other and a direct consequence is the paradoxical feature that more information is detrimental to the social cognition. On the other hand, if $\phi(N_1(-x)) < \frac{1}{2}$, the two effects work in opposite directions and the result is ambiguous.

Consider now a switch in the distribution of values inside the society. Under the assumption $2sx < c$, the threshold v_3 is explicitly defined by

$$v_3 = \frac{\frac{c}{s} - (2\pi - 1)x(2\phi(1 - F((2\pi - 1)x)) - 1)}{2(\phi(1 - F(0)) - \phi(1 - F((2\pi - 1)x)))}$$

And v_4 is equal to v_3 times a constant term. What matters in the distribution of values is only the number of individuals who are *moderately* right-biased $(1 - F(0))$ and the number of individuals who are *extremely* right-biased $(1 - F((2\pi - 1)x))$. *Ceteris paribus*, an increase in the former decreases v_3 and v_4 and worsens the social acceptance of the news $m = -x$: indeed, the incentive to forget the scientific message is higher since more agents would vote for $a = 1$ in the absence of such a message. An increase in the number of *extremely* right-biased agents leads to an increase in v_3 and v_4 : hence, the first agent who biases his opinion has a higher value v. However, the total number of agents with distorted cognition is ambiguous.

Consider now a **type B** equilibrium. As soon as $v_3 < (2\pi - 1)x$, the proportion of agents who vote for $a = 1$ after signal $m = -x$ is independent of π, since it is exactly the agents with $v \geq v_3$ and who remember no signal. Since, in addition, $s_1 = 0$, the value of π (or, equivalently, of x), has no influence on the equilibrium pattern.

The comparative statics of the distribution of the values is the same as in a type A equilibrium : *ceteris paribus* an increase in the number of *extremely* right-biased agents (those with value $v > (2\pi - 1)x$) increases v_3 and v_4, while an increase in the number of *moderate* individuals (those with value $0 < v < v_3$) leads to a decrease in v_3 and v_4.

Proposition 2. *1. In a **type A equilibrium**, if the society is sufficiently right-biased, i.e. if $\phi(\int_{(2\pi-1)x}^{+\infty} f(v)dv) > \frac{1}{2}$, then a marginal increase in the quality of the signal $m = -x$ decreases the acceptance of the scientific evidence $X = -x$. In a **type B equilibrium**, a marginal increase in π has no impact on the social cognition.*

2. *In both equilibria, the threshold v_3 increases with the number of extremely right-biased agents and decreases with the number of moderately right-biased agents.*

This paradoxical result states that a more significant signal can have no effect, or even worsen the social acceptance of the scientific consensus. This finding is in line with a great deal of experimental evidence that reveals that objective information does not lead the subjects to align their positions.

An important implication of this result is that providing the citizens with reliable information is not sufficient to spread the scientific messages in the society. This does not mean that transparent communication is not desirable, but it implies that the framing of the messages is crucial : official communication should try not to "hurt" any social group's values, by the choice of an appropriate vocabulary for instance. Recent experimental evidence by Hardisty et al. (2010) confirms, for instance, that the Republicans react much more positively to an increase in the price of various goods (airline tickets, gasoline etc.) designed to reduce the CO_2 emissions and finance mitigation measures, when such increase is labelled an "offset" rather than a "tax". In contrast, the Democrats react similarly to both terms. According to this theory, this happens because the wording "tax" conveys some negative overtone for some part of the political spectrum, while the label "offset" breaks the association between one decision and the ongoing political values.

4.4 Welfare analysis

What is the effect of the manipulation of cognition on society's welfare ? To answer this question, I suppose that the society has a free commitment device to force its member to accept the evidence, and I compare the individuals' welfare in this hypothetical situation with their welfare in the equilibrium of the game described above. In this subsection too, I will consider type A and type B equilibria under assumption 1.

To begin with, notice that the answer is straightforward in a type A equilibrium : since the manipulation of information does not change the social decision (only the extreme types bias their memory), the equilibrium strategies maximize welfare. The cognitive manipulations have no practical implications, except that they allow some group to form rosier expectations.

Things are more complicated in a type B equilibrium, since the decision is influenced by the agents' denial. Consider the welfare after message $m = -x$, and consider an agent of type v. If everyone is realistic (including himself), his expected payoff equals

$$\pi(v)^R = -\delta(2\pi - 1)x(2\phi(N_1^R(-x)) - 1) + (s_2 + \delta)v(2\phi(N_1^R(-x)) - 1)$$

In the denial equilibrium with threshold v_3 and voting parameter $N_1^D(-x)$, the welfare of an agent who would behave as a realist (for instance, $v < v_3$) is

$$\pi_{realism}(v)^D = -\delta(2\pi - 1)x(2\phi(N_1^D(-x)) - 1) + (s_2 + \delta)v(2\phi(N_1^D(-x)) - 1)$$

Since $N_1^D(x) > N_1^R(-x)$, there exists some threshold v^* such that the $\pi(v)^R < \pi_{realism}(v)^D$ if and only if $v > v^*$.

However, if the agent is such that $v > v_3$ and forgets the information, his welfare equals

$$\pi_{denial}(v)^D = -\delta(2\pi-1)x(2\phi(N_1^D(-x))-1)+s_2v(2\xi(v)\phi(N_1^D(\varnothing))+2(1-\xi(v))\phi(N_1^D(-x))-1)$$

Where $\xi(v)$ is the ex post probability attached to the original message $m = \varnothing$ (depending on agent v's cognitive strategy). There exists a threshold v^{**} such that the welfare in the realism equilibrium is higher if and only if $v < v^{**}$. Hence, the following conclusion :

Proposition 3. *In a type A equilibrium, realism has no influence on the welfare of agents such that $v \leq v_3$, and strictly decreases welfare for agents such that $v > v_3$. In a type B equilibrium, the welfare after message m is strictly higher in a realism equilibrium if and only if v is lower than a certain threshold, and strictly lower otherwise.*

The intuition for this result is straightforward : for small values of v, for instance $v < 0$, the welfare in the *realism* equilibrium is higher since the outcome of the vote is more favorable (it both leads to an *objectively* better decision and is more compatible with one's values). However, for large values of v, unanimous *realism* is not a good option as soon as the value-related incentive to deny is high enough.

5 Sustainable distorted cognition

This section extends the previous model to a multiperiod setting, in order to study the possibilities of maintaining one's biased beliefs in spite of the social interactions. The motivation for this section is that the pattern of social cognition, characterized by (v_1, v_2, v_3, v_4) was pinned down by the distribution of values in the society. Hence, in a multiperiod setting, a rational individual observing the distribution of votes at $t = 2$ should infer the true value and update his beliefs accordingly : the collective error should immediately vanish. Moreover, discrepancies in preexisting values can explain why different patterns of perception of scientific issues arise in different societies. However, why these perceptions differ accross time in one given country, in other words why collective mistakes appear or disappear, is not captured by the above theory.

This section studies two forces that help the *biased* agents to maintain their rosy beliefs in a multiperiod setting : the uncertainty surrounding the society's values, and the influence of the social network. All the results described in this section hold with more general models than the one studied in section 4, provided that the result of the period 1-manipulations is that some agents have formed a distorted belief due to their *ideological* preferences. For instance, the self-esteem model of section 6, relying on the fact that agents with uncertain preferences rationalize their past opinions and choices by selectively denying some information, provides the same starting point for this section.

5.1 Uncertain values and incomplete updating

This paragraph shows that the wrong belief can be (at least partially) sustained in a multiperiod setting, if the distribution of values in the society is not perfectly known by the agents. As an example, consider the following simple case, in which v can take only three values, and the distribution is drawn among two possibilities :

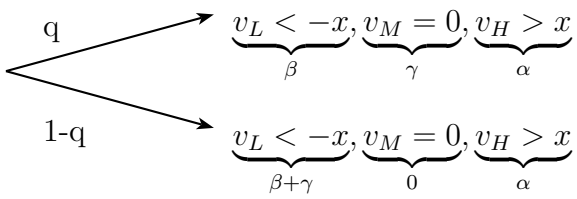


Figure 3: Distribution of values in the society

Hence, it is common knowledge that there exist exactly α extremely right-biased types, and at least β extreme left-biased types, but there exists some doubt over the values of the γ remaining agents. Suppose that the scientists produce some evidence $m = -x$ with quality π. The outcome of the vote is that, with probability 1, $\beta + \gamma$ agents vote for $a = -1$. Suppose that a v_H-type agent had chosen to deny the signal $m = -x$. Recalling $\hat{m} = \varnothing$ and observing that $N_{-1} = \alpha$ does not lead the agent to perfectly update his belief about X : indeed, there is some positive probability that the $\beta + \gamma$ agents who voted for $a = -1$ where composed only of extreme-left agents. More precisely, the ex post probability attached to the fact that $m = \varnothing$ has been sent is positive and equal to :

$$\mathbb{P}(m = \varnothing | \hat{m} = \varnothing, \beta + \gamma \text{ vote for } a = -1) = \frac{(1-q)(1-\lambda)}{(1-q)(1-\lambda) + \frac{\lambda}{2}}$$

Hence, if the equilibrium strategies are those described above, some information about X is necesarily lost forever in equilibrium by the types who have an incentive to deny the news,

even in a multiperiod setting. In words, observing that $1-\alpha$ citizens vote for $a=-1$ does not lead the v_H-type agents to completely update their beliefs, since there is a positive probability for this group to be composed only of extremely biased agents, in which case their opinion is uninformative about the state of the world.

Multiple equilibria The mechanism described in the previous paragraph also gives birth to multiple equilibria in a two-stage setting. Throughout this section, I assume that some psychic cost is not incurred when the cognition is manipulated, but when some *ex post* information contradicts the encoded memory. This assumption is very much in line with the theory of cognitive dissonance (see Festinger (1957) for the seminal work), described by the psychologists as a situation of internal fight between incompatible sources of cognition : the cognitive dissonance happens when one's "moral integrity" is threatened by some information. This modelling choice reflects the fact that the psychic cost of cognitive dissonance is incurred more than once, and that all the information related to the issue might be concerned. For a wrong belief to be sustained over time, it must be that the intertemporal cost of maintaining it in spite of contradictory evidence is not too high, otherwise this belief is revised.

More precisely, here I assume that the *ex post* information takes the form of the expression of agents' preferences (by voting, or by polls), which convey some valuable information about the original signal m. If an agent has some nonnull message m in his memory, the *ex post* psychic cost equals zero ; but if his memory is empty, and if the distribution of opinions leads him to infer that some nonnull message m was sent with probability p, he incurs a psychic cost of cp, where c is here the magnitude of the *ex post* of cognitive dissonance.

The distribution of values is described in figure 5.1, except that $0 < v_H < x$. Suppose also that, when no signal is received, half of the moderate agents (when they exist) vote for $a=1$, the other half for $a=-1$.

Consider the case where the message $m=-x$ has been produced, and suppose (as is shown in the previous section) that the neutral and left-biased types recall this evidence, i.e. that $\sigma_-(v_L)=\sigma_-(v_M)=1$. The next result characterizes the behavior of the right-biased types v_H.

Proposition 4. *If c is high, the unique equilibrium is the acceptance of the consensus, i.e. $\sigma_-(v_H)=1$. If c is low, the unique equilibrium features complete denial, i.e. $\sigma_-(v_H)=0$. Moreover, there exists an interval I such that, if $c \in I$, both strategies ($\sigma_-(v_H)=1$ and $\sigma_-(v_H)=0$) are sustainable in equilibrium.*

A complete proof is provided in the appendix. The presence of multiple equilibria relies on

the fact that the low types' decisions are strategic complements[5].The proposition states that there are two possible states of the world. In the first one, $\sigma_-(v_H) = 1$ and all the right-biased types accept the uncomfortable signal $m = -x$. Consequently, the outcome is that all agents vote for $a = -1$. Suppose that a v_H-type agent chooses to deviate and play the strategy $\sigma_- = 0$. At date $t = 1$, this agent will be the only one to express a preference for $a = 1$. Maintaining his belief biased is hence very costly, since the agent infers with probability 1 that the public message $m = -x$ had been sent. Hence, the associated psychic cost is high, equal to c.

In the second equilibrium, $\sigma_-(v_H) = 0$ and all the right-biased types deny the signal $m = -x$. Hence, the outcome of the vote is $N_1(-x) = \alpha$. *Ex post*, a v_H-type agent observing this outcome is able to rationalize this situation by thinking that the true distribution of values is the second one, with no moderate types, and that no public message was sent. Indeed, the posterior probability attached to the public message $m = -x$ now equals $\frac{\frac{\lambda}{2}}{(1-q)(1-\lambda)+\frac{\lambda}{2}}$, and the associated psychic cost equals $c\frac{\frac{\lambda}{2}}{(1-q)(1-\lambda)+\frac{\lambda}{2}} < c$. This cost is lower than in the full-consensus equilibrium described above, which means that more denial helps the agents who would like to manipulate their beliefs to sustain a distorted cognition. In this multiperiod setting where the opinions are expressed twice, it is perfectly possible to see a large share of the population (here, the α biased types) persist in their wrong belief accross time, and rationalize the arising distribution of opinions by attributing them to *biased* agents.

This result provides an explanation for why some collective beliefs quickly appear or disappear in a society. An interesting example is provided by Bronner (2004) who studies why the children stop believing in Santa Klaus. He reports that, since the abandon of the belief is very costly (since it might imply that no gift will be received in the future), a large fraction of the children experience an internal struggle during some period in order to maintain their conviction in face of contradictory evidence ; and that, generally, they give up when a sufficiently large number of other children (especially the oldest) start to spread an alternative theory (the presents come from their parents), which make it more difficult to sustain the cognitive dissonance. Similarly, when more people start to believe that the creationism is a sound theory, or that radioactivity will ultimately endanger human health (cf Boudia (2007)), it becomes easier for those who 'would like' to think the same to engage in self deception and to rationalize their own cognition. It is noteworthy that this result here relies on a rationalization process that occurs inside a social group, and remains distinct from arguments related

[5]Note that this strategic complementarity is endogeneous in this model, and does not rely on assumptions made about some cultural affiliation, where each individual would care only (or more) about having the opinion prescribed by some group membership.

to social pressure or private signals.

5.2 Personal interactions

This paragraph studies the influence of social interactions on the possibilities of sustaining one's distorted cognition. Instead of assuming that the agents fully observe the distribution of opinions, I suppose that they interact with M persons (family, friends, ...) before date 2. Two types of interaction will be considered : *weak* relationships, where the agents share only their preferred decision a, and *strong* interactions where they also are able to share some information if they want to. The former can be thought of as loose discussions involving few sophisticated arguments, while the latter imply a deeper conversation. For expositional simplicity, I present the case of a type A-equilibrium (the general case is equivalent).

Weak interactions This interaction is modelled in the simplest way : each agent v interacts with M persons whose values $(w_1, ..., w_M)$ are i.i.d. and drawn according to the density g. I assume, first, that the agent does not know precisely the value of his relationships : his prior information is limited to g. The information received at this stage takes the form of a sequence of opinions $(a_1, ..., a_M)$ or, equivalently, of an integer $k \in \{0, ..., M\}$ representing the number of persons expressing a preference for $a = 1$ in his poll ($M - k$ agents prefer decision $a = -1$). Weak interactions are more likely to be relevant for situations where the facts are difficult to communicate, for instance where information is *soft* rather than *hard*.

Consider the case of an agent v with biased cognition : assume for instance that $v > v_4$, which means that the agent forgets the messages $m = -x$ with probability 1. How does the shape of the agent's social network influence his opinion at date 2 (seen in expectation from date 1) ?

Conditional on the value of m, the original signal, the random draws of $(a_1, ..., a_M)$ follow a Bernouilli process with respective probabilities :

$$\mathbb{P}(a_j = 1 | m = \varnothing) = \int_0^{+\infty} g(v)dv$$
$$\mathbb{P}(a_j = -1 | m = \varnothing) = \int_{-\infty}^{0} g(v)dv$$

$$\mathbb{P}(a_j = 1 | m = -x) = \int_{(2\pi-1)x}^{+\infty} g(v)dv$$
$$\mathbb{P}(a_j = -1 | m = -x) = \int_{-\infty}^{(2\pi-1)x} g(v)dv$$

The information extracted from the sequence $(a_1, ..., a_M)$ about m is the most significant when $\mathbb{P}(a_j = 1|m = \varnothing)$ and $\mathbb{P}(a_j = 1|m = -x)$ are the furthest (see the appendix). Indeed, the intuition is that a *right-biased* type needs to meet many *moderately biased* right types who vote for $a = -1$ in order to realize that some information $m = -x$ had been sent. Meeting *left-biased* types or *extremely biased* right types is uninformative since their opinion is the same in both states $m = -x$ and $m = \varnothing$.

Proposition 5. *Fix $\int_{(2\pi-1)x}^{+\infty} g(v)dv$. At date $t = 2$, the opinion of a right-biased agent ($v > v_4$) towards X is more accurate, the higher is $\int_0^{(2\pi-1)x} g(v)dv$.*

For instance, an *extremely* biased agent who would meet only *extremely* biased agents like him (i.e. such that $\int_0^{(2\pi-1)x} g(v)dv = 0$) would not update his opinion at all after the social interactions.

Consider now the fact where an agent receives the opinion of only one source (official institution, firm, public expert...), whose value v is publicly known. If $v = 0$, the messenger is entirely trustworthy, since his opinion is based only on the existing evidence, and not on any idiosyncratic distortion. Another case of interest is when $v \neq 0$, but the messenger conveys a contradictory opinion, for instance $v > 0$ and $a = -1$. In this case again (what Glaeser and Sunstein (2013) call the *convert communicator*), the messenger proves that his opinion his reliable by revealing some signal that contradicts his preexisting wordlview. For instance, Glaeser and Sunstein (2013) report that in an experimental study, individuals tend to adopt more easily a position that hurts their values (for instance, Democrats tend to support conservative policies) when this opinion is held by the representatives of their affiliation political party.

Strong interactions To model stronger interactions between the individuals, I assume that they now have the choice, when they meet, to transfer their encoded memory $\hat{m}$ at no cost. They do so in order to convince those close to them to come round to their opinion. With this assumption, the *extreme* types selectively convey their information, while the moderate types always share all their information (if any). The result of the social interactions for an agent $v > v_4$ who recalls $\hat{m} = \varnothing$ is a sequence of signals $(m_1, ..., m_M)$, where $m_j \in \{-x, \varnothing\}$.

If the original message was $m = \varnothing$, no information is extracted from the interactions. However, if the message was $m = -x$, the probability for agent w_j to transfer m_j equals

$$\mathbb{P}(m_j = -x|m = -x) = \int_{-\infty}^{(2\pi-1)x} g(v)dv \tag{5}$$

The signals m_j are themselves subject to memory manipulation, as was the original message m : I assume that the agent can choose to deny the signals $m_j = -x$ when received,

at a cost c for each. When too many messages $m_j = -x$ are received, the cost of cognitive dissonance becomes too high and the agent chooses to keep the signal m and update his belief by coming round to the opinion $m = -x$. I assume that all the messages are received simultaneously, and that the agent chooses whether to deny all the messages such that $m_j = -x$ (at a cost $c \times |\{j, m_j = -x\}|$) or to recall the evidence. There exists a cutoff level $k \in \{1, ..., M\}$ such that the optimal decision for agent v is to encode the information $m = -x$ if and only if $m_j = -x$ is received at least k times. This, together with equation 5 proves :

Proposition 6. *The expected opinion of a right-biased agent ($v > v_4$) towards X at date $t = 2$ is more accurate, the higher is $\int_{-\infty}^{(2\pi-1)x} g(v)dv$.*

Propositions 5 and 6 show how the composition of the social network influences the possibility of persisting in one's biased opinion. An important aspect is that some updating might occur even though all the original information is public : the social interactions play some role not by spreading the information, but rather by limiting the possibility for some groups to engage in wishful thinking.

A first observation is that a biased agent ($v > v_4$) who would meet only biased agents like him would not update his belief at all. Hence, a society composed of independent groups whose members share the same values is very favorable to the persistence of ideological beliefs across time ; conversely, social mixity (in terms of values) is the most efficient device to fight ideological thinking. This finding is consistent with the evidence (provided by Erisen and Erisen (2012)) that a close-knit cohesive social network is detrimental to the sophistication of one's political thinking. A second observation is that the *strong* social interactions, where more information is shared among the agents, are more efficient than *weak* interactions to prevent wishful thinking. Erisen and Erisen (2012) also report that the individuals who have regular and deep political conversations with their peers are more likely to form coherent reasonings about policy in general. According to the present model, this happens because the objective information transmitted in these conversations helps to overcome the ideological biases, while it is easy to attribute some adverse opinion to other non-meaningful factors.

Endogeneous network formation An immediate implication of the model is that the individuals prefer to interact with people who share the same values, since it helps them to maintain their motivated beliefs and to avoid the revision of their opinion. If the social interactions described above are endogeneously chosen by the agents, a prediction of the model is that the individuals would try to form coherent groups of people endowed with the same v, endogeneously leading to the dispersion of beliefs in the society.

6 Personal identity, individual cognition and ex post rationalization

The previous section showed that wrong collective beliefs can be sustained in equilibrium in spite of the possibilities of updating offered by social interactions. This part explores another mechanism of manipulation of information based on the fact that individuals might want themselves to distort the ex post feedback on their actions (or opinions) in order to maintain a positive self-view. It provides another rationale for why beliefs towards objective phenomena tend to align with cultural preferences. I consider a situation where some binary decision is driven both by uncertain preferences (an imperfectly-recalled identity) and by some imperfect information. The main idea of this part is that the agents might want to suppress some information that, according to one's past behavior, casts some doubt about the agent's intrinsic characteristics.

6.1 Interpretation of the model

Individual decision In this framework, the choice $a = 1$ or $a = -1$ refers to an individual decision : for instance, smoking, mountaineering or consuming GMOs. The payoff to the agent is composed of two terms. The first is the *idiosyncratic* payoff, equal to $v.a$, where v denotes the personal benefit derived from making the decision $a = 1$. In the following, v will be referred to as the personal *identity*. The second term is the *specific* payoff, equal to $X.a$, where the random variable X can take values $x > 0$ or $x < 0$ with equal probabilities. All together, the payoff to a v-type individual deciding a equals $\pi(v, X, a) = Xa + va$.

Information structure Contrary to the previous section, v is derived from a discrete prior distribution. I assume that v can take only three values : $v \in V = \{v_L, v_M, v_H\}$. $v = v_H$ (high taste), $v = v_M$ (intermediate taste) and $v = v_L$ (low taste) with respective probabilities α, $(1 - \alpha - \beta)$ and β. I assume that :

$$v_L < -x < v_M < x < v_H$$

In other words, the extreme types'decisions are insensitive to the information received about X, since they have a dominant strategy : the high type always benefits from deciding $a = 1$, while the low type always chooses $a = -1$. In contrary, it is optimal for the intermediate type to choose $a = 1$ if and only if $X = x$.

Timing The timing of the game is identical to the baseline model, except that the agent receives some feedback over the relevance of the decision, and has to decide how to treat this information.

At $t = 0$, the identity v is fully known by the agent, who also receives a signal $m \in \{-x, x\}$ which is right ($m = X$) with probability $\pi > \frac{1}{2}$. π is interpreted here as the ex-ante probability of being the informed type : with probability π, the individual is skilled and receives the right signal. The decision a is made according to the information (v, m).

At $t = 1$, the agent makes a decision a according to the information (v, m).

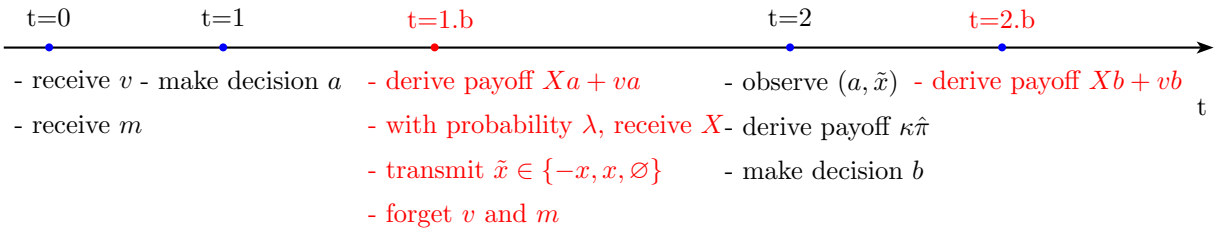
At $t = 1.b$, the agent receives with probability $0 < \lambda < 1$ some information about the payoff-relevant variable X. To simplify, I assume that the information takes the form of a perfectly informative signal equal to the realization of X. The signal $\tilde{x} \in \{-x, x, \varnothing\}$ transmitted to the next period is submitted to the cognitive manipulations described in section 3. Hence, a cognitive strategy is a pair of probabilities (σ_-, σ_+) denoting the probabilities of recalling the feedbacks $X = -x$ and $X = x$ respectively. With probability $1 - \lambda$, the agent does not receive any signal and consequently has no decision to make : $\tilde{x} = \varnothing$.

At $t = 2$, the agent has forgotten with probability 1 her identity v and the information m that directed the period 1-decision. The only observable variables are the period 0-decision a and the signal $\tilde{x}$ received from period 1[6]. The agent then draws inferences $\hat{\pi}(a, \tilde{x})$ (the probability of being the well-informed type), $\hat{v}(a, \tilde{x})$ (the probability of being the type v) and $\hat{X}(a, \tilde{x})$ (the probabilities attached to the values of X).

I also assume that there exists a positive instrumental value of correctly transmitting the feedback information. This is the case, for instance, for a repeated decision of whether to engage in a risky activity (e.g. smoking, mountaineering, consuming GMOs) for which some more precise information is provided to the agents after a past decision has already been made. More precisely, at date $t = 2$, the agent makes a second decision b leading to the payoff $Xb + vb$.

The (modified) timing of the game is now as follows :

[6]This rather extreme assumption is made in order to capture the fact that the agents' self-assessment process is based more on their past actions, which are easily remembered, than on their past preferences or information. The results are not modified if the agents remember their past preference with a probability lower than 1. This assumption is typical from the literature on identity-investment, where the players usually take actions in order to signal their type to their future self, for instance in Bénabou and Tirole (2011). Here, the signalling motivation remains the same, but the channel through which the signals are sent is the manipulation of one's information towards past behavior.



Period 1-decision In order to put the emphasis on the manipulation of beliefs, I consider the case where the discount factor between periods $1.b$ and 2 is sufficiently large for the period 1-choice to be unaffected by the following continuation game. More precisely, I suppose that :

Assumption 2. *At date 1, the agents play in their best short-term interest without anticipating the consequences.*

Moreover, I suppose that the prior of the player about the quality of the period 0-information, π, is high enough to induce a type v_M to always follow this recommendation :

Assumption 3. $|v_M| < (2\pi - 1)x$

Assumptions 2 and 3 together imply that :

$$\begin{cases} a(v_H, m) = 1 \text{ for all } m \\ a(v_M, m) = 1 \text{ if and only if } m = x \\ a(v_L, m) = -1 \text{ for all } m \end{cases} \tag{6}$$

Self-esteem concern At $t = 1.b$, when choosing whether to transmit the feedback X to the next period, the agent trades off two incentives. The first is the instrumental value of information in order to make the proper decision at $t = 2$. The second is related to a self-esteem concern. The key assumption in this framework is that the agents only remember a and $\tilde{x}$ when they evaluate themselves. Suppose for instance that self 2 remembers the information $\tilde{x} = -x$ and the decision $a = 1$. There is some doubt about whether the choice $a = 1$ was made by a misinformed intermediate type (type v_M having received the wrong signal $m = x$) or by a high type ($v = v_H$) insensitive to the information received. Hence, this memory profile entails a positive self-esteem cost. The balance of these two contradicting incentives provides the basic mechanism by which beliefs manipulation may occur.

6.2 Equilibrium

An equilibrium of the game is characterized by a set of first-period actions $a(v, m)$, a set of recall strategies $(\sigma_-(a, v), \sigma_+(a, v))$, a set of posterior beliefs $\hat{\pi}(a, \tilde{x})$, $\hat{v}(a, \tilde{x})$ and $\hat{X}(a, \tilde{x})$ and a set of last-period actions $b(a, \tilde{x})$. The standard equilibrium concept is that of Perfect Bayesian Equilibrium, which imposes that beliefs are derived by Bayes rule whenever possible, and that the strategies of all players are optimal given the set of beliefs and the other players' strategies. More precisely :

- $a(v, m)$ satisfies (6)
- $(\sigma_-(a, v), \sigma_+(a, v))$ and $b(a, \tilde{x})$ maximize the agent's expected utility according to the information held by the agent and the equilibrium strategies
- whenever possible, the beliefs are derived by Bayes rule.

In order to further simplify the analysis and to focus on the most intuitive equilibria, I use the following restrictions. First, I focus on the most informative equilibria. If two equilibra feature the same actions, and are such that all the players transmit more information in the second one, and at least one player transmits strictly more information, then the second one is selected and the first is ruled out.

Finally, in order to rule out equilibra where some agents play a dominated strategy, the equilibrium must also satisfy the extensive form trembling hand property, i.e. it must be the limit of a sequence of equilibria of perturbed games converging to the game studied (i.e., games where the players are allowed to play only totally mixed actions at every information set).

6.3 Self-esteem and ex-post rationalization

In this paragraph, I study the effect of the self-esteem concern κ defined above on the incentives to recall or deny the objective information. Without loss of generality, I assume that $a = 1$. Thus, when self 2 receives $a = 1$ and $\tilde{x} \in \{-x, x, \varnothing\}$, the probability attached to being the low type v_L is equal to zero. In this section, I drop the subscript a.

To sharpen the intuition, I begin by showing that full revelation of the signal (i.e $\tilde{x} = X$ with probability one when a signal X is received) is not an equilibrium of the game.

Lemma 7. *Full revelation of the information is not an equilibrium.*

Lemma 1 states that some feedback information must be lost due to the presence of contradictory incentives : providing the next self with accurate information about the state of

the world, and at the same time enhancing one's self esteem. The driving force of this result if the presence of a strict incentive to announce $X = x$ (the "good news" about the relevance of the decision) and a strict incentive to conceal $X = -x$ (the "bad news"). Indeed, in full revelation the ex post probabilities of being the informed type are given by :

$$\hat{\pi}(\tilde{x} = -x) = \pi \frac{\alpha}{\alpha + (1-\pi)(1-\alpha-\beta)} < \pi$$
$$\hat{\pi}(\varnothing) = \pi$$
$$\hat{\pi}(\tilde{x} = x) = \pi \frac{\alpha + (1-\alpha-\beta)}{\alpha + \pi(1-\alpha-\beta)} > \pi$$

The continuation game starting at date 1 after action $a = 1$ admits a multiplicity of equilibria. In order to illustrate the role played by the self-esteem term, however, I shall focus here on the special case where κ is large. In this case, there exists a unique equilibrium of the game, which is described in the next proposition.

Proposition 8. *Suppose that the first-period action is $a = 1$. There exists a threshold κ^* such that, if $\kappa > \kappa^*$, all the equilibria of the game are such that :*

(a) the "good news" ($X = x$) are always revealed by both players, i.e. $\sigma_+(v_M) = \sigma_+(v_H) = 1$.

(b) the high type reveals more "bad news" than the intermediate type, i.e. $\sigma_-(v_M) \leq \sigma_-(v_H)$.

The case $a = 0$ is the mirror case of $a = 1$. The first part of the proposition states that the feedbacks conveying good news about the relevance of the past decision are always revealed in equilibrium. The second part states that some information about the bad news must be destroyed. Suppose, indeed, that v_M and v_H both truthfully report the signal $X = -x$. After observing the signal $\tilde{x} = -x$ and the action $a = 1$, self 2 incurs a loss in self-esteem since these observations make it more plausible that self 1 has been misled by a wrong signal. Consequently, it cannot be optimal for v_H and v_M to reveal $X = -x$ with probability 1. Revealing $X = x$, on the contrary, enhances one's self-views, since it was optimal for both types to play $a = 1$ in this state.

Appendix B provides a complete characterization of the equilibria of the game.

The findings are consistent with many real-life observations, for instance the empirical literature on individual risk perceptions : people tend to bias their cognition in order to rationalize their past behavior, i.e. to think that there decisions relied on objective evidence (see for instance Slovic (2000) for several examples, Loewenstein (1999) for mountaineering or Peretti-Watel (2001) for cigarette and drug consumption). According to the present model,

this need for rationalization, which is costly for future decision-making, occurs because of the imperfect transmission of tastes accross times.

6.4 Identity transmission

This section studies a slightly different version of the previous model, in a case where individuals want to signal their identity v rather than their skills. For instance, some people engaged in risky activities (extreme sports, working in a dangerous environment...) might feel the need to show to their future incarnation (and to their environment) that their taste for this activity is high. From self 1's view point, the payoff derived at date $t = 1.b$ by a v-type agent is now equal to $\mu\mathbb{P}(V = v|a, \tilde{x})$, where $\mathbb{P}(V = v|a, \tilde{x})$ is the ex post probability attached by the agent to the true identity v after seeing history $(a, \tilde{x})$.

The following proposition shows that the incentive to transmit one's identity modifies the pattern of information denial, for an individual whose prior is sufficiently biased towards the dominant type.

Proposition 9. *If α is sufficiently close to 1, there exists a unique Pareto-dominant equilibrium of the game where $\sigma_-(v_M) = \sigma_+(v_M) = 1$ and $\sigma_-(v_H) = \sigma_+(v_H) = 0$.*

The underlying mechanism is that the identity transmission incentivizes the different types to completely separate. The proposition shows that the player with a dominant strategy denies all the feedbacks. The intuition for this result is that self 2, after receiving the signal $\tilde{x} = \varnothing$, attaches a probability close to 1 to the type v_H, and acts accordingly : $b(\varnothing) = 1$. Thus, the intermediate type v_M has more incentives to announce the signals (in order to induce $b(-x) = 0$ if $X = -x$, and in order to signal the identity v_M if $X = x$) than the high type.

The surprising result is that the high type denies even the good news about the relevance of the past decision. This finding is consistent with some facts reported by the literature on individual risky behaviors. For instance, Peretti-Watel (2001) reports that, in a survey, some workers in a chemical factory disregarded the information stating that their job was "safe". In this model, the rationale for doing so is to transmit to their future incarnation a signal of how much they value their activity.

7 Conclusion

In this paper, I introduced two notions mainly used by sociologists and social psychologists in economic models of decision-making in a situation of uncertainty. The main insight is to show how self-serving motives lead the individuals to manipulate their own beliefs in order to

protect their *identity* or their *values* ; and how the distortion can persist across time in spite of the possibilites of updating in the future.

The main message of this work is that the interest of providing the citizens with objective information is considerably weakened when the issue involves both scientific knowledge and personal or cultural worldviews. For instance, in the context of risk regulation, *identity* concerns prevent the individuals from appropriately assessing the risks they face, while the dichotomy of *values* translates into a similar polarization of judgements over the scientific evidence. This does not mean that transparent and reliable information is not desirable, but it shows, at least, that it is insufficient. It also opens new challenges for effective risk regulation, such as the possibility to adapt risk communication to various social groups, or the relevance of including such aspects in regulatory decisions.

References

Akerlof, George and Rachel Kranton (2000), "Economics and identity." *Quarterly Journal of Economics*, 115, 715–753.

Anderegg, William, James Prall, Jacob Harold, and Stephen Schneider (2010), "Expert credibility in climate change." *Proceedings of the National Academy of Sciences*, 107, 12107–12109.

Austen-Smith, David and Roland Fryer (2005), "An economic analysis of "acting white"." *Quarterly Journal of Economics*, 120, 551–583.

Bénabou, Roland (2008), "Ideology." NBER Working Paper number 13907.

Bénabou, Roland (2009), "Groupthink: Collective delusions in organizations and markets." NBER Working Papers 14764, National Bureau of Economic Research, Inc.

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

Bénabou, Roland and Jean Tirole (2006), "Belief in a just world and redistributive politics." *Quarterly Journal of Economics*, 121, 699–746.

Bénabou, Roland and Jean Tirole (2011), "Identity, morals, and taboos: Beliefs as assets." *Quarterly Journal of Economics*, 126, 805–855.

Benjamin, Daniel J., James J. Choi, and A. Joshua Strickland (2010), "Social identity and preferences." *American Economic Review*, 100, 1913–28.

Borraz, Olivier (2008), *Les Politiques du Risque*.

Boudia, Soraya (2007), "Naissance, extinction et rebonds dune controverse scientifique : les dangers de la radioactivité pendant la guerre froide." *Revue d'histoire intellectuelle*, 25, 157–170.

Breyer, Stephen (1995), *Breaking the Vicious Circle: Toward Effective Risk Regulation*.

Bronner, Gérald (2004), "Contribution á une théorie de l'abandon des croyances : la fin du pére noël." *Cahiers internationaux de sociologie*, 116, 117–140.

Brunnermeier, Markus, Christian Gollier, and Jonathan Parker (2007), "Optimal beliefs, asset prices, and the preference for skewed returns." *American Economic Review*, 97, 159–165.

Brunnermeier, Markus and Jonathan Parker (2005), "Optimal expectations." *American Economic Review*, 95, 1092–1118.

Carrillo, Juan and Thomas Mariotti (2000), "Strategic ignorance as a self-disciplining device." *Review of Economic Studies*, 67, 529–44.

Ditto, Peter, James Scepansky, Geoffrey Munro, Anne Marie Apanovitch, and Lisa Lockhart (1998), "Motivated sensitivity to preference-inconsistent information." *Journal of Personality and Social Psychology*, 75, 53–69.

Dixit, Avinash and Jörgen Weibull (2007), "Political polarization." *Proceedings of the National Academy of Sciences*, 104, 7351–7356.

Douglas, Mary (1994), *Risk and Blame : Essays in Cultural Theory*.

Douglas, Mary and Aaron Wildavsky (1982), *Risk and Culture: An Essay on the Selection of Technological and Environmental Dangers*.

Erisen, Elif and Cengiz Erisen (2012), "The effect of social networks on the quality of political thinking." *Political Psychology*, 33, 839–865.

Farnsworth, Stephen and Robert Lichter (2012), "The structure of scientific opinion on climate change." *International Journal of Public Opinion Research*, 24, 93–103.

Festinger, Leon (1957), *A Theory of Cognitive Dissonance*.

Glaeser, Edward and Cass Sunstein (2013), "Why does balanced news produce unbalanced views?" Working paper.

Gollier, Christian (2007), "Optimal expectations with complete markets." IDEI Working Papers 462, Institut d'Economie Industrielle (IDEI).

Grossman, Gene and Elhanan Helpman (1996), "Electoral competition and special interest politics." *Review of Economic Studies*, 63, 265–86.

Hardisty, David, Eric Johnson, and Elke Weber (2010), "A dirty word or a dirty world? attribute framing, political affiliation and query theory." *Psychological Science*, 21, 86–92.

Kahan, Dan M. (2012), "Cultural cognition as a conception of the cultural theory of risk." In *Handbook of Risk Theory: Epistemology, Decision Theory, Ethics and Social Implications of Risk* (Sabine Roeser, Rafaela Hillerbrand, Per Sandin, and Martin Peterson, eds.), Springer.

Kahan, Dan M., Donald Braman, Geoffrey L. Cohen, John Gastil, and Paul Slovic (2010), "Who fears the hpv vaccine, who doesnt, and why? an experimental study of the mechanisms of cultural cognition." *Law and Human Behavior*, 34, 501–16.

Kahan, Dan M., Ellen Peters, Maggie Wittlin, Paul Slovic, Lisa Larrimmore Ouellette, Donald Braman, and Gregory Mandel (2012), "The polarizing impact of science literacy and numeracy on perceived climate change risks." *Nature Climate Change*. Advanced online publication.

Levy, Raphaël (2012), "Soothing politics." Working paper.

Loewenstein, Goerge (1999), "Because it is there: the challenge of moutaineering... for utility theory." *Kyklos*, 52, 315–44.

Mullainathan, Sendhil and Andrei Shleifer (2005), "The market for news." *American Economic Review*, 95, 1031–1053.

Peretti-Watel, Patrick (2001), *La société du risque*.

Plous, Scott (1991), "Biases in the assimilation of technological breakdows: do accidents make us safer ?" *Journal of Applied Social Psychology*, 21, 1058–1082.

Pollak, Michaël (1988), *Les Homosexuels et le SIDA: Sociologie d'une Epidémie*.

Portney, Paul R. (1992), "Trouble in happyville." *Journal of Policy Analysis and Management*, 11, 131–2.

Salanié, Francois and Nicolas Treich (2009), "Regulation in happyville." *The Economic Journal*, 119, 665–670.

Slovic, Paul (2000), *The Perception of Risk*.

Tetlock, Philip E., Orie V. Kristel, S Beth Elson, Melanie C Green, and Jennifer S Lerner (2000), "The psychology of the unthinkable: Taboo trade-offs, forbidden base rates, and heretical counterfactuals." *Journal of Personality and Social Psychology*, 78, 853–70.

Viscusi, W. Kip and James T. Hamilton (1999), "Are risk regulators rational ? evidence from hazardous waste cleanup decisions." *American Economic Review*, 89, 1010–1027.

Appendix A

Proof of Proposition 1 A v-type agent recalling $\hat{m} = \varnothing$ infers from the equilibrium strategies the probabilities that the other agents had received some signal m. By Bayes' rule, the posterior probabilities attached to these states are :

$$\mathbb{P}(m = x|\hat{m} = \varnothing) = \frac{\frac{\lambda}{2}(1 - \sigma_+(v))}{1 - \lambda + \frac{\lambda}{2}(1 - \sigma_+(v)) + \frac{\lambda}{2}(1 - \sigma_-(v))} \tag{7}$$

$$\mathbb{P}(m = -x|\hat{m} = \varnothing) = \frac{\frac{\lambda}{2}(1 - \sigma_-(v))}{1 - \lambda + \frac{\lambda}{2}(1 - \sigma_+(v)) + \frac{\lambda}{2}(1 - \sigma_-(v))} \tag{8}$$

$$\mathbb{P}(m = \varnothing|\hat{m} = \varnothing) = \frac{1 - \lambda}{1 - \lambda + \frac{\lambda}{2}(1 - \sigma_+(v)) + \frac{\lambda}{2}(1 - \sigma_-(v))} \tag{9}$$

Moreover, an agent recalling signal $\hat{m}$ $in\{-x, x\}$ knows with probability 1 that $m = \hat{m}$ had been sent to all agents, since it is impossible to forge a signal. Hence, considering that the signal is of quality π, the ex post probability attached to the states X is equal to :

$$\mathbb{P}(X = x|\hat{m} = x) = \pi$$

$$\mathbb{P}(X = -x|\hat{m} = x) = 1 - \pi$$

$$\mathbb{P}(X = -x|\hat{m} = -x) = \pi$$

$$\mathbb{P}(X = x|\hat{m} = -x) = 1 - \pi$$

$$\mathbb{P}(X = x|\hat{m} = \varnothing) = \pi\mathbb{P}(m = x|\hat{m} = \varnothing) + (1 - \pi)\mathbb{P}(m = -x|\hat{m} = \varnothing) + \frac{1}{2}\mathbb{P}(m = \varnothing|\hat{m} = \varnothing)$$

$$\mathbb{P}(X = -x|\hat{m} = \varnothing) = \pi\mathbb{P}(m = -x|\hat{m} = \varnothing) + (1 - \pi)\mathbb{P}(m = x|\hat{m} = \varnothing) + \frac{1}{2}\mathbb{P}(m = \varnothing|\hat{m} = \varnothing)$$

Hence, the ex post expectations of the state of the world are given by :

$$\mathbb{E}(X|\hat{m} = x) = (2\pi - 1)x$$

$$\mathbb{E}(X|\hat{m} = -x) = -(2\pi - 1)x$$

$$\mathbb{E}(X|\hat{m} = \varnothing) = (2\pi - 1)x\frac{\lambda}{2}\frac{\sigma_-(v) - \sigma_+(v)}{1 - \lambda + \frac{\lambda}{2}(1 - \sigma_+(v)) + \frac{\lambda}{2}(1 - \sigma_-(v))}$$

Denote $p(v)$ the last expectation, which is the evaluation of the state of the world by a citizen endowed with value v, strategies $\sigma_-(v)$ and $\sigma_+(v)$ and who does not recall any signal.

Now, denote $N_1(m)$ the proportion of citizens who vote for $a = 1$ after the message m has been sent. For instance, after $m = x$, the agents who vote for $a = 1$ are those who recall the signal x and are such that $(2\pi - 1)x + v > 0$, and those who don't recall x and are such that $p(v) + v > 0$, i.e. :

$$N_1(x) = \int_{(2\pi-1)x}^{+\infty} f(v)dv + \int_{-(2\pi-1)x}^{(2\pi-1)x} (\sigma_+(v) + (1 - \sigma_+(v))\mathbb{1}_{v+p(v)>0})f(v)dv \tag{10}$$

And similarly :

$$N_1(\varnothing) = \int_{(2\pi-1)x}^{+\infty} f(v)dv + \int_{-(2\pi-1)x}^{(2\pi-1)x} \mathbb{1}_{v+p(v)>0}f(v)dv \tag{11}$$

And

$$N_1(-x) = \int_{(2\pi-1)x}^{+\infty} f(v)dv + \int_{-(2\pi-1)x}^{(2\pi-1)x} (1 - \sigma_+(v))\mathbb{1}_{v+p(v)>0}f(v)dv \tag{12}$$

Equations (10), (11) and (12) show that $N_1(x) \geq N_1(\varnothing) \geq N_1(-x)$.

Consider now the interim anticipatory utility that the agent receives when he forms his beliefs about the future. When a signal $m = -x$ has been received, if the agent recalls it he knows with probability 1 that $N_1(-x)$ people will vote for $a = 1$. Hence, his interim utility is given by :

$$\pi_{\text{interim}}(\text{recalling}) = s_1(-(2\pi - 1)x)(2\phi(N_1(-x)) - 1) + s_2 v(2\phi(N_1(-x)) - 1)$$

The interim utility from denying stems from the inference by the agent of the possible messages sent to the public, given his equilibrium memory manipulations :

$$\begin{aligned}\pi_{\text{interim}}(\text{denying}) = {} & \mathbb{P}(m = -x|\hat{m} = \varnothing)(-s_1(2\pi - 1)x + s_2 v)(2\phi(N_1(-x)) - 1) \\ & + \mathbb{P}(m = \varnothing|\hat{m} = \varnothing)s_2 v(2\phi(N_1(\varnothing)) - 1) \\ & + \mathbb{P}(m = x|\hat{m} = \varnothing)(s_1(2\pi - 1)x + s_2 v)(2\phi(N_1(x)) - 1)\end{aligned}$$

Hence, the incentive $I(v)$ to deny the signal $m = -x$ is given by :

$$\begin{aligned}I(v) &= \pi_{\text{interim}}(\text{denying}) - \pi_{\text{interim}}(\text{recalling}) \qquad (13)\\ &= \mathbb{P}(m = x|\hat{m} = \varnothing)[s_1(2\pi - 1)x(2\phi(N_1(-x)) - 1 + 2\phi(N_1(x)) - 1) + s_2 v(2\phi(N_1(x)) - 2\phi(N_1(-x)))] \\ &+ \mathbb{P}(m = \varnothing|\hat{m} = \varnothing)[s_1(2\pi - 1)x(2\phi(N_1(-x)) - 1) + s_2 v(2\phi(N_1(\varnothing)) - 2\phi(N_1(-x)))] \qquad (14)\end{aligned}$$

To interpret this equation, notice that :

- $\mathbb{P}(m = x|\hat{m} = \varnothing)$ is the ex post probability that the agent attaches to the message $m = x$ having been sent (it depends on the recall strategies σ_- and σ_+)
- $\mathbb{P}(m = \varnothing|\hat{m} = \varnothing)$ is the ex post probability that the agent attaches to the message $m = \varnothing$ having been sent
- $(2\pi - 1)x(2\phi(N_1(-x)) - 1)$ is positive if and only if $\phi(N_1(-x)) > \frac{1}{2}$, i.e. if many people in the society vote for $a = 1$ in spite of some evidence that $X = -x$. Hence, in this case this term enters positively in the incentive to deviate since it reflects a correction for wrong decision making.
- $(2\pi - 1)x(2\phi(N_1(x)) - 1)$ is positive if and only if $\phi(N_1(x)) > \frac{1}{2}$. This term reflects the anticipation of the social decision if message $m = x$ has been sent.
- The terms $v(2\phi(N_1(\varnothing)) - 2\phi(N_1(-x)))$ and $v(2\phi(N_1(x)) - 2\phi(N_1(-x)))$ are of the sign of v. If $v > 0$, there are both positive, and a decision maker endowed with $v > 0$ has a positive incentive to forget $m = -x$ in order to improve his beliefs about the future social decision.

Hence, focusing on the terms depending on v shows that the incentive to forget $m = -x$ tends to ∞ when v tends to ∞, and to $-\infty$ when v tends to $-\infty$. Hence, the recall strategy σ_- must be equal to 1 for low values of v, and to 0 for high values of v.

Claim 1. The function σ_- is non-increasing in v.

Proof. Suppose on the contrary that there exist a pair of values $v_1 < v_2$ such that $\sigma_-(v_1) < \sigma_-(v_2)$. Equations 7 and 9 show that $\mathbb{P}(m = x|\hat{m} = \varnothing)$ and $\mathbb{P}(m = \varnothing|\hat{m} = \varnothing)$ are strictly increasing functions of $\sigma_-(v)$. Consequently, equation (13) shows that $I(v_1) < I(v_2)$. But $\sigma_-(v_1) < \sigma_-(v_2)$ implies that $\sigma_-(v_1) < 1$, and hence $I(v_1) \geq c$, and also implies that $\sigma_-(v_2) > 0$ and hence $I(v_2) \leq c$, which contradicts $I(v_1) < I(v_2)$. □

Claim 2. The function σ_- is continuous in v.

Proof. Suppose that there exists v such that σ_- is discontinuous at v. Since σ_- is decreasing, there is a downward jump at v. Hence, $\sigma_-(v^+) < \sigma_-(v^-)$. Equations 7, 9 and 13 show that $I(v_-) > I(v^+)$, which contradicts $\sigma_-(v^+) < \sigma_-(v^-)$. □

Combining claims 1 and 2 show that there exist some thresholds v_3 and v_4 such that σ_- is equal to 1 on $(-\infty, v_3)$ and to 0 on $(v_4, +\infty)$, and everywhere continuous. On (v_3, v_4), the agent is indifferent between denying or recalling and

$$\forall v \in (v_3, v_4), I(v) = c$$

This proves the first part of the proposition.

Consider now assumption 1. If 1a) or 1 b) holds, it is easy to show that $I(v) < c$ for all $v \leq 0$ (see equation 13). Hence, $\sigma_-(v) = 1$ for all $v \leq 0$, and $v_3 > 0$. By symmetry, $\sigma_+(v) = 1$ for all $v \geq 0$, and the incentive to deny the news $m = -x$ for $v \geq 0$ simplifies to :

$$I(v) = \frac{1-\lambda}{1-\lambda+\frac{\lambda}{2}(1-\sigma_-(v))}[s_1(2\pi-1)x(2\phi(N_1(-x))-1) + 2s_2 v(\phi(N_1(\varnothing)) - \phi(N_1(-x)))] \tag{15}$$

This expression is constant and equal to c over (v_3, v_4). Thus, $\sigma_-(v)$ on (v_3, v_4) is uniquely determined by 15, and is affine in v. Moreover, v_4 satisfies the equations $\sigma_-(v_4) = 0$ and $I(v_4) = c$, while v_3 satisfies $\sigma_-(v_3) = 1$ and $I(v_3) = c$, which implies

$$s_1(2\pi-1)x(2\phi(N_1(-x))-1) + 2s_2 v_3(\phi(N_1(\varnothing)) - \phi(N_1(-x))) = c \tag{16}$$

An equilibrium of the game is entirely characterized by a value for v_3 (and v_2 by symmetry), which satisfies (16) (where $N_1(-x)$ and $N_1(\varnothing)$ are implicit functions of v_3) and pins down all the cognitive strategies and voting behaviors.

Claim 3. Under assumption 1, 16 has a unique solution.

Proof. I first prove the claim under the assumption $2\max(s_1, s_2)x < c$. Consider a candidate equilibrium threshold $v_3 \leq (2\pi-1)x$. Since ϕ is bounded between 0 and 1, we have the following sequence of inequalities :

$$\begin{aligned}
&I(v_3) = s_1(2\pi-1)x(2\phi(N_1(-x))-1) + 2s_2 v_3(\phi(N_1(\varnothing)) - \phi(N_1(-x))) \\
&\leq s_1(2\pi-1)x(2\phi(N_1(-x))-1) + 2s_2(2\pi-1)x(\phi(N_1(\varnothing)) - \phi(N_1(-x))) \\
&\leq (2\pi-1)x((2s_1-2s_2)\phi(N_1(-x)) + 2s_2\phi(N_1(\varnothing)) - s_1) \\
&< c
\end{aligned}$$

Hence, a solution v_3 of 16 (if it exists) is such that $v_3 > (2\pi-1)x$. Consider now the candidate values v_3 on the interval $((2\pi-1)x, +\infty)$. On this interval, $N_1(\varnothing) = \int_0^{+\infty} f(v)dv$ and $N_1(-x) = \int_{(2\pi-1)x}^{+\infty} f(v)dv$ are independent of v_3. Hence, $I(v_3)$ is affine in v_3 with positive slope and such that $I((2\pi-1)x) < c$. As a conclusion, there exists a unique solution v_3 of the fixed point problem on the interval $((2\pi-1)x, +\infty)$.

Consider now the case where $s_1 = 0$. Equation 16 becomes $2s_2v_3(\phi(N_1(\varnothing)) - \phi(N_1(-x))) = c$. For $v \in (v_3, v_4)$, $\sigma_-(v)$ is derived from the incentive constraint $I(v) = c$, which leads to $v_4 = v_3\frac{1-\frac{lambda}{2}}{1-\lambda}$ and $p(v) = -(2\pi - 1)x(1 - \frac{v_3}{v})$.

Similarly, it can be shown that $N_1(-x, v_3)$ is nonincreasing with v_3, which is intuitive : the number of agents who vote for $a = 1$ in spite of the signal $m = -x$ increases with the amount of denial, which is inversely related with v_3.

Hence, the function $s(v_3) = 2s_2v_3(\phi(N_1(\varnothing, v_3)) - \phi(N_1(-x, v_3)))$ is strictly monotonic in v_3 and equation 16 defines a unique threshold value v_3. □

In the following, I will focus on two types of equilibrium :

- An equilibrium of type A where $2\max(s_1, s_2)x < c$ and $v_3 > (2\pi - 1)x$.
- An equilibrium of type B where $s_1 = 0$ and where I make the following additional assumptions : $v_3 < v_4 < (2\pi - 1)x$ and $v + p(v) > 0$ for all $v > 0$. This is for instance equivalent to a condition on λ being high enough, and ensures that the votes are monotonic with respect to the value parameter. In this equilibrium, $N_1(\varnothing) = \int_0^{+\infty} f(v)dv$ and $N_1(-x) = \frac{2(1-\lambda)}{\lambda}\int_{v_3}^{v_3\frac{1-frac\lambda 2}{1-\lambda}}(\frac{v}{v_3} - 1)f(v)dv + \int_{v_3\frac{1-frac\lambda 2}{1-\lambda}}^{+\infty} f(v)dv$.

Comparative statics In an A equilibrium, v_3 is simply defined by the explicit equation

$$(2\pi - 1)xs_1(2\phi(\int_{(2\pi-1)x}^{+\infty} f(v)dv) - 1) + 2v_3s_2(\phi(\int_0^{+\infty} f(v)dv - \phi(\int_{(2\pi-1)x}^{+\infty} f(v)dv) = c \tag{17}$$

The derivative of the left-hand side with respect to π equals :

$$2x(2\phi(\int_{(2\pi-1)x}^{+\infty} f(v)dv) - 1) + 4x(v_3 - (2\pi - 1)x)f((2\pi - 1)x)\phi'(\int_{(2\pi-1)x}^{+\infty} f(v)dv) \tag{18}$$

The second term of 18 is positive. If the first term is also positive, namely if $\phi(\int_{(2\pi-1)x}^{+\infty} f(v)dv) > \frac{1}{2}$, then the left hand side of 17 is strictly increasing with π. In this case, a marginal increase in π, the quality of the signal, leads to a decrease in v_3, which means a deterioration of the pattern of social cognition.

In a B equilibrium, v_3 is independent of π.

Proof of Proposition 4 I derive first the conditions under which $\boldsymbol{\sigma_-(v_H) = 1}$ is sustained in equilibrium. The interim payoff to a v_H-type agent playing $\sigma_- = 1$ and recalling $\hat{m} = -x$ equals

$$\begin{aligned}\pi_{interim}(\hat{m} = -x) &= s(-x + v_H)(2\phi(0) - 1) \\ &= s(x - v_H)\end{aligned} \tag{19}$$

And the *ex post* cognitive dissonance cost equals *zero*, since a v_H-type agent who voted for $a = -1$ infers that he recalled the evidence $m = -x$.

Now, the interim anticipatory utility to a v_H-type agent who deviates an play $\sigma_- = 0$ is the sum of the anticipation of the social decision following messages $m = -x$, $m = \varnothing$ and $m = x$ weighted by the

corresponding *posterior* probabilitities attached to these states :

$$\pi_{interim}(\hat{m}=\varnothing) = s \underbrace{\frac{1-\lambda}{1-\frac{\lambda}{2}}}_{\mathbb{P}(m=\varnothing|\hat{m}=\varnothing)} \underbrace{v_H[2q\phi(\alpha+\frac{\gamma}{2})+2(1-q)\phi(\alpha)-1]}_{\pi_{interim}(m=\varnothing)} \tag{20}$$
$$+ s \underbrace{\frac{\frac{\lambda}{2}}{1-\frac{\lambda}{2}}}_{\mathbb{P}(m=-x|\hat{m}=\varnothing)} \underbrace{(-x+v_H)(2\phi(0)-1)}_{\pi_{interim}(m=x)}$$

And the *ex post* cognitive dissonance cost equals c, since a v_H type-agent who voted for $a=1$ infers that he deliberately denied the signal $m=-x$.

Hence, comparing equations 19 and 20 shows (after some algebra) that $\sigma_-(v_H)=1$ is sustainable in equilibrium if and only if

$$\frac{1-\frac{\lambda}{2}}{1-\lambda}c > 2sv_Hq\phi(\alpha+\frac{\gamma}{2}) - sx \tag{21}$$

Now, I describe the conditions under which $\boldsymbol{\sigma_-(v_H)=0}$ is an equilibrium strategy. The outcome of the vote, in this equilibrium, is that $1-\alpha$ agents vote for $a=-1$. Hence, the cost for a v_H-type agent of expressing the opinion $a=1$ equals $c\frac{\frac{\lambda}{2}}{(1-q)(1-\lambda)+\frac{\lambda}{2}}$, since $\frac{\frac{\lambda}{2}}{(1-q)(1-\lambda)+\frac{\lambda}{2}}$ is the posterior probability attached to the message $m=-x$ after seeing $N_1=\alpha$. The anticipatory payoff after playing strategy $\sigma_-=0$ equals

$$\pi_{interim}(\hat{m}=\varnothing) = s \underbrace{\frac{1-\lambda}{1-\frac{\lambda}{2}}}_{\mathbb{P}(m=\varnothing|\hat{m}=\varnothing)} \underbrace{v_H[2q\phi(\alpha+\frac{\gamma}{2})+2(1-q)\phi(\alpha)-1]}_{\pi_{interim}(m=\varnothing)} \tag{22}$$
$$+ s \underbrace{\frac{\frac{\lambda}{2}}{1-\frac{\lambda}{2}}}_{\mathbb{P}(m=-x|\hat{m}=\varnothing)} \underbrace{(-x+v_H)(2\phi(\alpha)-1)}_{\pi_{interim}(m=x)}$$

Similarly, the payoff received by a v_H-type agent who would choose to deviate and play the strategy $\sigma_-=1$ is

$$\pi_{interim}(\hat{m}=-x) = s(-x+v_H)(2\phi(\alpha)-1) \tag{23}$$

The condition for $\sigma_-(v_H)=0$ to be an equilibrium strategy is derived from 22 and 23 :

$$\frac{1-\frac{\lambda}{2}}{1-\lambda}\frac{\frac{\lambda}{2}}{(1-q)(1-\lambda)+\frac{\lambda}{2}}c < 2sv_Hq\phi(\alpha+\frac{\gamma}{2}) - sx \tag{24}$$

Comparing equations 21 and 24 proves the proposition.

Proof of Proposition 5 Define $\alpha_{-x}=\int_{(2\pi-1)x}^{+\infty} g(v)dv$ the probability for each draw a_j to equal 1 if the signal received is $m=-x$. Similarly, $\alpha_\varnothing=\int_{(2\pi-1)x}^{+\infty} g(v)dv$ is the same probability conditionned on the signal $m=\varnothing$. Among the M draws, the agent observes K opinions $a_j=1$ with probability $C_M^K\alpha_m^K(1-\alpha_m)^{M-K}$, where α_m is conditionned on the true value of m. Hence, if K opinions $a_j=1$ are observed, Bayes rule leads to

$$\mathrm{P_v}(m=-x|m=\varnothing, k\text{opinions } a_j=1) = \frac{\frac{\lambda}{2}\alpha_{-x}^K(1-\alpha_{-x})^{M-K}}{\frac{\lambda}{2}\alpha_{-x}^K(1-\alpha_{-x})^{M-K}+(1-\lambda)\alpha_\varnothing(1-\alpha_\varnothing)^{M-K}}$$

If the true value of the signal is $m=-x$, K positive draws are observed with probability $C_K^M\alpha_{-x}^K(1-\alpha_{-x})^{M-K}$. Hence, in expectation, the posterior probability attached to the signal $m=-x$ by a right-biased

agent is

$$\mathbb{E}(\mathrm{P_v}(m=-x|m=\varnothing)|m=-x)=\sum_{K=0}^{M}C_M^K\frac{\frac{\lambda}{2}\alpha_{-x}^{2K}(1-\alpha_{-x})^{2(M-K)}}{\frac{\lambda}{2}\alpha_{-x}^{K}(1-\alpha_{-x})^{M-K}+(1-\lambda)\alpha_{\varnothing}(1-\alpha_{\varnothing})^{M-K}} \tag{25}$$

Differentiating 25 with respect to $\alpha_{\varnothing}$ leads to

$$\frac{\partial\mathbb{E}(\mathrm{P_v}(m=-x|m=\varnothing)|m=-x)}{\partial\alpha_{\varnothing}}=\sum_{K=0}^{M}C_M^K\frac{\frac{\lambda}{2}(1-\lambda)\alpha_{-x}^{2K}(1-\alpha_{-x})^{2(M-K)}(M\alpha_{\varnothing}-K)\alpha_{\varnothing}^{K-1}(1-\alpha_{\varnothing})^{M-K-1}}{(\frac{\lambda}{2}\alpha_{-x}^{K}(1-\alpha_{-x})^{N-K}+(1-\lambda)\alpha_{\varnothing}(1-\alpha_{\varnothing})^{M-K})^2} \tag{26}$$

The sequence of denominators of 26 is positive and bounded from below by a constant γ. Hence, we have

$$\frac{\partial\mathbb{E}(\mathrm{P_v}(m=-x|m=\varnothing)|m=-x)}{\partial\alpha_{\varnothing}}\geq\frac{1}{\gamma}\sum_{K=0}^{M}C_M^K\frac{\lambda}{2}(1-\lambda)\alpha_{-x}^{2K}(1-\alpha_{-x})^{2(M-K)}(M\alpha_{\varnothing}-K)\alpha_{\varnothing}^{K-1}(1-\alpha_{\varnothing})^{M-K-1} \tag{27}$$

Using Newton's formula shows that

$$\sum_{K=0}^{M}C_M^K\frac{\lambda}{2}(1-\lambda)\alpha_{-x}^{2K}(1-\alpha_{-x})^{2(M-K)}(M\alpha_{\varnothing}-K)\alpha_{\varnothing}^{K-1}(1-\alpha_{\varnothing})^{M-K-1}=$$
$$\frac{\lambda}{2}(1-\lambda)M[\alpha_{-x}^2\alpha_{\varnothing}+(1-\alpha_{-x})^2(1-\alpha_{\varnothing})]^{M-1}(\alpha_{-x}^2+(1-\alpha_{-x})^2)$$

And hence, $\frac{\partial\mathbb{E}(\mathrm{P_v}(m=-x|m=\varnothing)|m=-x)}{\partial\alpha_{\varnothing}}\geq 0$.

Q.E.D.

Appendix B

In the following, for convenience I consider that the action taken in the first period was $a = 1$ and I drop the subscript a : all the probabilities are now conditioned on seeing the action $a = 1$. I denote $\hat{\pi}(\tilde{x})$ the ex-post probability of being the informed type after recalling action $a = 1$ and feedback $\tilde{x} \in \{-x, x, \varnothing\}$. Similarly, $b(\tilde{x})$ is the action taken at date 2 following history $(a = 1, \tilde{x})$. Finally, $(\sigma_-(v), \sigma_+(v))$ are player v's equilibrium strategies.

The following preliminary results will be useful to prove the proposition.

Claim 4. $b(x) = 1$

Proof. Since $a = 1$, self 2 attaches 0 probability of being the low type $v = v_L$. Thus, since $v = v_M$ or $v = v_H$, $\mathbb{E}(v + X|X = x) = \mathbb{E}(v|X = x) + x > v_M + x > 0$, where the last inequality follows from assumption 2. □

Lemma 10. *1.*

$$\hat{\pi}(x) \geq \pi$$

2.

$$\hat{\pi}(-x) \leq \pi$$

3.

$$\hat{\pi}(\varnothing) \begin{cases} < \pi & \textit{iff } \sigma_-(v_M) < \sigma_+ v_M \\ = \pi & \textit{iff } \sigma_-(v_M) = \sigma_+ v_M \\ > \pi & \textit{iff } \sigma_-(v_M) > \sigma_+ v_M \end{cases}$$

Proof. 1. Applying Bayes' rule leads to :

$$\hat{\pi}(x) = \pi \frac{\alpha\sigma_+(v_H) + (1 - \alpha - \beta)\sigma_+(v_M)}{\alpha\sigma_+(v_H) + \pi(1 - \alpha - \beta)\sigma_+(v_M)}$$

Note that this expression is well defined only if the denominator is positive, namely only if either $\sigma_+(v_H) > 0$ or $\sigma_+(v_M) > 0$. If this is so, then it is straightforward to see that $\hat{\pi}(x) \geq \pi$ (with strict inequality if $\pi < 1$ and $(1 - \alpha - \beta) > 0$). Suppose now that the signal $X = x$ is never transmitted in equilibrium, namely that $\sigma_+(v_H) = \sigma_+(v_M) = 0$, and fix an arbitrary out-of-equilibrium belief c defined by : $c = \mathbb{P}(v = v_H|\tilde{x} = x)$ when $\tilde{x} = x$ is not revealed in equilibrium. The probability of being the informed type is now given by :

$$\begin{aligned} \mathbb{P}(\text{being well informed}|\tilde{x} = x) &= \underbrace{\mathbb{P}(\text{being well informed}|\tilde{x} = x, v = v_H)}_{\pi} \underbrace{\mathbb{P}(v = v_H|\tilde{x} = x)}_{c} \\ &+ \underbrace{\mathbb{P}(\text{being well informed}|\tilde{x} = x, v = v_M)}_{1} \underbrace{\mathbb{P}(v = v_M|\tilde{x} = x)}_{1-c} \\ &= \pi c + 1 - c \\ &\geq \pi \text{ since } \pi < 1 \end{aligned}$$

This result is intuitive : in the worst case (attributing the signal $\tilde{x} = x$ to the high type, who was insensitive to the information in the first period), seeing the signal $\tilde{x} = x$ is uninformative about whether self 0 received the right signal before making a decision. Thus, in this case, the posterior equals the prior (π) and in all other cases it exceeds it.

2. Applying Bayes' rule again leads to :

$$\hat{\pi}(-x) = \pi \frac{\alpha\sigma_-(v_H)}{\alpha\sigma_-(v_H) + (1-\pi)(1-\alpha-\beta)\sigma_-(v_M)}$$

Thus, $\hat{\pi}(-x) < \pi$ if the expression above is well defined. Otherwise, suppose that the signal $X = -x$ is never transmitted in equilibrium. Fix an arbitrary out-of-equilibrium belief c defined by : $c = \mathbb{P}(v = v_H|\tilde{x} = -x)$. The probability of being the informed type is now given by :

$$\begin{aligned}
\mathbb{P}(\text{being well informed}|\tilde{x} = -x) &= \underbrace{\mathbb{P}(\text{being well informed}|\tilde{x} = -x, v = v_H)}_{\pi}\underbrace{\mathbb{P}(v = v_H|\tilde{x} = -x)}_{c} \\
&\quad + \underbrace{\mathbb{P}(\text{being well informed}|\tilde{x} = -x, v = v_H)}_{0}\underbrace{\mathbb{P}(v = v_M|\tilde{x} = -x)}_{1-c} \\
&= \pi c \\
&\leq \pi
\end{aligned}$$

This result means that the posterior probability of being the informed type cannot exceed the prior if $X = -x$, since the action $a = 1$ might have been taken by an informed intermediate type v_M who had received the incorrect information $m = x$.

3. By Bayes' rule, the probability of being the informed type after observing no signal is given by :

$$\hat{\pi}(\varnothing) = \pi \frac{\alpha + \frac{1-\alpha-\beta}{2} - \frac{\lambda\alpha}{2}\sigma_-(v_H) - \frac{\lambda\alpha}{2}\sigma_+(v_H) - \frac{\lambda(1-\alpha-\beta)}{2}\sigma_+(v_M)}{\alpha + \frac{1-\alpha-\beta}{2} - \frac{\lambda\alpha}{2}\sigma_-(v_H) - \frac{\lambda\alpha}{2}\sigma_+(v_H) - \pi\frac{\lambda(1-\alpha-\beta)}{2}\sigma_+(v_M) - (1-\pi)\frac{\lambda(1-\alpha-\beta)}{2}\sigma_-(v_M)}$$

Which is strictly higher than π iff $\sigma_-(v_M) > \sigma_+(v_M)$, equal to γ iff $\sigma_-(v_M) = \sigma_+(v_M)$ and strictly lower than γ iff $\sigma_-(v_M) < \sigma_+(v_M)$. The result means that, for the signal $\tilde{x} = \varnothing$ to improve the probability of being the informed type, the intermediate player who made the wrong decision (i.e. the player v_M having received $m = x$ instead of $m = -x$) must reveal the information strictly more than the intermediate player who made the right decision (i.e. the player v_M having correctly received $m = x$).

□

Proof of Lemma 7 Suppose that all the information is transmitted by the players, i.e. suppose that $\sigma_-(v_M) = \sigma_-(v_H) = \sigma_+(v_M) = \sigma_+(v_H) = 1$. These equations state that it is optimal for all the players to truthfully report their signals. However, lemma 10 shows that $\hat{\pi}(-x) < \hat{\pi}(\varnothing) = \pi < \hat{\pi}(x)$. Hence, both players v_H and v_M have a strict self-esteem incentive to lie and to conceal the feedback $X = -x$. The incentive to reveal $X = -x$ for proper second-period decision making must then be positive, which is impossible since in state $X = -x$, v_H wants to play $b = 1$ while v_M wants to play $b = -1$.

Proof of Proposition 8 Lemma 7, together with the assumption that κ is large, shows that some players must have a nonnegative self-esteem incentive to conceal their feedback information. Hence, the case $\hat{\pi}(\varnothing) < \hat{\pi}(-x)$ is ruled out.

First, notice that there always exists an equilibrium in which $\hat{\pi}(-x) < \hat{\pi}(\varnothing) < \hat{\pi}(x)$, and $\sigma_-(v_H) = \sigma_-(v_M) = 0, \sigma_+(v_H) = \sigma_+(v_M) = 1$. In this equilibrium, good news are always revealed, while bad news are always concealed. This equilibrium is sustained with an out-of-equilibrium belief $\mathbb{P}(v = v_M|\tilde{x} = -x) = c$ such

that $\hat{\pi}(-x) < \hat{\pi}(\varnothing)$, i.e. $\pi c < \pi \frac{\alpha + \frac{1-\alpha-\beta}{2} - \frac{\lambda\alpha}{2} - \frac{\lambda(1-\alpha-\beta)}{2}}{\alpha + \frac{1-\alpha-\beta}{2} - \frac{\lambda\alpha}{2} - \frac{\lambda\pi(1-\alpha-\beta)}{2}}$, and exists for all values of the parameters. Moreover, it is the most informative equilibrium satisfying $\hat{\pi}(-x) < \hat{\pi}(\varnothing)$.

Now, examine the case $\hat{\pi}(-x) = \hat{\pi}(\varnothing) = \hat{\pi}(x)$. Lemma 7 shows that the common value of these expressions is π. $\hat{\pi}(-x) = \pi$ implies that $\sigma_-(v_M) = 0$, and $\hat{\pi}(x) = \pi$ implies that $\sigma_+(v_M) = 0$. Moreover, $b(\varnothing) = 1$ since otherwise $\sigma_+(v_M) = 0$ would be a strictly dominated strategy. Suppose that $b(-x) = 0$. Then, $\sigma_-(v_H) = 0$ and this equilibrium is strictly less informative than the pure-strategy equilibrium defined above. Conversely, if $b(-x) = 1$, all the players are indifferent between transmitting or concealing their information, and the strategies σ can take any value between 0 and 1. If $\sigma_-(v_H) = 0$, the equilibrium is also strictly less informative than the pure-strategy equilibrium. If $\sigma_-(v_H) > 0$, however, the equilibrium does not satisfy the trembling hand perfection requirement. Indeed, if with some small positive probability ϵ, v_M reveals the feedback $X = -x$, the self-esteem incentives turn to $\hat{\pi}(-x) < \hat{\pi}(\varnothing)$, and $\sigma_-(v_H) = 0$ becomes a best reponse that does not converge to the positive candidate equilibrium value.

The last possible case is $\hat{\pi}(-x) = \hat{\pi}(\varnothing) < \hat{\pi}(x)$. This implies that $\sigma_+(v_M) = \sigma_+(v_H) = 1$. Moreover, $\hat{\pi}(-x) = \hat{\pi}(\varnothing)$ is equivalent to

$$\lambda\alpha\sigma_-(v_H) = [\alpha(2-\lambda) + (1-\alpha-\beta)(1-\lambda)]\sigma_-(v_M) \tag{28}$$

Which first shows that $\sigma_-(v_H) > \sigma_-(v_M)$, since $\lambda\alpha < \alpha(2-\lambda) + (1-\alpha-\beta)(1-\lambda)$. A necessary condition for this equilibrium to hold is that $b(\varnothing) = b(-x)$: indeed, since $(v_M, -x)$ and $(v_H, -x)$ have opposite preferences over the decision b, $b(\varnothing) \neq b(-x)$ together with $\hat{\pi}(-x) = \hat{\pi}(\varnothing)$ would imply that they would play opposite pure strategies, which contradicts 28. $b(-x) = 1$ is equivalent to $\mathbb{E}(X + v|\tilde{x} = -x) > 0$, which is (after some algebra) :

$$v_H[\alpha(2-\lambda) + (1-\alpha-\beta)(1-\lambda)] > x[\alpha(2-\lambda) + (1-\alpha-\beta)(1-\lambda\pi)] \tag{29}$$

And $b(\varnothing) = 1$ is equivalent to

$$\begin{aligned} &v_H\alpha(1-\frac{\lambda}{2}) - \frac{x}{2}[\lambda\alpha + (1-(2-\lambda)\pi)](1-\alpha-\beta) > \\ &\sigma_-(v_M)[v_H[\alpha(2-\lambda) + (1-\alpha-\beta)(1-\lambda)] - x[\alpha(2-\lambda) + (1-\alpha-\beta)(1-\lambda\pi)]] \end{aligned} \tag{30}$$

An equilibrium such that $b(-x) = b(\varnothing) = 1$ exists if and only if conditions 29 and 30 hold simultaneously for some value of $\sigma_-(v_M) \in (0, \frac{\lambda\alpha}{\alpha(2-\lambda)+(1-\alpha-\beta)(1-\lambda)})$. One can easily show that this is true if and only if α is sufficiently large.

Conversely, an equilibrium such that $b(-x) = b(\varnothing) = -1$ exists if and only if the complementary conditions of 29 and 30 hold simultaneously for some value of $\sigma_-(v_M) \in (0, \frac{\lambda\alpha}{\alpha(2-\lambda)+(1-\alpha-\beta)(1-\lambda)})$. One can see that this imposes a higher bound on α.

Finally, the two types of equilibria are :

- For all values of the parameters, a pure strategy equilibrium in which $\sigma_-(v_M) = \sigma_-(v_H) = 0$ and $\sigma_+(v_M) = \sigma_+(v_H) = 1$, where $\hat{\pi}(-x) < \hat{\pi}(\varnothing) < \hat{\pi}(x)$.
- For α sufficiently close to 1 or to 0, a class of equilibria in which $\sigma_+(v_M) = \sigma_+(v_H) = 1$ and $\sigma_-(v_M) < \sigma_-(v_H) \leq 1$, $\hat{\pi}(-x) = \hat{\pi}(\varnothing) < \hat{\pi}(x)$. If α is large (extreme identity), these equilibria are such that and $b(-x) = b(\varnothing) = 1$. If α is low (moderate identity), they are such that $b(-x) = b(\varnothing) = -1$.