Games and Economic Behavior 85 (2014) 127–158

Contents lists available at ScienceDirect

Games and Economic Behavior

www.elsevier.com/locate/geb

Imperfect memory and choice under risk ☆

Daniel Gottlieb

Wharton School, University of Pennsylvania, United States

ARTICLE INFO

Article history:
Received 8 July 2010
Available online 4 February 2014

JEL classification:
D03
C70

Keywords:
Memory
Self deception
Behavioral economics

ABSTRACT

This paper presents a model of choice based on imperfect memory and self-deception. I assume that people have preferences over their own attributes (e.g., skill, knowledge, or competence) and can manipulate their memories. The model provides a prior-dependent theory of regret aversion and allows for prior-dependent information attitudes. It implies that behavior will converge to the one predicted by expected utility theory after a choice has been faced a sufficiently large number of times.

1. Introduction

Choices with uncertain outcomes are an unavoidable part of a person's life. The outcomes often depend on the person's own attributes (e.g., skill, knowledge, or competence) and, therefore, influence the individual's self-views. Choices that turn out to be wrong typically lead to self-doubt, while choices that turn out to be right enhance the person's self-image. Hence, a person who cares about self-image will desire to manipulate recollections and beliefs. Indeed, there is abundant psychological evidence that people value a positive self-image and manipulate their memories (see Section 2).

This paper focuses on how the concern for self-image affects an individual's behavior under risk when memory is imperfect. I consider a model based on two basic premises: First, individuals have preferences over their own attributes; Second, they can influence what they will remember. Both assumptions are largely supported by psychological evidence. Apart from these two assumptions, individuals are assumed to behave as in standard economic models. Their preferences satisfy the axioms of expected utility theory. Furthermore, individuals follow Bayes' rule and, therefore, are aware of their memory imperfection. The model ties the concept of self-deception together with several deviations from standard expected utility theory, such as non-linear probability weights, risk aversion over lotteries with small stakes, regret aversion, and the competence hypothesis. It also generates endowment and sunk cost effects.

In its simplest version, the model consists of a two-period decision problem. In the first period, an individual observes the outcome of a signal, which is informative about her attributes. Then, she chooses the probability of forgetting the outcome of the signal. In the second period, the individual applies Bayes' rule to her recollection of the signal. Because

☆ I thank Muhamet Yildiz for insightful guidance and detailed suggestions, and Bengt Holmstrom and Drazen Prelec for valuable comments and suggestions. I also thank Eduardo Azevedo, Abhijit Banerjee, Roland Benabou, Moshe Cohen, Drew Fudenberg, Xavier Gabaix, Michael Grubb, Lucas Maestri, Wolfgang Pesendorfer, Andy Postlewaite, Jean Tirole, an Editor, and seminar participants at Boston University, Fucape, EPGE/FGV, Kellogg MEDS, NYU, MIT, Princeton, PUC Rio, Stanford Institute for Theoretical Economics, Toulouse School of Economics, UCSD, UCSB, Wharton, and Yale SOM for comments, and Yiwei Zhang for outstanding research assistance. I gratefully acknowledge financial support from the Dorinda and Mark Winkelman Distinguished Scholar Award.

E-mail address: dgott@wharton.upenn.edu.

http://dx.doi.org/10.1016/j.geb.2013.11.013
0899-8256/

Bayes' rule implies that, on average, the individual's interpretation of her recollections are correct, self-deception does not change her (ex-ante) expected self-views. Hence, from an ex-ante point of view, memory manipulation is wasteful and, therefore, the agent would prefer not to observe the realization of the signal. Nevertheless, after observing the signal, the individual has an incentive to manipulate her memory in order to improve her self-image.

The model generates a preference for avoiding information: people prefer not to acquire information if the expected benefit from making an informed decision is lower than the costs of self-deception. Because individuals anticipate these costs, they may prefer to make uninformed decisions if the objective value of information is sufficiently low. This result contrasts with Blackwell's celebrated theorem, which states that having additional information never hurts. It is consistent, however, with the large psychology literature that connects self-deception and information avoidance. For example, people often avoid health exams, especially if the value of information is not high enough (e.g. the disease is not easily treatable) and if being diagnosed with the disease significantly affects their self-image. Individuals also engage in "self-handicapping" strategies, such as under-preparing for an exam or getting too little sleep before physical exercise, in order to reduce the informational content of the signal. They may also have a "fear of competition," since outcomes from competitors are often informative about their own attributes.

The model formalizes the theory of regret aversion based on self-perception proposed by Josephs et al. (1992). According to this theory, individuals with low self-image are more likely to make choices that minimize the possibility of regret. The model is also consistent with behavior that Eliaz and Spiegler (2006) have shown to be inconsistent with models of "utility from beliefs."

When outcomes involve money, the individual may reject gambles with small but positive expected value unlike predicted by expected utility theory. The divergence from expected utility in my model is directly related to the decision maker's self-perceived attributes. This result is consistent with experimental evidence suggesting that deviations from expected utility theory are associated with the lotteries' being correlated with the decision maker's skill or knowledge.[1]

In a repeated setting in which the person observes a sequence of signals and manipulates her memory after observing of them, the attitude towards risk converges to the one implied by expected utility theory. This result is consistent with the arguments that people do not exhibit ambiguity aversion over events that have been observed several times and that experts may be subject to less bias than beginners (e.g. List, 2003; List and Haigh, 2005).

Two applications illustrate the theory. Successful trading usually requires certain skills or knowledge. At the very least, a potential buyer must form expectations about how much the good is worth. In more complex markets, future prices of the good must also be estimated. Thus, the outcome of the trade is informative about the person's skills or knowledge. Since decision makers avoid information correlated with skills or knowledge, they will accept to buy an object if the expected benefit exceeds a certain positive threshold. Therefore, self-deception generates an endowment effect.

The second application considers the influence of sunk decisions on behavior. In several contexts, revising one's decision usually involves admitting that a wrong decision was made and, therefore, it is often informative to the person about her own skills or knowledge. The model provides a self-deception explanation for the influence of sunk decisions on behavior that is consistent with arguments from the literature in psychology.

The structure of the paper is as follows. Section 2 briefly reviews the psychological evidence on the memory and the related literature in economics. Section 3 introduces and discusses the general framework. In Section 4, I describe the implications for information acquisition. Section 5 considers a repeated version of the model. Section 6 summarizes the main results and discusses possible extensions. In the appendix, I consider lotteries over money (Appendix A), and I present the two applications of the model (Appendix B).

2. Related literature

2.1. An overview of the psychology literature

> Ego-involvement, or its absence, makes a critical difference in human behavior. When a person reacts in a neutral, impersonal, routine atmosphere, his behavior is one thing. But when he is behaving personally, perhaps excitedly, seriously committed to a task, he behaves quite differently. In the first condition his ego is not engaged; in the second, it is. (Gordon W. Allport, 1943, p. 459).

Psychologists have largely documented a human tendency to deny or misrepresent reality to oneself (i.e., engage in self-deception). In general, people consider themselves to be "smart," "knowledgeable," and "nice." Information conflicting with this image is usually ignored or denied. Greenwald (1980, p. 605), for example, argued that "[o]ne of the best established recent findings in social psychology is that people perceive themselves readily as the origin of good effects and reluctantly as the origin of ill effects." Similarly, Gollwitzer et al. (1982, p. 702), claimed that the "asymmetrical attributions after success and failure" are a "firmly established finding."

People are also more likely to remember successes than failures (Korner, 1950). After choosing between two different options, they tend to recall the positive aspects of the chosen option and the negative aspects of the forgone option (Mather et al., 2003). Relatedly, individuals overestimate their achievements and readily find evidence that they possess attributes

[1] See, for example, Heath and Tversky (1991), Josephs et al. (1992), Fox and Tversky (1995), Goodie (2003), and Goodie and Young (2007).

which they believe to be correlated with success in personal or professional life (Kunda and Sanitioso, 1989; Quattrone and Tversky, 1984). Success is usually attributed to one's own ability and effort, whereas failure tends to be attributed to bad luck or other external variables (Gollwitzer et al., 1982; Zuckerman, 1979). In group settings, where each individual's contribution cannot be unequivocally determined, people tend to attribute to themselves a larger share of the group's outcome after a success and a smaller share after a failure (Johnston, 1967).

Self-assessments and the memory are intrinsically connected. In his *Essay Concerning Human Understanding*, Locke (1690) identified the self with memory. Mill (1829, Vol. 2, p. 174) argued that "[t]he phenomenon of Self and that of Memory are merely two sides of the same fact." Modern cognitive psychologists define the self as the "mental representation of oneself, including all that one knows about oneself" (Kihlstrom et al., 2003). Therefore, a model of self-views should devote considerable attention to memory.

In psychology, the memory is typically viewed as imperfect and manipulable. Rapaport (1961), for example, conceived "memory not as an ability to revive accurately impressions once obtained but as the integration of impressions into the whole personality and their revival *according to the needs of the whole personality*." Allport (1943) believed that self-deception was a mechanism of ego defense and the maintenance of self-esteem. Hilgard (1949, p. 374) argued that "the need for self-deception arises because of a more fundamental need to maintain or to restore self-esteem. Anything belittling the self is to be avoided." Festinger (1957) suggested that individuals seek consistency among their cognitions (i.e., beliefs and opinions). He labeled the discomfort felt when one is presented with evidence that conflicts with one's beliefs and the resulting effort to distort those beliefs or opinions cognitive dissonance. In a review of the recent literature in social psychology, Sedikides et al. (2004, p. 165) described people as "striving for a positive self-definition or the avoidance of a negative self-definition (…) at the expense of accuracy and truthfulness." According to them, "[m]emory serves the function of shielding a positive self-definition from negativity."

There are many reasons why people may want to believe in things that are not true. First, there may be a hedonic value of positive self-views so that people simply like to think that they have these attributes.[2] Second, as argued by Compte and Postlewaite (2004), they may benefit from having overconfident beliefs in situations in which emotions affect performance. Third, manipulating one's own beliefs may facilitate the deception of others. Thus, having positive self-views may help convincing others of one's own value.[3] Fourth, there may be a motivational value of belief manipulation. As Benabou and Tirole (2002) argued, higher self-confidence may help people set more ambitious goals and persist in adverse situations.

In this paper, I abstract from the exact reason why one may value a positive self-image. The model developed here is based on the two basic ideas discussed above. First, individuals have preferences over their attributes. Second, they can affect what they will remember. The paper focuses on how memory manipulations affect choices under risk.

As the opening quote from Allport illustrates, psychologists have long realized that self-deception may change a person's behavior. Festinger (1957, p. 3), for example, argued that "[w]hen dissonance is present, in addition to trying to reduce it, the person will actively avoid situations and information which would likely increase the dissonance." More recently, Josephs et al. (1992, p. 27) argued that "[r]isky decisions are potentially threatening to self-esteem because the chosen alternative will occasionally yield a less desirable outcome than would some other alternative. When a less desirable outcome does occur, it can sometimes lead people to doubt their judgment and ability, especially when the decision is an important one."

This paper shows that incorporating self-deception in a standard model of choice can lead to a unified theory of choice under risk that is consistent with economic phenomena such as ambiguity aversion, risk aversion over lotteries with small stakes, regret, and the competence hypothesis. It also provides a rationale for endowment and sunk cost effects.

2.2. An overview of the literature on imperfect memory

The economic literature on imperfect memory can be divided in two strands. The first assumes that decision makers are naive and act as if they have not forgotten anything (Mullainathan, 2002). The other strand assumes that decision makers are sophisticated and draw Bayesian inferences given that they might have forgotten things. This paper follows the latter approach and considers the case of rational decision makers subject to imperfect recall.[4] As suggested by Piccione and Rubinstein (1997), the resulting game of imperfect recall is solved by the principle of "multiself consistency," whereby decisions made in different stages are viewed as being made by different incarnations of the decision maker.

Models of limited memory are a special case of imperfect memory. They were originally proposed by Robbins (1956) in the mathematical statistics literature. He suggested a decision rule for choosing between two lotteries with unknown distributions that was conditional on a finite number of outcomes (finite memory).[5] More recently, economists have independently studied optimal decision making subject to limited memory. Dow (1991) considered the behavior of a consumer looking for the lowest price. Wilson (2003) studied how limited memory leads to certain biases in belief formation.

[2] For example, in Schelling's (1985) theory of the mind as a consuming organ, self-views have a hedonic value.

[3] As argued by Trivers (2000, p. 115), "[b]eing unconscious of ongoing deception may more deeply hide the deception. Conscious deceivers will often be under the stress that accompanies attempted deception." This argument is modeled formally by Byrne and Kurland (2001) in an evolutionary game.

[4] The online appendix considers the case of naive decision makers.

[5] In a series of papers, Cover and Hellman characterized optimal solutions to some finite memory problems. See Hellman and Cover (1973) for a review of the main results in this literature.

In a sequence of papers, Benabou and Tirole have used imperfect memory frameworks to study questions from the psychology literature. Based on the assumption that agents recalled actions but not their motivations, they have proposed theories of personal rules and internal commitments (Benabou and Tirole, 2004), prosocial behavior (Benabou and Tirole, 2006b), and identity and taboos (Benabou and Tirole, 2006c). Using a model of self-deception, they have analyzed the provision of self-motivation (Benabou and Tirole, 2002), the formation of collective beliefs and ideologies (Benabou and Tirole, 2006a; Benabou, 2013). Kopczuk and Slemrod (2005) have employed a similar model of self-deception to show denial of death may explain certain biases in intertemporal decision-making. Hirshleifer and Welch (2002) considered informational cascades generated by players who observe actions but not the information leading to such actions.

The model of memory presented here is based on Benabou and Tirole's self-deception framework. It encompasses a static version of the limited memory framework as a special case. This paper is also connected to the economic literature on cognitive dissonance (Akerlof and Dickens, 1982; Rabin, 1994). The cognitive dissonance literature assumes that agents derive utility from their beliefs and that they can, at some cost, choose their beliefs. Separately, Lowenstein (1987), Caplin and Leahy (2001, 2004), and Kőszegi (2006) studied models with anticipatory emotions.[6]

3. General framework

3.1. Decision problem

The model examines a decision maker (DM) with preferences over her own attributes θ. Attributes θ may be interpreted as skills, knowledge, or competence as well as a parameter of anticipatory utility. Let Θ be a non-empty subset of $\mathbb{R}$ representing the possible values of θ and let $F(.)$ denote the agent's prior distribution of θ.[7]

The DM acts in 3 periods ($t = 0, 1, 2$). In period 0, she chooses an (ex-ante) action a from a finite set A. For example, a can be an investment decision or a decision of whether to undertake some medical examination. The set A can also be a singleton, in which case the agent makes no choice in period 0. In some applications, the set A may also include the possibility of not observing a signal. I allow the DM to randomize between ex-ante actions, and denote a random action by $\phi_a \in \Delta(A)$.

In period $t = 1$, an outcome σ_a, which can be either high (H) or low (L), is observed. The outcome σ_a may be a purely informative signal, entering the agent's preferences only indirectly through her beliefs about her attributes θ. It may also affect the agent's preferences directly. For example, a medical exam consists of a purely informative signal, whereas the outcome of an investment affects an individual not only through its informational content but also through the monetary payment associated with it. Let $q_a \in (0, 1)$ denote the probability of observing a high outcome given action $a \in A$. I assume that a high outcome is more favorable than a low outcome in the sense of first-order stochastic dominance:

$$F(\theta|\sigma_a = H) \leqslant F(\theta|\sigma_a = L) \quad \text{for all } \theta \in \Theta, \tag{1}$$

with strict inequality for some value of θ, and for all $a \in A$.[8]

Following Rabin (1994) and Benabou and Tirole (2002, 2006a), I assume that the individual can, at a cost, influence her recollections. The DM remembers the outcome $\sigma \in \{H, L\}$ with probability $\eta_\sigma + m_\sigma$, where the parameter $\eta_\sigma \in [0, 1]$ is the agent's "natural" rate of remembering outcome σ. This rate determines the probability that the DM remembers the outcome if she does not attempt to manipulate her memory. However, exerting effort $m_\sigma \in [-\eta_\sigma, 1 - \eta_\sigma]$ in period $t = 1$ allows the DM to depart from this natural rate at a cost of $\psi_\sigma(m_\sigma) \geqslant 0$. Let $\hat{\sigma}_a \in \{H, L, \varnothing\}$ denote the recollection of outcome σ_a, where $\hat{\sigma}_a = \varnothing$ denotes a forgotten outcome.

In period $t = 2$, the DM chooses an (ex-post) action b from finite set B. For example, b can be a decision of whether to continue with the previous investment or whether to undertake some medical treatment. B can also be a singleton, in which case the DM does not act after observing the outcome. As in the case of ex-ante actions, I allow the DM to randomize, and denote a random ex-post action by $\phi_b \in \Delta(B)$. Fig. 1 presents the informational structure.

Preferences satisfy the axioms of expected utility theory over the set of attributes, actions, and realized signals. Therefore, there exists a utility function $u : \Theta \times A \times B \times \{H, L\} \to \mathbb{R}$ representing the DM's preferences.[9] Furthermore, $u(\theta, a, b, \sigma)$

[6] Brunnermeier and Parker (2005) proposed a theory of "optimal expectations," according to which individuals choose their beliefs balancing the gains from anticipating a higher future utility with the losses from suboptimal decision-making. Similarly, Hvide (2002) proposed the notion of "pragmatic beliefs," which are the beliefs that maximize the individual's utility. Bernheim and Thomadsen (2005) showed that memory imperfections and anticipatory emotions may lead to a resolution of Newcomb's Paradox and sustain cooperation in the Prisoners Dilemma.

[7] Θ can be continuous or discrete, as long as it contains at least two elements (otherwise, θ cannot be random). Since I have not assumed that decision makers hold a 'correct' prior distribution over θ, they are allowed to hold optimistic or pessimistic beliefs about their attributes.

[8] Note that inequality (1) does not hold if A includes the possibility of not observing a signal. In applications in which the agent may choose not to observe a signal, I will assume that (1) holds for all other ex-ante actions.

[9] Attributes such as skill or knowledge are not typically considered to be objects of choice. Nevertheless, people often invest in developing those attributes. Therefore, we can, in principle, elicit preferences using only choice data on people's willingness to pay for improving their attributes.

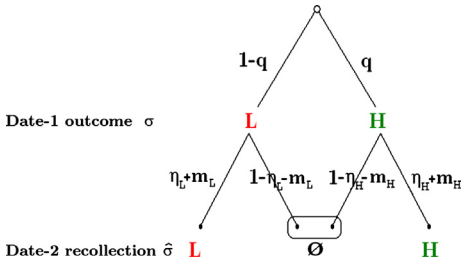

Fig. 1. Informational structure.

is strictly increasing in θ for all a, b, and σ. The utility function is extended to random actions ϕ_a and ϕ_b by taking expectations:

$$u(\theta, \phi_a, \phi_b, \sigma) = \sum_{\tilde{a} \in A} \sum_{\tilde{b} \in B} \phi_a(\tilde{a}) \phi_b(\tilde{b}) u(\theta, \tilde{a}, \tilde{b}, \sigma).$$

When $u(\theta, a, b, H) = u(\theta, a, b, L)$ for all (θ, a, b), we refer to outcomes as signals since they do not affect the agent's utility directly. In this case, we say that the model has *purely informative signals*. When signals are purely informative and A and B are singletons, we say that signals have a *purely hedonic value*. In models where signals have a purely hedonic value, the DM does not need to make any decision and the only reason for memory manipulation is the improvement of her self-views.

The recollection of an outcome $\hat{\sigma}$ always affects the agent's utility through beliefs about attributes θ. As described previously, however, the agent may also have preferences for the outcomes σ themselves holding their informational content fixed. For example, the outcome of an investment σ may combine both information about the investor's skills (which affects utility since its recollection $\hat{\sigma}$ will lead the agent to update beliefs about her attributes θ) and a monetary payment.[10] Consistently, we refer to the case where $u(\theta, a, b, H) \neq u(\theta, a, b, L)$ for some (θ, a, b) as a model of *monetary outcomes*. In this case, the realized outcomes σ are interpreted as monetary payments and each outcome affects the agent's utility both directly and through beliefs about θ.[11]

The cost of memory manipulation ψ_σ can be related to psychic costs (stress from repression of negative information or effort to focus on positive information), time (searching for reassuring information or excuses, lingering over positive feedback), or real resources (avoiding certain cues and interactions or eliminating evidence). It can also be interpreted as the shadow cost of memory in a limited information framework. Remembering an outcome with probability above its natural rate η_σ requires an individual to focus on it and on information correlated with it. In turn, this restricts the amount of attention available to store other information (which has shadow cost ψ_σ). Similarly, forgetting an outcome with probability above the natural rate $1 - \eta_\sigma$ requires an individual to focus on confronting evidence which again restricts the amount of attention available to other potentially useful information.[12]

Assumption 1. The cost of memory manipulation $\psi_\sigma(m_\sigma)$ is strictly decreasing in $m_\sigma < 0$, strictly increasing in $m_\sigma > 0$, convex, twice-continuously differentiable, and such that $\psi_\sigma(0) = 0$, $\sigma \in \{H, L\}$.

Fig. 2 depicts the costs of memory manipulation implied by Assumption 1. I further assume that the agent forgets a high outcome with some positive probability if she does not exert any effort[13]:

Assumption 2. $\eta_H < 1$.

The model can be seen as a conflict between an "interim self" and an "ex-post self." Holding ex-post actions constant, utility is increasing in self-image. Knowing this, the interim self cannot commit not to manipulate signals that contribute to

[10] Note that the outcome itself, and not its recollection, is the argument of the utility function. Therefore, we assume that the DM cannot perfectly infer the observed outcome from her income (either because consumption happens soon after the signal realization and, therefore, its recollection can be manipulated or because the DM does not know exactly how much income she had before the outcome). In fact, a previous version of the model considered a DM who had non-degenerate beliefs about her pre-outcome income, which were updated using both her post-outcome income and her recollection $\hat{\sigma}$. Since all results remained unchanged, I opted for the current, simpler version.

[11] When $u(\theta, a, b, H) > u(\theta, a, b, L)$ for all θ, a, b, the outcome associated with a higher monetary payment also provides favorable news about the DM's attributes. Although this is the most intuitive situation, a high monetary payment may also be associated with unfavorable news about the DM's attributes. This would be the case, for example, if the outcome is an employee's compensation for being laid off.

[12] For example, Steele's (1988) self-affirmation theory argues that people cope with negative outcomes in one domain by focusing on other, unrelated domains.

[13] The model becomes trivial when $\eta_H = 1$. Since the agent always remembers high outcomes, she will perfectly infer that a low outcome was observed if she recollects $\hat{\sigma} = \varnothing$. Therefore, she will never engage in memory manipulation.

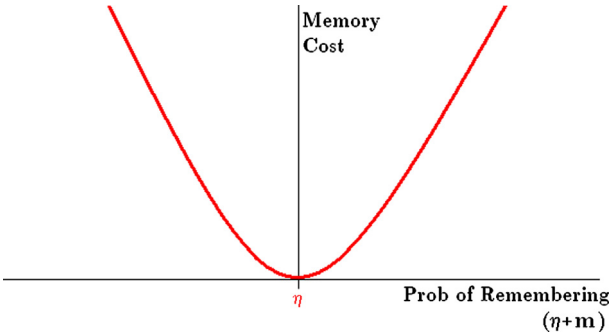

Fig. 2. Cost of memory.

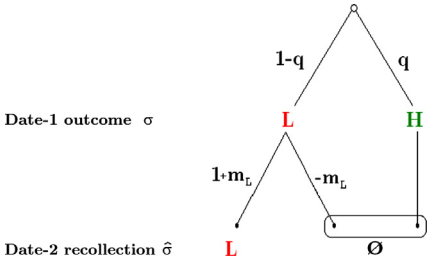

Fig. 3. Forgetfulness model.

self-image when given the chance.[14] The ex-post self, however, being a good Bayesian, makes a rational inference about the signal and updates her beliefs about attributes accordingly.[15]

As the following examples show, the general framework encompasses other models of imperfect memory.

Example 1 *(Forgetfulness model).* Take $\eta_L = 1$, $\eta_H = 0$, and $\psi_H(m_H) = +\infty$ for all $m_H > 0$ so that high outcomes are always forgotten (i.e., $\eta_H + m_H = 0$). Let ψ_L be strictly convex. Fig. 3 presents the informational structure in this case. This is the memory framework from Benabou and Tirole (2002). It can be interpreted as a model of bad news or no news. If the agent receives bad news, she can exert an effort $m_L \in [-1, 0]$ in order to forget them.[16]

If we reinterpret the state $\varnothing$ as the recollection of a high outcome, then the model from Example 1 becomes one where the agent is able to convince herself that a low outcome was a high outcome.[17] Hence, memory manipulation would allow the DM to believe that she observed a high outcome. This reinterpretation is compatible with neurological evidence from Prelec (2008), who showed that subjects experience heavy brain activity only when they try to convince themselves that a bad outcome was actually a good one. In the other states (both when they acknowledge a mistake or when they believe to have been correct), no such activity is detected. Example 1 can then be interpreted as the agent incurring psychological costs when she tries to convince herself that a bad outcome was actually a good one.

Example 2 *(Limited memory model).* Take $\eta_L = \eta_H = 0$ so that the DM forgets any outcome if she does not employ memory efforts and let ψ_σ be strictly convex. Then, the framework becomes a model of limited memory. In this model, the DM must allocate a limited amount of memory in order to store information. By spending a memory cost $\psi_\sigma(m_\sigma)$, she remembers an

[14] This interpretation assumes that the interim self is rational in the sense of taking into account the benefits and costs of memory manipulation. Several papers in social psychology have documented that individuals tend to be more realistic and impartial when making important decisions (cf., Taylor and Gollwitzer, 1995, and references therein). Therefore, self-deception seems to decrease when the cost of a mistake increases. Mijovic-Prelec and Prelec (2010) presents experimental evidence suggesting that self-deception responds positively to its expected benefits.

Similarly to this interpretation, Bodner and Prelec (2002) present a signaling model between an agent's privately informed gut and the agent's uninformed mind.

[15] The model can be interpreted as a formalization of the neurophysiological argument put forth by Trivers (2000). According to this interpretation, the interim self would be the person's unconscious process of information manipulation. In the context of intertemporal choice, several papers have proposed dual self models (cf. Thaler and Shefrin, 1981; Fudenberg and Levine, 2006, and Brocas and Carrillo, 2008).

[16] In fact, the model of Benabou and Tirole (2002) is more general than the one from Example 1. Since they assume that the cost of memory is U-shaped, their model also allows for $\eta_L \in (0, 1]$. The model of Benabou and Tirole (2006a) also allows for $\eta_L \in (0, 1]$ but assumes that the cost function is piecewise linear.

[17] In this model, the agent would never choose to believe that a high outcome was actually low.

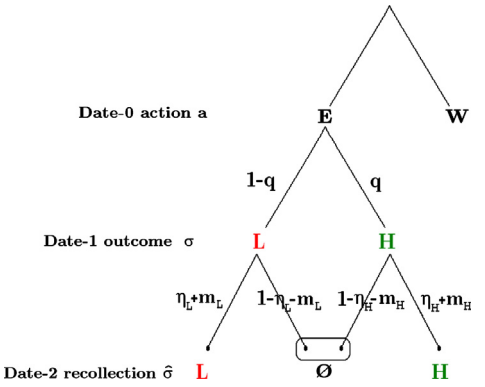

Fig. 4. Entrepreneurship example.

outcome $\sigma \in \{H, L\}$ with probability m_σ. A higher effort m_σ can be interpreted as having greater memory resources used to store the information.[18]

The following example presents an application of the general framework to an economic environment:

Entrepreneurship example. An employed individual is considering quitting her job and starting a new company. Building a successful company requires entrepreneurial skills. The individual does not know her exact skills. Therefore, a success improves the individual's self-assessment of her skills. If she decides not to quit her job, she obtains a wage $w \in \mathbb{R}_+$ and does not learn anything about her skills.

In this paper, this situation is modeled as follows. Let the individual's career choice be denoted by $a = E$ if she becomes an entrepreneur and by $a = W$ if she remains a worker, and let θ denote the individual's entrepreneurial skills. The outcome from starting a company is denoted by σ, which is equal to H in the case of success and L in the case of failure. After the outcome σ is observed, the entrepreneur may engage in memory manipulation. In this model, there is no ex-post choice (B is a singleton). The agent's decision tree is presented in Fig. 4.

Appendix C considers a more realistic model, where outcomes are vectors $\sigma = (s, r)$. The binary variable $s \in \{S, F\}$ reflects whether the company succeeded or failed and $r \in \mathbb{R}$ is a variable that is independent of the agent's attributes but affects outcomes (e.g. general market conditions or economy-wide shocks). The entrepreneur always remembers whether the company succeeded or failed but may forget the prevailing external conditions r.

Succeeding under adverse conditions indicates higher skills. Similarly, failing under favorable conditions indicates lower skills. In this model, the agent will manipulate her memory to forget positive external shocks and remember negative shocks. This result is consistent with the psychological literature described in Section 2, which shows that success is usually attributed to one's own attributes whereas failure tends to be attributed to bad luck or other external variables.

Appendix A shows that self-deception will prevent some individuals from becoming entrepreneurs even when the expected monetary payoffs from starting a new company are higher than the payoff from remaining on the previous job.

3.2. Modeling as a multiself game

This paper follows Piccione and Rubinstein (1997) in modeling a decision problem with imperfect memory as a game between different selves. The decision maker is treated as a collection of selves, each of them unable to control the behavior of future selves.

There are two players: self 1 and self 2. Both players share the same utility function but have different information sets. In period 0, self 1 chooses a probability distribution over ex-ante actions ϕ_a. Then, nature plays an ex-ante action $a \in A$ according to this probability distribution, and plays a high outcome with probability q_a and a low outcome with probability $1 - q_a$. In period 1, conditional on the outcome $\sigma \in \{H, L\}$, self 1 decides the amount of memory manipulation m_σ. Then, given the outcome σ and the manipulation effort m_σ, nature plays $\hat{\sigma} = \sigma$ or $\hat{\sigma} = \varnothing$ with probabilities $\eta_\sigma + m_\sigma$ and $1 - \eta_\sigma - m_\sigma$. In period 2, self 2 observes the recollection $\hat{\sigma}$ and chooses a probability distribution over ex-post actions ϕ_b. Then, nature plays an ex-post action $b \in B$ according to this distribution and both selves get payoff $E[u(\theta, a, b, \sigma)|\hat{\sigma}] - \psi_\sigma(m_\sigma)$.

[18] Dow (1991) considers a consumer who searches sequentially for the lowest price, but who only remembers each price as belonging to a finite number of categories. Wilson (2003) considers a decision-maker who must act after a large number of periods but whose memory is restricted to a finite number of states.

Because the DM has preferences over θ, she has an interim incentive to manipulate her beliefs. However, the set of possible beliefs that an agent can hold is restricted by the assumption that recollections are interpreted according to Bayes' rule. Thus, the agent makes correct inferences about her attributes θ given her recollections $\hat{\sigma}$.

Remark 1 *(Conservatism).* There is sizeable psychological evidence that people update their beliefs too slowly compared to the update implied by Bayes' rule (Conservatism Bias).[19] The model predicts that decision makers will display a conservatism bias. Formally, let θ_{σ_a} denote the expected value of θ conditional on the observed outcome σ_a and let $\hat{\theta}_{\hat{\sigma}_a}$ denote the expected attributes conditional on the recollection $\hat{\sigma}_a$. Appendix D shows that $\hat{\theta}_{\hat{\sigma}_a}$ is "less variable" than θ_{σ_a} in the sense of second-order stochastic dominance. Therefore, because θ_{σ_a} is the Bayesian estimate of θ given the outcome σ_a, forgetfulness implies that the decision maker updates observed outcomes σ_a less than implied by Bayes' rule.

3.3. Solution concept

The decision made by the agent with imperfect recall is modeled as the perfect Bayesian equilibrium (PBE) of the multiself game. Let $\mu(.|\hat{\sigma})$ denote the DM's posterior beliefs about θ given $\hat{\sigma}$ and let $E_\mu[.|\hat{\sigma}]$ denote the expectation operator with respect to $\mu(.|\hat{\sigma})$. Given a vector of memory manipulation (m_L, m_H), let $E_{\hat{\sigma}_a}[.|m_L, m_H]$ denote the expectation with respect to the distribution of $\hat{\sigma}_a$.

Definition 1. A PBE of the game is a strategy profile $(\phi_a^*, \phi_b^*, m_H^*(a), m_L^*(a))$ and posterior beliefs $\mu(.|\hat{\sigma}_a)$ such that:

1. $\phi_a^* \in \arg\max_{a\in A}\{E_{\hat{\sigma}_a}[E_\mu[u(a, \phi_b^*(a, \hat{\sigma}_a), \theta, \sigma_a)|\hat{\sigma}_a]|m_L^*(a), m_H^*(a)] - q\psi_H(m_H^*(a)) - (1-q)\psi_L(m_L^*(a))\}$;
2. $m_\sigma^*(a) \in \arg\max_{m_\sigma}\{(\eta_\sigma + m_\sigma)E_\mu[u(a, \phi_b^*(a, \hat{\sigma}_a), \theta, \sigma)|\hat{\sigma}_a = \sigma] + (1 - \eta_\sigma - m_\sigma)E_\mu[u(a, \phi_b^*(a, \hat{\sigma}_a), \theta, \sigma)|\hat{\sigma}_a = \varnothing] - \psi_\sigma(m_\sigma)\}$, $\sigma \in \{H, L\}$, $a \in A$;
3. $\phi_b^*(a, \hat{\sigma}_a) \in \arg\max_{b\in B}\{E_\mu[u(a, b, \theta, \sigma_a)|\hat{\sigma}_a = \hat{\sigma}]\}$, $a \in A$;
4. $\mu(\theta|\hat{\sigma}_a = \hat{\sigma})$ is obtained by Bayes' rule if $\Pr(\hat{\sigma}_a = \hat{\sigma}|m_L^*(a), m_H^*(a)) > 0$, $\hat{\sigma} \in \{L, H, \varnothing\}$, $a \in A$.

Conditions 1–3 are the standard perfection conditions. Condition 1 states that self 1 chooses an ex-ante action a with positive probability if and only if this action maximizes the agent's ex-ante expected utility given the behavior of self 2. Condition 2 states that self 1 chooses the amount of manipulation that maximizes the agent's expected payoff conditional on each outcome $\sigma \in \{H, L\}$. Condition 3 states that self 2 takes an ex-post action b with positive probability if and only if this action maximizes her utility given the beliefs she holds about the manipulation employed by self 1.

Condition 4 is a consistency condition, requiring that beliefs of self 2 satisfy Bayes' rule given the strategy of self 1. This condition is stronger than Bayesian consistency since it requires beliefs to be consistent even for certain events that may not occur on the equilibrium path (namely, events associated with ex-ante actions that are not played in equilibrium). This condition ensures that the PBE satisfies subgame perfection, which would not be the case if one only restricted beliefs on the equilibrium path to be consistent with Bayes' rule.[20]

The properties of the PBE can be studied in two parts by applying a backward induction algorithm. First, we compute the PBE of the continuation games starting after each ex-ante action $a \in A$. Second, we obtain the distributions over ex-ante actions ϕ_a^* chosen by self 1 given the PBE of the continuation games. The existence of a PBE is established in the online appendix.

Fix a PBE. For each ex-ante action $a \in A$, under every recollection $\hat{\sigma}$ that occurs with positive probability under $\{m_H^*(a), m_L^*(a)\}$, Condition 4 implies that posterior beliefs μ must be consistent with Bayes' rule. Therefore, Condition 3 yields

$$\phi_b^*(a, \hat{\sigma}_a) \in \arg\max_{b\in B} \int u(a, b, \theta, \sigma)\, dF(\theta|\hat{\sigma}_a = \hat{\sigma}),$$

for any recollection $\hat{\sigma}$ that is reached with positive probability in the continuation game following the ex-ante action a.

Denote the expected utilities given $\sigma_a = H$ and $\sigma_a = L$ by

$$u_H(a, \phi_b, \sigma_a) \equiv \int u(a, \phi_b, \theta, \sigma_a)\, dF(\theta|\sigma_a = H), \quad \text{and}$$
$$u_L(a, \phi_b, \sigma_a) \equiv \int u(a, \phi_b, \theta, \sigma_a)\, dF(\theta|\sigma_a = L). \tag{2}$$

[19] According to Edwards (1968, p. 359) beliefs change "very orderly, and usually proportional to numbers calculated from Bayes' theorem – but in insufficient amount." See Edwards (1982) for a review of this literature.

[20] See Fudenberg and Tirole (1991, pp. 331–333) for a related discussion. Here, the additional restrictions on the updating rule are straightforward since the ex-ante action is observed by both selves and, therefore, each $a \in A$ induces a proper subgame.

If self 2 recalls a high outcome ($\hat{\sigma}_a = H$), then she must assign probability one to the event $\sigma_a = H$. Hence, apart from manipulation costs, self 1 obtains expected utility $u_H(a, \phi_b(a, H), H)$ if self 2 recalls a high outcome. Similarly, her expected utility conditional on self 2 recalling a low outcome is $u_L(a, \phi_b(a, L), L)$.

Let $m_L^*(a)$ and $m_H^*(a)$ denote the amounts of memory manipulation self 2 believes that self 1 has exerted in period 1 conditional on action $a \in A$. The PBE concept implies that self 1 takes these amounts as given when choosing how much memory manipulation to exert. When the DM forgets the period-1 outcome, self 2 assigns probability $\alpha_a(m_L^*(a), m_H^*(a))$ to a high outcome and $1 - \alpha_a(m_L^*(a), m_H^*(a))$ to a low outcome, where

$$\alpha_a(m_L, m_H) \equiv \frac{q_a(1 - \eta_H - m_H)}{q_a(1 - \eta_H - m_H) + (1 - q_a)(1 - \eta_L - m_L)}$$

is the conditional probability of a high outcome given by Bayes' rule. Then, the expected utility when the outcome is forgotten is

$$u_\varnothing\big(a, \phi_b^*(a, \varnothing), \sigma_a\big) \equiv \alpha_a\big(m_L^*, m_H^*\big) u_H\big(a, \phi_b^*(a, \varnothing), \sigma_a\big) + \big[1 - \alpha_a\big(m_L^*, m_H^*\big)\big] u_L\big(a, \phi_b^*(a, \varnothing), \sigma_a\big). \tag{3}$$

After observing a high outcome, self 1 chooses m_H to maximize[21]

$$\begin{aligned}(\eta_H + m_H)\big\{&\big[1 - \alpha_a\big(m_L^*(a), m_H^*(a)\big)\big]\big[u_H\big(a, \phi_b^*(a, \varnothing), H\big) - u_L(a, \phi_b^*(a, \varnothing), H)\big] \\ &+ u_H\big(a, \phi_b^*(a, H), H\big) - u_H\big(a, \phi_b^*(a, \varnothing), H\big)\big\} + u_\varnothing\big(a, \phi_b^*(a, \varnothing), H\big) - \psi_H(m_H).\end{aligned} \tag{4}$$

Self 1 takes three factors into account when choosing the amount of memory manipulation. First, remembering a high signal leads to a higher utility through a more favorable inference about θ since $u_H(a, \phi_b^*(a, \varnothing), H) > u_L(a, \phi_b^*(a, \varnothing), H)$ (*self-image* factor). Second, it leads to better decision-making since $u_H(a, \phi_b^*(a, H), H) \geqslant u_H(a, \phi_b^*(a, \varnothing), H)$ (*decision-making* factor). Third, it leads to a cost of $\psi_H(m_H)$ (*memory cost* factor).

Both the self-image and the decision-making factors induce the DM to try to remember a high signal. Therefore, because small amounts of memory manipulation have second-order costs, the DM always remembers a high signal with probability greater than her natural rate η_H:

Proposition 1 *(Remembering good news). In any PBE, $m_H^*(a) > 0 \ \forall a \in A$.*

After observing a low outcome, self 1 chooses m_L to maximize

$$\begin{aligned}&u_L\big(a, \phi_b^*(a, \varnothing), L\big) + (\eta_L + m_L)\big[u_L\big(a, \phi_b^*(a, L), L\big) - u_L\big(a, \phi_b^*(a, \varnothing), L\big)\big] \\ &\quad + (1 - \eta_L - m_L)\alpha_a\big(m_L^*(a), m_H^*(a)\big)\big[u_H\big(a, \phi_b^*(a, \varnothing), L\big) - u_L\big(a, \phi_b^*(a, \varnothing), L\big)\big] - \psi_L(m_L).\end{aligned} \tag{5}$$

As in the case of a high signal, self 1 takes self-image, decision-making, and memory costs into account when deciding the amount of memory manipulation. However, in the case of low signals, the self-image and the decision-making factors have opposite signs.[22] Therefore, the DM faces a conflict between forgetting a low signal and having a better self-image or remembering it and making better decisions. In the special case where there are no ex-post decisions, the decision-making factor vanishes and the DM always remembers a low signal with probability below the natural rate η_L:

Proposition 2 *(Forgetting bad news with no ex-post decisions). Let $B = \{b_0\}$ be a singleton. Then in any PBE, $m_H^*(a) > 0 \geqslant m_L^*(a)$ for any $a \in A$. Furthermore,*

$$u_H(a, b_0, H) - u_L(a, b_0, H) < \psi_H'(1 - \eta_H) \quad \Longrightarrow \quad 0 < m_H^*(a) < 1 - \eta_H \quad \text{and} \quad m_L^*(a) < 0.$$

If the marginal benefit of remembering good news is high enough relative to its marginal cost, high signals are remembered with probability one. Then, there is no point in trying to forget a low signal since the DM perfectly infers that a low signal was observed if the signal is forgotten. In this case, the equilibrium features full manipulation after a high signal and no manipulation after a low signal ($m_H^* = 1 - \eta_H$ and $m_L^* = 0$). Proposition 2 shows that when the marginal benefit of remembering good news is lower than its marginal cost for some m_H, the DM forgets high signals with positive probability. Then, because the cost of a small amount of memory manipulation is of second-order, bad news are remembered with probability below the natural rate η_L, i.e., $m_L^* < 0$.

[21] This expression is obtained by substituting Eq. (3) on self-1's expected utility

$$(\eta_H + m_H) u_H\big(a, b_a(H), H\big) + (1 - \eta_H - m_H) u_\varnothing\big(a, b_a(\varnothing), H\big) - \psi_H(m_H).$$

[22] The self-image and decision-making effects of remembering a low signal are captured by $u_L(a, b_a(\varnothing), L) - u_H(a, b_a(\varnothing), L) < 0$ and $u_L(a, b_a(L), L) - u_L(a, b_a(\varnothing), L) \geqslant 0$, respectively.

Example 1 *(with purely hedonic signals).* Consider the Forgetfulness Model under purely hedonic signals. Let $\Delta u \equiv u_H - u_L$ denote the benefit of memory manipulation. This model has an essentially unique PBE.[23] Because self 1 chooses memory manipulation by comparing its costs and benefits, it follows that the amount of manipulation $|m_L^*|$ is: (i) increasing in the benefit of manipulation Δu (for u_L fixed), (ii) decreasing in the marginal cost of manipulation, and (iii) increasing in q, the probability of not observing a signal.[24] These comparative statics results are consistent with the experimental evidence presented by Mijovic-Prelec and Prelec (2010), which suggests that self-deception is increasing in the benefits of manipulation.

Example 2 *(with purely hedonic signals).* Consider the Limited Memory model under purely hedonic signals. The set of PBE memory manipulations is characterized by:

$$\frac{(1-q)\Delta u}{1-q+q(1-m_H^*)} = \psi_H'(m_H^*) \quad \text{if } \Delta u \leqslant \left[1+\frac{q}{1-q}(1-m_H^*)\right]\psi_H'(1), \quad \text{and}$$
$$m_H^* = 1 \quad \text{if } \Delta u \geqslant \psi_H'(1). \tag{6}$$

Since both sides of the first equation in (6) are increasing in m_H^*, there may be multiple interior equilibria. It is also possible to have an interior and a corner equilibrium.[25] Individuals who believe that they often forget good signals are not hurt much by failing to recall good signals. Therefore, they will not manipulate their memories enough and, in equilibrium, they will often forget good signals. On the other hand, individuals who usually remember good signals are severely hurt by forgetting a good signal. Therefore, they will have greater incentives to remember good signals. As I will show next, these equilibria are welfare ranked (from an ex-ante perspective): The equilibrium with the lowest amount of memory manipulation is preferred. The individual may be caught in a self-trap where she exerts more manipulation effort because self 2 believes that she will have engaged in more memory manipulation.[26]

3.4. Ex-ante utility

As in other decision problems with imperfect recall, the timing of decisions matters. If the agent could commit to a strategy at an ex-ante stage, she would generally choose a different amount of memory manipulation.

Consider, for example, the model of purely hedonic signals. Bayesian updating implies that the individual is not fooled on average and the effects of memory manipulation on the expected self-image cancel out. Then, since there are no decisions to be made, the individual would be better off by exerting no memory manipulation and saving its costs. However, after the signal is observed, the individual has an incentive to manipulate her self-image by trying to forget a low signal and remember a high signal. Therefore, the ex-ante optimal strategy is time-inconsistent.[27]

When the individual has to take actions ex-post, it may be ex-ante optimal to exert some positive amount of memory manipulation.[28] Nevertheless, because self 1 also takes self-image into account when deciding the amount of memory manipulation, and the self-image effects cancel out from an ex-ante perspective, the individual would prefer to commit to a lower probability of remembering good news and a lower probability of forgetting bad news. Denote the ex-ante expected utility given the ex-ante action $a \in A$ by

$$\mathcal{U}_a(m_H, m_L) \equiv \max_{\{b(\hat{\sigma})\in B\}_{\hat{\sigma}=L,H,\varnothing}} E\left[u_{\hat{\sigma}}\left(a, b(\hat{\sigma}), \sigma\right)\middle| m_L, m_H\right] - q\psi_H(m_H) - (1-q)\psi_L(m_L). \tag{7}$$

Proposition 3 establishes this result formally:

Proposition 3 *(Excessive manipulation). Let $(\tilde{m}_H(a), \tilde{m}_L(a))$ be a maximizer of $\mathcal{U}_a$ and suppose $\mathcal{U}_a$ is concave.[29] Then, in any PBE with manipulations $m_H^*(a)$ and $m_L^*(a)$,*

[23] See Appendix D.2 for the characterization of the PBE. It is essentially unique in the sense that all PBE feature the same choices of actions a and b, the same manipulation efforts m_L and m_H, and the same beliefs for all recollections that are reached with positive probability. Equilibria may diverge only with respect to beliefs at recollections that are not reached with positive probability.

[24] In the model of Example 1, no news is good news. Therefore, when the probability of not observing a signal q is higher, it becomes more credible that the DM has not manipulated her beliefs into forgetting a low signal. Hence, an increase in q increases the marginal benefit of self-deception, which in turn leads to an increase in the amount of memory manipulation $|m_L^*|$.

[25] For example, if $\psi_H'(1) \leqslant \Delta u \leqslant \psi_r'(1)[1+\frac{q}{1-q}(1-m_H^*)]$, there exist both an equilibrium with $m_H^* = 1$ and an interior equilibrium with m_H^* implicitly defined by Eq. (6).

[26] The possibility of multiple equilibria has been previously established by Benabou and Tirole (2002). It is interesting since there seems to be a large heterogeneity in the amount of self-deception across different people (cf., Prelec, 2008). However, since the results presented here hold in all PBE, they would also be obtained if one applied a selection criterion.

[27] This argument resembles the signal-jamming models of Holmstrom (1999) and Dewatripont et al. (1999). See Piccione and Rubinstein (1997) for a discussion of decision problems with imperfect recall.

[28] More precisely, let $b(\hat{\sigma})|_{m_L,m_H}$ denote an action that maximizes the DM's utility given recollection $\hat{\sigma}$ and conditional on manipulation efforts m_L and m_H. Then, whenever $b(H)|_{m_L=m_H=0} \neq b(\varnothing)|_{m_L=m_H=0}$, the manipulation effort m_H that maximizes the ex-ante expected utility is strictly positive. Analogously, if $b(L)|_{m_L=m_H=0} \neq b(\varnothing)|_{m_L=m_H=0}$ then m_L that maximizes the ex-ante expected utility is strictly positive.

[29] It is straightforward to show that $\mathcal{U}$ is always concave when B is a singleton.

$$m_H^*(a) \geqslant \tilde{m}_H(a) \quad \text{and} \quad m_L^*(a) \leqslant \tilde{m}_L(a)$$

for all $a \in A$, with at least one of the inequalities being strict.

4. Purely informative signals and information acquisition

Suppose the decision maker can choose whether or not to observe a signal. When would she prefer to do so?

The standard theory of information acquisition states that it is optimal to observe a signal when the value of information (defined as the expected payoff gained by observing the signal) is greater than the signal's cost. Similarly, I will show that the DM prefers to observe a signal if the "objective value of information" exceeds the expected cost of self-deception. In particular, when information has purely hedonic value, the DM always prefers not to observe any signal.

The *objective value of information* is defined as the expected payoff from observing the signal:

$$V \equiv \max_{\phi_a \in \Delta(A), \{\phi_b(\hat{\sigma}_a)\} \in \Delta(B)} E_{\hat{\sigma}_a}\left[E_\mu\left[u\left(\phi_a, \phi_b(\hat{\sigma}_a), \theta\right)\middle|\hat{\sigma}_a\right]\right] - \max_{\phi_a \in \Delta(A), \phi_b \in \Delta(B)} E\left[u(\phi_a, \phi_b, \theta)\right] \geqslant 0, \tag{8}$$

where $\max_{\phi_a, \phi_b} E[u(\phi_a, \phi_b, \theta)]$ is the expected payoff if the DM could not observe $\hat{\sigma}_a$. From Eq. (7), the ex-ante expected utility from observing the signal, $U(\Sigma)$, equals

$$\left\{\max_{\{\phi_a \in \Delta(A), \phi_b \in \Delta(B)\}} E\left[u(\phi_a, \phi_b, \theta)\right]\right\} + V - \int_{a \in A} \phi_a^*\left[q_a \psi_H\left(m_H^*(a)\right) - (1-q)\psi_L\left(m_L^*(a)\right)\right] da. \tag{9}$$

Thus, the DM prefers to observe the signal if the objective value of information V is greater than the expected cost of memory manipulation $\int_{a \in A} \phi_a^*[q_a \psi_H(m_H^*(a)) - (1-q)\psi_L(m_L^*(a))]\, da$.

Proposition 4 *(Information acquisition). Fix a PBE. Let $U(\Sigma)$ denote the expected utility of observing the signal in this PBE and let $E[u]$ denote the expected utility of not observing the signal. Then, $U(\Sigma) - E[u] = V - \int_{a \in A} \phi_a^*[q_a \psi_H(m_H^*(a)) - (1-q)\psi_L(m_L^*(a))]\, da < V$.*

When information has a purely hedonic value, the objective value of information V is equal to zero. In equilibrium, when a signal is forgotten ($\hat{\sigma} = \varnothing$), the DM knows that there is a probability $\alpha_a(m_L^*, m_H^*)$ that there was a high signal and $1 - \alpha_a(m_L^*, m_H^*)$ that there was a low signal. Bayesian updating implies that on average, the only effects of engaging in self-deception are the manipulation costs $\psi_L(m_L^*)$ and $\psi_H(m_H^*)$. Of course, there is still an interim incentive to manipulate beliefs after she observes the signal. *The inability to commit not to engage in self-deception leads to a loss in (ex-ante) expected utility*:

Corollary 1. *When information has purely hedonic value, the DM is strictly better off by not observing the signal: $E[u] > U(\Sigma)$.*

Proposition 4 and Corollary 1 show that memory manipulation leads to preferences for avoiding information when individuals care about their attributes (i.e., they have "ego utility"). The most standard model of ego utility one could formulate consists of a basic application of expected utility theory. Let the space of possible attributes Θ be a non-empty subset of $\mathbb{R}$ and let $F(.)$ denote the agent's prior distribution of θ. The DM has preferences that are represented by a strictly increasing von Neumann–Morgenstern utility function $u : \Theta \to \mathbb{R}$.

In this basic model, if the individual does not observe a signal that is informative about θ, her utility is $\int u(\theta)\, dF(\theta)$. If she observes a signal σ, the utility conditional on σ is $\int u(\theta)\, dF(\theta|\sigma)$. Hence, the expected utility of observing the signal is $\int_\sigma \int_\theta u(\theta)\, dF(\theta|\sigma)\, dG(\sigma)$, where G is the distribution of signals σ. By the law of iterated expectations, we have

$$\int u(\theta)\, dF(\theta) = \int_\sigma \int_\theta u(\theta)\, dF(\theta|\sigma)\, dG(\sigma),$$

so that *an individual with perfect memory and who behaves as an expected utility maximizer is always indifferent between observing the signal or not* when signals do not affect actions. In other words, in this standard model of ego utility, the fact that an individual has preferences over her expected attributes does not influence her decision of whether to acquire information. In particular, as in Blackwell's theorem, more information never hurts.

The result above holds regardless of the shape of the utility function u. In order to affect the decision of whether to acquire information, the utility function must be a non-linear function of *probabilities*. Several models of information acquisition have, thus, assumed that utility functions are non-linear in probabilities.[30] Our model also leads to a utility

[30] For example, Philipson and Posner (1995) and Caplin and Eliaz (2003) analyze the case of testing for sexually transmitted diseases, Kőszegi (2003) considers a model of patient decision-making, Kőszegi (2006) studies information acquisition and financial decisions, and Caplin and Leahy (2004) study strategic information transmission. With the exception of Philipson and Posner (1995), who do not provide a justification for the assumption of a utility function that is non-linear in probabilities, all these papers depart from the standard expected utility model by adopting the Psychological Expected Utility model.

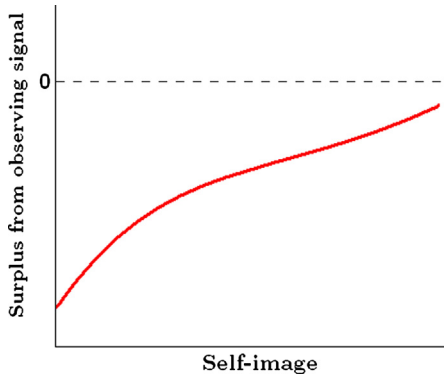

Fig. 5. Regret aversion based on self-perceptions.

function that is non-linear in probabilities. However, the non-linearity arises endogenously through memory manipulation. Therefore, the memory manipulation model can be seen as providing a cognitive foundation for a model of information avoidance.

Proposition 4 shows that the DM will prefer not to collect some information if its objective value V is lower than the expected costs from memory manipulation. In particular, she will always prefer not to observe a signal that is informative about her attributes θ but does not affect her actions b. For example, people will prefer not to know the outcome of a medical exam if the value of information is not sufficiently high (e.g. if a detected disease is not treatable) and if the exam has a potentially large impact on the person's self-image. Dawson et al. (2006) present experimental evidence supporting this result.[31]

An immediate consequence of avoiding information correlated with one's skills is the possible desirability of "self-handicapping" strategies such as under-preparing for an examination or getting too little sleep before a physical exercise (Berglas and Baumeister, 1993). Self-handicapping strategies reduce the informational content of the signal, and therefore, the model predicts that a person may engage in such strategies if the expected costs are not too high.

In several environments, competition allows for more precise information about one's abilities. Thus, individuals may display a "fear of competition" and prefer environments their outcomes are not directly comparable to other people's outcomes. More generally, the model predicts that in environments where information is correlated with one's attributes, individuals typically face a trade-off between the objective value of information and the costs of self-deception. Coarser information structures reduce the objective value of information but cause lower self-deception costs.

4.1. Regret aversion

This subsection studies how the decision markers' utility from the signal changes as a function of their prior distribution about their attributes. This allows us to show that the model developed in this paper provides a formalization for the (informal) theory of regret aversion based on self-evaluation proposed by Josephs et al. (1992).

The theory of regret aversion based on self-perceptions. The theory of choice based on regret aversion was simultaneously proposed by Bell (1982) and Loomes and Sugden (1982). According to this theory, agents base their decisions not only on expected payoffs but also on the payoffs that they would have obtained if they had made other decisions. Agents anticipate feeling regret or delight over their choice and take these feelings into account when making a decision.

Josephs et al. (1992) argued that the feeling of regret arises from the self-evaluation that follows an outcome.[32] They suggested that people with worse self-perceptions are more severely harmed by negative outcomes than those with better self-perceptions. Therefore, *individuals with low self-image would be more likely to make choices that minimize the possibility of regret*.

According to the theory of regret aversion based on self-perception, the premium required to observe a signal σ_a that is informative about the DM's attributes should be decreasing in the favorableness of the agent's prior distribution (see Fig. 5). Denote by $U(\Sigma_a)$ the expected utility of observing signal σ_a and, as in Proposition 4, let $E[u]$ denote the expected utility from not observing the signal. Then, the theory predicts that $E[u] - U(\Sigma_a)$ should be decreasing in the favorableness of the agent's prior distribution over her attributes.

The model. Since the theory presented by Josephs et al. (1992) considers situations with no ex-post actions, let B be a singleton. As in the other parts of this section, suppose that signals are purely informative.[33] For simplicity, consider either the forgetfulness model of Example 1 or the limited memory model of Example 2.

[31] Dunning (1995) also provides experimental evidence of this phenomenon in the domain of academic ability.

[32] See also Larrick (1993) for a similar discussion.

[33] The generalization to monetary lotteries is presented in Appendix A (see Remark 4).

In order to determine how the agent's attitude toward information is affected by her prior, let κ be a parameter that indexes her prior distribution. A higher parameter κ leads to a more favorable prior in the sense of first-order stochastic dominance:

$$\kappa' > \kappa \quad \Longrightarrow \quad F(\theta;\kappa') \leqslant F(\theta;\kappa), \tag{10}$$

for all $\theta \in \Theta$, with strict inequality for some θ.

Denote the gain from observing a high signal instead of a low signal by

$$\Delta u(\kappa, a) = \int u(\theta)\, dF(\theta|\sigma_a = H;\kappa) - \int u(\theta)\, dF(\theta|\sigma_a = L;\kappa).$$

The assumption that individuals with worse self-perceptions are more severely harmed by negative outcomes than those with better self-perceptions states can be stated as[34]:

Assumption 3 *(Decreasing differences).* $\Delta u(\kappa, a)$ is decreasing in κ for all $a \in A$.

Although Josephs et al. (1992) and Larrick (1993) do not discuss conditions under which Assumption 3 may fail, we should not expect it to hold globally (although there are many distributions and utility functions for which it does). In particular, Assumption 3 should probably fail for extremely self-confident individuals. For them, a high signal may be expected and would, therefore, not affect their self-views much. A bad signal, however, would be unexpected and severely affect their beliefs.[35]

Recall that $U(\Sigma_a)$ and $E[u]$ were defined as the expected utility of observing signal and the expected utility from not observing the signal, respectively. Then, the prediction of the theory of regret aversion based on self-perceptions can be stated as follows:

Conjecture 1. *(See Josephs et al., 1992.)* $E[u] - U(\Sigma_a)$ *is positive and decreasing in* κ*, for all* $a \in A$.

Since there are no ex-post actions, the only benefit from memory manipulation is the change in the DM's self-perceptions. As seen previously, the amount of memory manipulation is increasing in the self-image gain from observing a high signal $\Delta u(\kappa, a)$ in the models of Examples 1 and 2. Because, under Assumption 3, $\Delta u(\kappa, a)$ is decreasing in κ, we obtain the result stated in Conjecture 1:

Proposition 5 *(Regret aversion). Suppose Assumption 3 holds and consider either the forgetfulness model of Example 1 or the limited memory model of Example 2. For any* $a \in A$*, the premium required to observe the signal* σ_a *is decreasing* (*in the sense of strong set order*) *in the decision maker's prior over her attributes, indexed by parameter* κ.

Remark 2. Conjecture 1 is not true in the general model due to an indirect effect from m_L on m_H through the Bayesian posterior probability α. The probabilities of remembering good news and forgetting bad news are increasing in the gain from observing a high signal Δu holding beliefs fixed. However, in equilibrium, beliefs adjust to incorporate changes in these probabilities. Increasing the amount of memory manipulation reduces the probability of having observed a high signal given the recollection of no signal (i.e., α decreases). A smaller α reduces the payoff from forgetting a signal $u_\emptyset$, thereby increasing the incentive to remember good news and decreasing the incentive to forget bad news. In some cases, the decrease of the incentive to forget bad news may be large enough to dominate, leading to a decrease in total manipulation.[36] In those cases, the premium required to observe the signal σ_a will increase in the parameter κ.

In the limited memory model, the individual never manipulates her memory after bad news ($m_L = 0$), which prevents any indirect effect from m_L on m_H. Similarly, she never manipulates her memory after good news in the forgetfulness model ($m_H = 0$), preventing the indirect effect on m_H.

[34] Assumption 3 states that, for each $a \in A$ fixed, the function $\tilde{u}(\sigma_a, \kappa; a) \equiv \int u(\theta)\, dF(\theta|\sigma_a;\kappa)$ has decreasing differences in (σ, κ).

[35] For example, let $\theta = \Pr(\sigma_a = H|\theta)$ and $u(\theta, a) = \alpha + \beta\theta$. Then,

$$\Delta u(\kappa, a) = \beta \operatorname{Var}_\kappa(\theta)\left(\frac{1}{E_\kappa[\theta]} + \frac{1}{1 - E_\kappa[\theta]}\right),$$

where E_κ and $\operatorname{Var}_\kappa$ denote the expectation and variance operators. By stochastic dominance, $E_\kappa[\theta]$ is an increasing function (of κ). If κ does not affect the variance of θ, Assumption 3 holds if and only if $E_\kappa[\theta] < \frac{1}{2}$ (which is true if and only if κ is "not too large"). If, instead, we take θ to be uniformly distributed on $[0, \kappa]$, then Assumption 3 holds globally.

[36] For example, if $\psi_H(m_H) = C \times |m_H|$ for $C > 0$, the maximization of (4) will involve a bang-bang solution. Raising Δu may change the manipulation from $m_H = 0$ to $m_H = 1 - \eta_H$. However, if the date-1 self chooses to remember a high signal with probability 1, she will not engage in memory manipulation after observing a low signal. More precisely, note that this is a game of imperfect information, where the date-1 self has two "types": a type who chooses m_H after observing a high signal and a type who chooses m_L after a low signal. The low type's best response to full manipulation by the high type, $m_H = 1 - \eta_H$, is zero manipulation $m_L = 0$ (since $u_\emptyset = u_L$ in the maximization of (5)). Thus, if the cost of forgetting bad news is sufficiently high compared to C, the total cost of manipulation may decrease.

4.2. Prior-dependent attitude towards information

As shown in Proposition 4, the memory manipulation model explains why people may prefer to avoid information. Alternatively, Caplin and Leahy (2001) proposed the Psychological Expected Utility model (PEU), which generalizes expected utility to allow for different attitudes towards information.

Eliaz and Spiegler (2006) showed that PEU cannot account for certain situations in which the individual's preference for information varies with her prior distribution. In one example, they describe a patient who prefers more accurate medical tests when she is relatively certain of being healthy, yet she avoids these tests when she is relatively certain of being ill. In another, they describe a manager that asks for their employees' opinion only when she is sufficiently certain that the new information will not cause her to change her views much. They showed that such behaviors are inconsistent with PEU. As a result, Eliaz and Spiegler suggested that we should drop the assumption of Bayesian updating.

The next example illustrates that the memory manipulation model is consistent with the situations described by Eliaz and Spiegler. Therefore, unlike PEU, the self-deception model allows for prior-dependent attitudes towards information while retaining Bayesian updating.

Example 3. An individual must choose whether or not to take a medical exam. The exam has two possible diagnoses (signals): good or bad. Signals are ordered in terms of stochastic dominance and, therefore, a good signal increases the individual's expected health. For simplicity, we normalize the health type θ to be measured in terms of the probability of the good diagnosis q.

Let $h(q)$ denote the individual's utility conditional on the good diagnosis and let $\gamma(q)$ denote her utility conditional on the bad diagnosis. If the individual remembers receiving a bad diagnosis, she can undertake medical treatment, which increases her utility to $\tau(q) > \gamma(q)$. Consider the memory system from the Forgetfulness Model (Example 1) and suppose that memory manipulation is binary: $m_L \in \{-1, 0\}$, with $\psi_L(0) = 0$ and $\psi_L(-1) = \bar{\psi}$.

Let α denote the probability that self 2 assigns to a good diagnosis conditional on not remembering it ($\hat{\sigma} = \emptyset$). Forgetting a bad diagnosis is optimal for self 1 if the benefit from obtaining treatment is lower than the expected utility from not remembering it net of the manipulation costs:

$$\tau(q) \leqslant \alpha h(q) + (1-\alpha)\gamma(q) - \bar{\psi}.$$

Remembering a bad signal is optimal if the inequality is reversed.

In equilibrium, α is determined by Bayes' rule given self 1's strategy: $\alpha = 1$ if self 1 remembers a bad signal, whereas $\alpha = q$ if self 1 forgets it. Therefore, there is an equilibrium with no memory manipulation if

$$\tau(q) \geqslant h(q) - \bar{\psi},$$

and there is an equilibrium with manipulation if

$$\tau(q) \leqslant qh(q) + (1-q)\gamma(q) - \bar{\psi}.$$

In particular, there is an equilibrium with manipulation when $q = \frac{1}{2}$ and no manipulation when $q \approx 1$ if the following condition holds:

$$h(1) - \tau(1) < \bar{\psi} < \frac{h(\frac{1}{2}) + \gamma(\frac{1}{2})}{2} - \tau\left(\frac{1}{2}\right).$$

The first inequality states that a bad signal (followed by the appropriate treatment) does not harm the self-views of a sufficiently healthy individual much relative to the cost of memory manipulation. The second inequality concerns someone who is relatively uncertain about her health ($q = \frac{1}{2}$). It states that the harm of a bad signal on her self-views exceeds the manipulation costs.

Since there is no manipulation when the individual is relatively certain of being healthy, she prefers to take the exam. Moreover, because there is full manipulation when she is uncertain of her health, undertaking the exam does not affect actions. Then, because of the positive manipulation costs, she prefers not to taking the exam (Proposition 4).

With PEU, individuals have preferences defined over posterior beliefs. Therefore, they must be indifferent between signal structures that generate the same posteriors. Using the fact that PEU individuals with different priors who end up with the same posteriors must have the same utility, Eliaz and Spiegler show that the model cannot address several anomalies for which it was originally developed. Here, because signals can be manipulated at a positive cost, signal structures that generate the same posteriors through different amounts of belief distortion are ranked differently. Hence, manipulation cost breaks what would otherwise be an indifference in PEU, circumventing the results from Eliaz and Spiegler.[37]

[37] Eliaz and Spiegler also show that it is impossible to distinguish a PEU individual who prefers one state to another from an individual with the exact opposite preferences. In the present model, it is possible to identify preferences over attributes by eliciting the individual's recollections. From Propositions 1 and 2, the individual will recall more frequently signals that increase the posterior probability of preferred attributes and less frequently those that increase the probability of less preferred attributes. Therefore, a comparison of the recall probabilities of different signals identifies the ranking over attributes.

5. Repeated model

The previous sections considered an individual who observes an outcome once and makes inferences about her attributes based on her recollection of this outcome. In many situations, however, individuals participate in this process repeatedly. A professional investor, for example, is constantly deciding which investment to undertake and receives feedback about the success or failure of these investments quite frequently. It is often argued that the biases in decision-making that we observe in experimental settings would be attenuated as individuals gain experience. This section presents a repeated version of the general model described in Section 3 and shows that this is indeed the case in this model. More precisely, I show that the behavior of the DM converges to the one predicted by expected utility theory as the number of observed signals grows.

Consider a (frequently) repeated version of the general model described in Section 3. In each period $n \in \{1, 2, 3, \ldots, N\}$, the DM chooses an ex-ante action $a \in A$ and observes an independent draw of the outcome $\sigma_n \in \{H, L\}$. Each outcome is observed with probabilities $\Pr(\sigma_n = H|\theta, a)$ and $\Pr(\sigma_n = L|\theta, a)$, where θ is the agent's true attributes. The parameter θ is not known. Instead, the DM has beliefs about θ captured by a prior distribution $F(\theta)$ with full support on Θ.

I assume that a high outcome is more likely if the agent has higher attributes:

Assumption 4. $q_a(\theta) \equiv \Pr(\sigma = H|\theta, a)$ is a strictly increasing function of θ.

In particular, Assumption 4 implies that, as in the model from Section 3, high outcomes are more favorable than low outcomes in the sense of first-order stochastic dominance.

After observing $\sigma_n \in \{H, L\}$, the DM engages in memory manipulation m_H or m_L. She remembers the outcome with probability $\eta_{\sigma_n} + m_{\sigma_n}$, and forgets it with probability $1 - (\eta_{\sigma_n} + m_{\sigma_n})$. As before, we denote her recollection of the period-N signal by $\hat{\sigma}_n \in \{H, L, \varnothing\}$. Note that the DM can only manipulate the recollection of a signal in the period that the signal occurred. After the recollection has been registered into the agent's memory, she can no longer distort it.[38]

In the static framework, we assumed that a different self acted every time information was forgotten. In the repeated framework, there are different possible formulations of the decision problem under imperfect recall, depending on whether we allow memory manipulation to depend on forgotten information.

The most natural formulation assumes that a new incarnation of the stage-1 self is born every time a signal is forgotten. This captures the idea that, because the true signal was forgotten and only its recollection is known, the stage-1 self can only condition its actions on the recollections of the signal. Then, because both stage-1 and stage-2 selves only have access to recollections of signals but not the signals themselves, a history at time $n > 1$ is a sequence of recollections and actions:

$$h^n = \big((a_1, \ldots, a_{n-1}), (\hat{\sigma}_1, \ldots, \hat{\sigma}_{n-1}), (b_1, \ldots, b_{n-1})\big) \in \mathcal{H}^n,$$

where $\mathcal{H}^n \equiv A^{n-1} \times \{H, L, \varnothing\}^{n-1} \times B^{n-1}$ is the set of possible histories. A history in the initial period is the null set $h^1 = \varnothing$. The payoffs of each incarnation of the stage-1 self corresponds to the discounted sum of all incarnations of stage-1 selves.[39]

Alternatively, we could envision a single stage-1 self facing a single stage-2 self in all periods, where their payoffs correspond to the discounted sum of payoffs. In the language of reputation games, this corresponds to a situation where both selves are long-lived players. In this case, the stage-1 self would be able to condition her actions not only on previous recollections, but also on the previous signals. That is, self 1 would have access to the private history

$$h_1^n = \big((a_1, \ldots, a_{n-1}), (\sigma_1, \ldots, \sigma_{n-1}), (\hat{\sigma}_1, \ldots, \hat{\sigma}_{n-1}), (b_1, \ldots, b_{n-1})\big).$$

This formulation captures a compartmentalized decision maker, whose stage-1 self is able to remember all the history of observed signals but the state-2 self only has access to the manipulated recollections of such signals.

Although these two formulations typically give rise to different equilibria, the asymptotic results that will be obtained here hold in both. In fact, the proofs are identical in both cases. For clarity, however, I will focus on the formulation with public histories.

In each period, the stage-1 self chooses an ex-ante action $\phi_{a,n} : \mathcal{H}^n \to \Delta(A)$. Then, she observes an outcome $\sigma_a \in \{H, L\}$ and exerts memory manipulations $(m_{H,n}, m_{L,n}) : \mathcal{H}^n \times A \to [-\eta_H, 1 - \eta_H] \times [-\eta_L, 1 - \eta_L]$ to maximize the discounted sum of payoffs. The discount rate is $\delta \in [0, 1)$. Then, the stage-2 self applies Bayes' rule and chooses an ex-post action $\phi_b : \mathcal{H}^n \times A \times \{\varnothing, L, H\} \to \Delta(B)$. The prior distribution at the beginning of each period is the posterior distribution of skills conditional on the history and the equilibrium strategies. I am interested in the PBE of the game when N is large for a fixed $\delta \in [0, 1)$.

[38] This assumption captures the psychological finding that most information loss occurs soon after it is obtained. Nevertheless, it is clearly an extreme assumption. In general, forgetting rates seem to follow a power law (Anderson, 1995). Therefore, a large fraction of the information is lost right after learning, and over time, the rate of forgetting slows down.

[39] The formal definition of a PBE for the repeated game is analogous to Definition 1 and is presented in Appendix D (see Definition 2). Since the only decision of self 2 is to select an ex-post action conditional on the observed history, it is immaterial whether she is modeled as a long-run player or a sequence of short-run players who maximize the sum of discounted payoffs of all players in the sequence.

Let $\hat{\theta}_n(h^n) \equiv E_\mu[\theta|h^n]$ denote the conditional expectation of θ given history h^n according to the DM's beliefs. The first issue is whether the DM's conditional expectation converges to the true parameter θ. In other words, does the DM eventually learn her true attributes after observing a sufficiently large number of outcomes?

We will assume that *one* of the following conditions is satisfied:

Assumption 5a. $\eta_H > 0$.

Assumption 5b. For all a, b, σ, there exists $\bar{m}_{a,b,\sigma} > -\eta_L$ with $\psi_L(\bar{m}_{a,b,\sigma}) \geqslant \sup_\theta\{u(a,b,\theta,\sigma)\} - \inf_\theta\{u(a,b,\theta,\sigma)\}$.

Assumption 5a implies that the DM never forgets a good outcome with probability 1. Assumption 5a, which is satisfied, for example, if $\lim_{m_L \to -\eta_L} \psi(m_L) = +\infty$, implies that the DM never forgets a bad outcome with probability 1.[40] When either of them holds, there exists a probability bounded away from zero that the DM learns something about her skills. Then, her beliefs converge to the truth:

Proposition 6 *(Consistency). Let $N \to \infty$. Then, $\hat{\theta}_n \to \theta$ for almost all histories.*

Intuitively, because self 2 knows the equilibrium strategies, she knows the probability of each outcome conditional on the recollection. Therefore, she updates her beliefs correctly given the equilibrium amount of memory manipulation.

Proposition 6 shows that the DM eventually learns her true attributes θ. Then, the benefit of memory manipulation converges to zero, and therefore, memory manipulation converges to zero as the number of observed signals increases[41]:

Proposition 7 *(No manipulation in the long run). Let $N \to \infty$. Then, $m_H^* \to 0$ and $m_L^* \to 0$ for almost all histories.*

Suppose outcomes are purely informative signals. As in Section 4, omit the signal σ from the agent's utility function. Let a^O and b^O denote the set of actions that maximize the agent's utility when her attributes θ are known: $(a^O(\theta), b^O(\theta)) = \arg\max_{a\in A, b\in B} u(a,b,\theta)$. Proposition 7 implies that $(a_n, b_n) \to (\bar{a}, \bar{b}) \in (a^O, b^O)$ for almost all histories. Therefore, in the limit, the DM chooses the same actions as an expected utility maximizer who knows θ.[42]

Therefore, when signals are observed frequently enough, agents will eventually not engage in memory manipulation and their behavior will converge to the behavior of standard expected utility maximizers. This is consistent with the usual intuition that people do not exhibit ambiguity aversion over frequently observed events or that experts are subject to less biases than beginners (e.g. List, 2003; List and Haigh, 2005).[43] Proposition 7 also implies that individuals prefer to face decisions that require the same set of skills to decisions that require a different set of skills in each period when $\delta \approx 1$. Therefore, memory manipulation generates gains from specialization.

The results from this section rely only on two properties: the DM does not forget all signals with probability one, and beliefs are obtained by Bayes' rule on the set of histories that are consistent with the DM's behavior. Although these two properties are true in a PBE under Assumptions 4 and 5, they hold much more generally. In particular, they do not require each self to be optimizing (as long as the DM does not forget all signals with probability one).

Ex-ante actions and finitely many states. The key assumption for Propositions 6 and 7 is that all ex-ante actions $a \in A$ are informative about θ (Assumption 4). As the next example shows, convergence may not occur if the DM has the possibility of taking an action a whose outcomes are uninformative about θ.

Example 4 *(Non-convergence).* Consider a repeated version of the entrepreneurship example. For simplicity, assume that the DM is completely impatient ($\delta = 0$). There exists a strictly positive $\bar{\tau}$ such that the DM will not start a new company in every period and in all PBE if the expected monetary payoffs are smaller than $\bar{\tau}$. Thus, whenever $q\tau(H) + (1-q)\tau(L) < \bar{\tau}$, the DM never learns about θ.

Convergence under Naiveté. The convergence of beliefs relied on the fact that the DM takes into account the equilibrium amount of manipulation when interpreting her recollections. The online appendix considers individuals who are unaware of their manipulation and interpret recollections as if they had not engaged in memory manipulation. At any point in time, the individual is always optimistic about her abilities (in the sense of her beliefs first-order stochastically dominating the beliefs

[40] Assumptions 5a or 5b ensure identification. If $\eta_H = 0$ and $m_L(h^n) = \eta_L$, then $m_H(h^n) = 0$ for all h^n would imply that $\hat{\sigma}_n = \varnothing$. In this case, the posterior distribution would be equal to the prior and there would be no hope for it to converge to the true parameter.

[41] Appendix C generalizes the convergence results from Propositions 6 and 7 to any finite number of outcomes.

[42] In the model of monetary lotteries considered in Appendix A, the DM's utility from observing an additional signal converges to $qu(H,\theta) + (1-q)u(L,\theta)$, which is the same utility of an expected utility maximizer who knows her attributes θ.

[43] List (2003) also showed that experienced traders of sports paraphernalia show smaller endowment effects for everyday goods used in lab studies than novice traders. This result is also consistent with the model above if the ability to trade sports paraphernalia is correlated with the ability to trade other goods.

determined by Bayes' rule). Surprisingly, however, beliefs still converge to the true parameter θ under some regularity conditions. Therefore, the behavior of a naive individual also converges to the behavior of an expected utility maximizer who knows her attributes θ.

6. Conclusion

This paper proposed a model of choice under risk based on imperfect memory and self-deception. The model provides a unified explanation for a number of biases in decision-making. In particular, it provides a prior-dependent theory of regret aversion and allows for prior-dependent information attitudes.

The model can be enriched in several directions by incorporating strategic components. Principal-agent relationships seem like a natural application of the theory. Since the outcome of the relationship is typically informative about the agent's skill or knowledge, principals may prefer to offer contracts that do not completely reveal the outcome to the agent. Therefore, firms may prefer not to condition wages on economy-wide shocks. Similarly, CEOs may be "rewarded for luck."

Another direction is in the field of incomplete contracts. Contracts may be incomplete due to the contracting parties' preferences for avoiding information correlated with their skills or knowledge. However, because parties understand the consequences of contract forms and post-contractual decisions, the allocation of rights may matter for the outcomes. Therefore, the general framework proposed here may provide a behavioral model for a theory of ownership based on incomplete contracts. The model can also be embedded in a general equilibrium model. Since self-deception leads to endowment effects, the model may provide an explanation for the low volume of trades of uncertain assets occurring in equilibrium.

Finally, the model may lead to interesting predictions when θ is interpreted as a parameter of anticipatory utility as in Benabou (2013) or Brunnermeier and Parker (2005). Because anticipatory utility typically leads to a first-order gain from memory manipulation but only second-order costs through suboptimal decision-making, individuals will forget negative news and remember positive news with probability above their natural rates. For example, a model of portfolio allocation where signals σ are informative about the profitability of a risky asset may provide an explanation for why most investors hold extremely under-diversified portfolios and overinvest in stocks issued by the their employing firm.

Appendix A. Lotteries over money

> We propose that the consequences of each bet include, besides monetary payoffs, the credit or blame associated with the outcome. Psychic payoffs of satisfaction or embarrassment can result from self-evaluation or from an evaluation by others. (Heath and Tversky, 1991, pp. 7–8)

In Section 4, outcomes consisted of purely informative signals, which affected the DM's utility only through her beliefs about her own attributes θ. This appendix considers outcomes that affect the DM's utility not only by providing information about θ but also directly through monetary payments. The model leads to behavior similar to ambiguity aversion. The DM may reject gambles with small but positive expected value. Moreover, the model is consistent with the competence hypothesis proposed by Heath and Tversky (1991).

In order to focus on the implications of the model for the DM's preferences over monetary lotteries, I take B to be a singleton so that the agent does not take any ex-post actions. Each ex-ante action corresponds to a lottery that pays H_a with probability q_a and L_a with probability $1-q_a$. Therefore, the observable objects of choice are lotteries $(H_a, q_a; L_a, 1-q_a)$, where $a \in A$. The DM faces a two-stage problem, which can be solved by a backward induction procedure. In the stage 2, she plays the memory manipulation game for each fixed lottery $(H_a, q_a; L_a, 1-q_a)$. In stage 1, she picks the lottery that maximizes ex-ante expected utility given the equilibrium played in stage 2.

For notational simplicity, I omit a and b from the DM's von Neumann–Morgenstern utility function and focus on the second stage. Thus, the DM's utility function is $u(\theta, x)$, where $x \in \mathbb{R}$ denotes the amount of money that she has. If $H > L$, a high outcome not only provides favorable information about the agent's attributes θ but also pays a higher amount. This is the natural assumption since, in most cases, the outcome associated with higher monetary payments is also associated with better attributes. If $L > H$, a high outcome provides favorable information about θ but provides a lower payment. The results in this paper hold for any L and H.

I will refer to a lottery as an *objective lottery* if its payments are uninformative about the DM's characteristics θ (that is, $F(\theta|\sigma) = F(\theta)$ for all θ, σ). Because I want to focus on how self-image affects preferences through memory manipulation, I will assume that preferences over objective lotteries are independent of θ. That is, the utility function is additively separable between characteristics and money:

$$u(\theta, x) = v(\theta) + \tau(x),$$

for a strictly increasing function $v : \Theta \to \mathbb{R}$ and a function $\tau : \mathbb{R} \to \mathbb{R}$. Additive separability is equivalent to requiring that DM's with all possible characteristics θ rank lotteries over money in the same way.[44]

[44] The online appendix presents results for general utility functions.

Let $v_{\hat{\sigma}}$ denote the expected payoff from attributes conditional on recollection $\hat{\sigma} \in \{H, L, \varnothing\}$. Under additive separability, monetary payments can be factored out of self 1's memory manipulation choice. Given an outcome $\sigma \in \{H, L\}$, she maximizes:

$$(\eta_\sigma + m_\sigma)v_\sigma + (1 - \eta_\sigma - m_\sigma)v_\varnothing + \tau(\sigma) - \psi_\sigma(m_\sigma).$$

Therefore, self 1 chooses the same amount of memory manipulation as in the case of purely hedonic signals. Proposition 2 then implies that the DM will never choose to remember a low outcome or forget a high outcome. That is, in any PBE, $m_H^* > 0 \geqslant m_L^*$.

The DM's ex-ante expected utility (i.e., the utility at the first of the two-stage problem) is the sum of the expected payoff from attributes, the expected monetary payoffs, and the expected cost of memory manipulation:

$$U(\Sigma) = q\big[v_H + \tau(H) - \psi_H\big(m_H^*\big)\big] + (1-q)\big[v_L + \tau(L) - \psi_L\big(m_L^*\big)\big]. \tag{11}$$

Let U^I denote the utility of an objective lottery with the same distribution over monetary outcomes as the one. Then, the DM's ex-ante expected utility can be expressed as

$$U(\Sigma) = U^I - q\psi_H\big(m_H^*\big) - (1-q)\psi_L\big(m_L^*\big). \tag{12}$$

Because there are no ex-post actions, the objective value of information is zero. Therefore, the DM strictly prefers the lottery with uninformative outcomes to the one with informative outcomes.

Remark 3. Consider the entrepreneurship model described in Section 3.1. The DM will choose to become an entrepreneur if the expected monetary payoffs are greater than the expected costs of self-deception:

$$q\tau(H) + (1-q)\tau(L) \geqslant q\psi_H\big(m_H^*\big) + (1-q)\psi_L\big(m_L^*\big).$$

Baron (1999) presents evidence that individuals who become entrepreneurs find it easier to admit past mistakes to themselves. In a static environment, our model may easily lead to this result. Suppose, for example, that individuals have heterogeneous concerns for self-image or that homogeneous individuals play different equilibria of the game. Then those with a lower concern for self image or those who play equilibria with lower amounts of self-deception are precisely the ones who benefit the most from becoming entrepreneurs. Alternatively, Section 5 will establish that the expected cost of self-deception converges to zero as experience grows. Therefore, it could be the case that entrepreneurs were not different from other individuals ex-ante, but, as they have gained experience, their cost of admitting past mistakes decreased.

Remark 4. Under the additive separability assumption, it is immediate to extend Proposition 5 to monetary lotteries. Let κ index the DM's prior distribution as defined in Eq. (10). As in Assumption 3, assume that $\Delta v(\kappa)$ is decreasing in κ and consider either the forgetfulness model of Example 1 or the limited memory model of Example 2. Then, the premium $U^I - U(\Sigma)$ is positive and decreasing (in the sense of strong set order) in κ.

A.1. Probability weights

For each fixed equilibrium, there exists a weighting function $w : [0, 1] \to \mathbb{R}$ such that the utility from participating in the lottery is

$$U(\Sigma) = w(p) \times u_L + \big[1 - w(p)\big] \times u_H,$$

where $p \equiv \Pr(\sigma = L)$ and $u_s \equiv \int u(\theta, s)\, dF(\theta|\sigma = s)$. If $w(p) = p$, then the decision maker is an expected utility maximizer.

The DM prefers a lottery whose outcomes are uncorrelated with skills or competence θ to one whose outcomes are correlated with θ if and only if $w(p) > p$. Therefore, although the model does not feature ambiguity in the sense of an imprecise distribution of probabilities, it may be useful to think of the dependence of outcomes on skills or competence as "ambiguity" in this model. Accordingly, I will refer to an individual as "ambiguity averse" if $w(p) > p$.

Rearranging Eqs. (11) and (13), it follows that for each fixed PBE the DM's expected utility from the monetary lottery can be written as

$$U(\Sigma) = w(p)u_L + \big[1 - w(p)\big]u_H, \tag{13}$$

where

$$w(p) = p + \frac{(1-p)\psi_H(m_H^*) + p\psi_L(m_L^*)}{u_H - u_L}. \tag{14}$$

Note that $w(0) = 0$, $w(1) = 1$, and $w(p) > p$ for all $p \in (0, 1)$. Therefore, the DM always displays "ambiguity aversion" when the outcomes from the lottery are informative about her attributes.

Remark 5. When there are multiple equilibria, each equilibrium will be associated with a different weighting function. In the model of Example 1, the equilibrium is unique and, therefore, the probability weight is unique.

Remark 6. Eqs. (13) and (14) resemble a rank-dependent utility representation, except for the non-separability between probabilities and the utility u_σ. Since the departure from linear probability weights is caused by memory manipulation, individuals who engage in more memory manipulation have higher probability weights $w(p)$. Furthermore, because the amount of memory manipulation is increasing in the marginal utility from attributes, the deviation from linear weighting is itself a function of u_σ.

A.2. Discussion

The model presented here associates ambiguity aversion to the lottery outcomes' being informative about the DM's attributes. Several experimental papers have related ambiguity aversion with the lotteries' being influenced by an individual's skill or knowledge.[45] First, some experiments have contradicted the idea that ambiguity aversion is related to the imprecision of the probability distribution of the events as is usually argued. Budescu et al. (1988), for example, compared decisions based on numerically, graphically (the shaded area in a circle), and verbally expressed probabilities. Numerical descriptions of a probability are less vague than graphic descriptions which, in turn, are less vague than verbal descriptions. Thus, if agents had a preference for more precise distributions, they should rank events whose probabilities have a numerical description first, graphic descriptions second, and verbal descriptions last. However, unlike ambiguity aversion would predict, subjects were indifferent between these lotteries. Indeed, the authors could not reject that the agents behaved according to subjective expected utility theory and weighted events linearly.[46]

Heath and Tversky argued that people's preferences over ambiguous events arise from the anticipation of feeling knowledgeable or competent.[47] Their interpretation of the Ellsberg paradox is as follows:

> People do not like to bet on the unknown box, we suggest, because there is information, namely the proportion of red and green balls in the box, that is knowable in principle but unknown to them. The presence of such data makes people feel less knowledgeable and less competent and reduces the attractiveness of the corresponding bet. (Heath and Tversky, 1991, p. 8)

Fox and Tversky (1995, p. 585) proposed that ambiguity is caused by comparative ignorance. They have argued that "ambiguity aversion is produced by a comparison with less ambiguous events or with more knowledgeable individuals." As in Heath and Tversky's (1991) competence hypothesis, this "comparative ignorance hypothesis" states that ambiguity aversion is driven by the feeling of incompetence. Similarly, Goodie (2003) proposed the *perceived control* hypothesis, according to which ambiguity aversion is generated by an agent's belief that the distribution of outcomes is influenced by attributes such as knowledge or skill.[48]

A.3. Small-stakes risk aversion

Standard expected utility maximizers exhibit second-order risk aversion. An individual with second-order risk aversion always accepts small gambles with positive expected value. Then, if the agent has reasonable levels of risk aversion with respect to lotteries with small stakes, she must display unrealistically high levels of risk aversion with respect to lotteries with large stakes (Samuelson, 1963; Rabin, 2000). In this subsection, I show that the model allows us to reconcile risk aversion with respect to small lotteries with sensible levels of risk aversion with respect to large lotteries.

The certainty equivalent of a lottery is defined by the monetary amount $CE \in \mathbb{R}$ that makes the agent indifferent between participating in the lottery or receiving CE for sure. For the lottery considered in this section, the certainty equivalent CE solves

$$\int u(\theta, CE)\, dF(\theta) = qu_H(H) + (1-q)u_L(L) - q\psi_H\left(m_H^*\right) - (1-q)\psi_H\left(m_L^*\right). \tag{15}$$

The risk premium is defined as the difference between the expected payment and the certainty equivalent of the lottery: $\pi = qH + (1-q)L - CE$.

Let $s \in \{H, L\}$ be a binary random variable such that $E[s] = qH + (1-q)L = 0$. Consider a lottery that pays εs for $\varepsilon > 0$, and let $\pi(\varepsilon)$ denote the risk premium associated with this lottery. A decision maker has risk preferences of second order if $\lim_{\varepsilon \to 0_+} \pi(\varepsilon)/\varepsilon = 0$. She is first-order risk averse if $\lim_{\varepsilon \to 0_+} \pi(\varepsilon)/\varepsilon > 0$ is finite. She is zeroth-order risk averse if $\lim_{\varepsilon \to 0_+} \pi(\varepsilon)/\varepsilon = +\infty$. An individual with second-order risk aversion accepts small gambles with positive expected value.

[45] See Goodie and Young (2007) for a detailed discussion of this literature.
[46] See also Budescu et al. (2002).
[47] Appendix A.4 defines Heath and Tversky's "competence hypothesis" more precisely and also briefly reviews the empirical evidence related to it.
[48] There is a large experimental literature on the effect of perceived control on risk-taking (cf., Chau and Phillips, 1995, or Horswill and McKenna, 1999).

Segal and Spivak (1990) show that an individual with first-order risk aversion rejects small gambles as long as the positive expected value is sufficiently small. Clearly, an individual with zeroth-order risk aversion also rejects these gambles.

Note that the model of monetary lotteries converges to a model of purely hedonic signals as ε approaches zero. Then, as shown in Corollary 1, the DM demands a strictly positive participation premium in order to observe the signal. Hence, the certainty equivalent of the lottery converges to $CE(0) < 0$ and

$$\lim_{\varepsilon \to 0_+} \frac{\pi(\varepsilon)}{\varepsilon} = -\lim_{\varepsilon \to 0_+} \frac{CE(\varepsilon)}{\varepsilon} = +\infty.$$

Thus, the individual exhibits zeroth-order risk aversion. This result is established formally in the following proposition:

Proposition 8 *(Zeroth-order risk aversion). In any PBE, the DM exhibits zeroth-order risk aversion.*

Since outcomes are informative about the DM's attributes, the DM engages in memory manipulation. Therefore, even when the monetary payoffs converge to zero, she still demands a strictly positive risk premium. However, as shown in Corollary 1, when the expected monetary stakes are larger than the DM's participation premium, she will accept to participate in the lottery. Thus, as the following example shows, the DM may be risk averse over lotteries with small stakes without displaying an unreasonable degree of risk aversion over lotteries with large stakes:

Example 5. Suppose individuals reject a lottery that pays either -100 or 110 with equal probability for any wealth level W. Rabin (2000) shows that if these individuals are expected utility maximizers, they must also reject a lottery that pays -1000 and any arbitrarily large amount of money X with equal probability. The model in this paper is consistent with rejecting the first lottery and accepting the second for moderate values of X.

Suppose both lotteries have the same informational content about the DM's attributes θ. For simplicity, take the forgetfulness model of Example 1 with binary manipulation efforts $m_L \in \{-\frac{1}{2}, 0\}$ and let $\tau(x) = x$. Suppose that $\frac{1}{3}(v_H - v_L) > \psi_L(-\frac{1}{2})$ so that self 1 engages in memory manipulation: $m_L^* = -\frac{1}{2}$. Then, the DM rejects the first lottery for all wealth levels W if, for all W,

$$\frac{1}{2}(v_L + W - 100) + \frac{1}{2}(v_H + W + 110) - \frac{1}{2}\psi_L\left(-\frac{1}{2}\right) < \frac{1}{2}v_L + \frac{1}{2}v_H + W,$$

and she accepts the second lottery if

$$\frac{1}{2}(v_L + W - 1000) + \frac{1}{2}(v_H + W + X) - \frac{1}{2}\psi_L\left(-\frac{1}{2}\right) > \frac{1}{2}v_L + \frac{1}{2}v_H + W.$$

These conditions are satisfied if and only if

$$10 < \psi_L\left(-\frac{1}{2}\right) < \min\left\{\frac{1}{3}(v_H - v_L); X - 1000\right\}. \tag{16}$$

Therefore, when $10 < \psi_L(-\frac{1}{2}) < \frac{1}{3}(v_H - v_L)$, the DM accepts the first lottery and rejects the second lottery for all wealth levels W whenever $X > 1010$.

A.4. The competence hypothesis

Consider two lotteries with the same distribution over monetary outcomes. In the first lottery, outcomes are informative about the decision maker's skills or knowledge whereas in the second they are not. If the information about one's skills or knowledge is not useful (i.e., the objective value of information from the first lottery is zero) and the individual is an expected utility maximizer, she should be indifferent between these lotteries. Since one's attributes are ambiguous, an ambiguity averse individual should prefer the lottery whose outcomes are uninformative about her skills or knowledge.

Heath and Tversky (1991) have studied this choice in a series of experiments. In one experiment, for example, subjects were asked to answer several questions. Subjects also revealed (in an incentive-compatible way) their expected probability of answering the questions correctly. Afterwards, they chose between betting on their answers or participating in a lottery with the same expected probability of winning.

If decision makers are expected utility maximizers and the value of information is zero, they should be indifferent between these two lotteries. Since Prospect Theory does not distinguish between sources of uncertainty in the specification of probability weighting function, it also predicts that people should be indifferent between these two lotteries. Assuming that individuals choose each lottery with equal probability when indifferent, the proportion of individuals who bet on the knowledge-based lottery should be roughly constant at 50% when the value of information is zero. If the value of information is positive, individuals who behave according to either Expected Utility Theory or Prospect Theory should prefer the knowledge-based lottery. Hence, in this case, the proportion of individuals who bet on the knowledge-based lottery should be constant at 100%.

Heath and Tversky found a remarkably different pattern. The proportion of people who preferred to bet on the knowledge-based lottery instead of a knowledge-independent lottery *with the same expected probability of winning* was increasing in the judged probability.[49] In situations where the expected probability of winning was small, people preferred to bet on the knowledge-independent lottery. On the other hand, when the expected probability of winning was large, individuals preferred to bet on the knowledge-based lottery. Heath and Tversky have labeled this preference for the skill- or knowledge-dependent lottery only in contexts where individuals feel knowledgeable or competent the Competence Hypothesis. The following example shows that our model is consistent with the Competence Hypothesis:

Example 6. Consider the forgetfulness model of Example 1 and let memory manipulation be a binary variable $m_L \in \{-\frac{2}{3}, 0\}$, with $\psi_L(-\frac{2}{3}) = \frac{1}{4}$ and $\psi_L(0) = 0$. Suppose that the DM does not face ex-ante choices. However, in order to have a positive objective value of information, suppose that she chooses between b_L and b_H ex-post. Let $v_\sigma(b)$ denote the expected payoff from attributes conditional on outcome $\sigma \in \{H, L\}$ and take the following payoffs:

$$v_H(b_H) = 6, \qquad v_H(b_L) = 5, \qquad v_L(b_L) = 1, \qquad v_L(b_H) = 0, \qquad \tau(H) = 1, \qquad \tau(L) = 0.$$

In the online appendix, I show that the DM prefers the attribute-dependent lottery if $q > \frac{11}{23}$ and prefers the attribute-independent lottery if $q < \frac{11}{23}$.

Appendix B. Applications

This appendix presents two applications of the model. The first application provides a self-deception model of the endowment effect. The second application provides a self-deception rationale for people taking sunk investments into consideration when making decisions.

B.1. The endowment effect

An individual that satisfies the axioms of expected utility theory does not display a difference between the maximum willingness to pay for a good and the minimum compensation demanded to sell the same good (willingness to accept) when income effects are small. However, several empirical works have documented a discrepancy between these values. An individual tends to value one good more when the good becomes part of that person's endowment. Thaler (1980) labeled this phenomenon an "endowment effect."

Kahneman et al. (1990) argued that the endowment effect was caused by loss aversion.[50] This subsection proposes an alternative explanation for the endowment effect based on the idea that, in most markets, trading requires certain skills or knowledge. At the very least, potential buyers must form an expectation about how much the good is worth. In more complex markets, they must also estimate the future price of the good (which determines the opportunity cost of trading). As we have seen previously, individuals who care about self-image and are subject to imperfect memory demand a strictly positive premium in order to engage in an activity that reveals information about their skills. Therefore, they may prefer not to buy the good if the price is only slightly above its expected value.

The model is a special case of the general framework described in Section 3. Let $a \in \{Object, x\}$ denote the DM's choice between keeping an object and receiving x dollars. The object may have two possible values: L and H. Decision makers may be either buyers or sellers.

Fig. 6 depicts the Buyer's decision tree, which is identical to the one in the entrepreneur example (Section 3.1). Buyers do not know the value of the object. If they decide to keep the object, they observe its value. Then, they engage in memory manipulation m_L and m_H. As in Appendix A, the DM's preferences over skills θ and money x is represented by an additively separable function $v(\theta) + \tau(x)$.

Because sellers know the value of the object before making their decision, they do not learn any new information if they keep the object. Therefore, sellers never engage in memory manipulation. Buyers, on the other hand, make inferences about their skill or competence based on the value of the object, therefore, engage in memory manipulation. Therefore, buyers require greater compensation in order to keep the object:

Proposition 9 *(Endowment effect). Fix a PBE and let CE_B and CE_S denote the monetary payment x required by buyers and sellers in order to keep the object. Then, $CE_B < CE_S$.*

In this model, the endowment effect occurs due to the self-evaluation that follows acquiring an object. Since the value of the good is informative about the agent's skills or knowledge and therefore leads to costly self-deception, the DM requires

[49] A number of other experiments have confirmed the predictions of the competence hypothesis (cf., Keppe and Weber, 1995; Taylor, 1995; Kilka and Weber, 2000; Chow and Sarin, 2001; Fox and Weber, 2002; Kuehberger and Perner, 2003; Di Mauro, 2008).

[50] According to loss aversion, losses are weighed substantially more than gains. Then, the cost of losing a good is much higher than the benefit of winning a good.

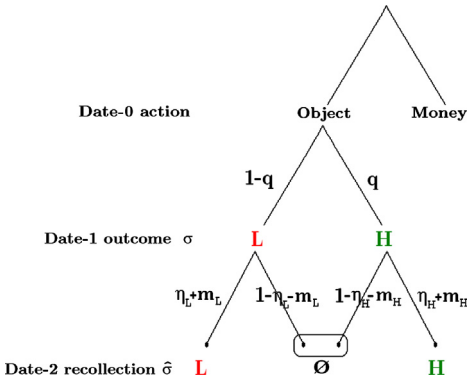

Fig. 6. Buyer's decision tree.

some strictly positive premium in order to buy the object. Therefore, unlike in the explanation based on loss aversion, it is not the *ownership* of an object but its *physical possession* that leads to endowment effects. A few recent papers have documented exactly this result. Reb and Connolly (2007), present two experiments in which they separate factual ownership from the physical possession of an object. Interestingly, they have found that physical possession increases one's valuation of the object whereas ownership does not. Wolf et al. (2008) find that the amount of time participants are exposed to coffee mugs increases the amount they are willing to pay for the mugs. Relatedly, Strahilevitz and Loewenstein (1998) show that the duration of ownership increases one's valuations for the good.

B.2. Sunk cost effects

> The consequences of any single decision (...) can have implications about the utility of previous choices as well as determine future events or outcomes. This means that sunk costs may not be sunk psychologically but may enter into future decisions. (Staw, 1981, p. 578)

Standard decision theory shows that only incremental costs and benefits should influence decisions. Historical costs, which have already been sunk, should be irrelevant. However, evidence suggests that people often take sunk costs into account when making decisions.[51] Genesove and Mayer (2001), for example, studied the Boston housing market. They have shown that when expected prices fall below the original purchase price, sellers set an asking price that exceeds the asking price of other sellers by between 25 and 35 percent of the difference.

This subsection shows that the self-deception model leads to sunk-cost effects. Psychologists have long argued that self-deception may be an important cause of why sunk costs affect choice. For example, Staw (1976) has shown that being personally responsible for an inefficient investment is an important factor in choosing to persist on it. Brockner et al. (1986) have documented that persisting on an inefficient allocation of resources is increased when subjects are told that outcomes reflected their "perceptual abilities and mathematical reasoning."[52]

Whether previous investments succeed or fail has important effects on the decision maker's self-views. Then, as the opening quote suggests, a past choice may be associated with not simply sunk *monetary* costs but also real *psychological* costs. Abandoning a project usually involves admitting that a wrong decision was made. Therefore, when a decision maker reconsiders a previous decision, she may learn something about her own ability. As shown in Section 4, the DM will prefer to avoid such information if the cost of making an uninformed decision is not high enough. But, in this case, some inefficient projects will not be terminated.

The model is a special case of the general framework described in Section 3. As in Appendix A, I assume the DM's utility function is additively separable over attributes θ and money x. For simplicity, I also assume that the DM is risk neutral so that $u(\theta, x) = v(\theta) + x$.

The timing of the model is presented in Fig. 7. First, the DM chooses whether to invest in a project that costs $K > 0$ and gives a random monetary payoff of π. Let $a_0 \in \{I, NI\}$ denote the investment choice, where $a_0 = I$ represents undertaking the investment and $a_0 = NI$ represents not undertaking it. After making the sunk investment, the DM can reevaluate the value of the project at zero cost. Let $a_1 = E$ denote the case where DM reevaluates the project and $a_1 = NE$ otherwise. Reevaluating the project leads to a (purely informative) signal $\sigma \in \{H, L\}$. A high signal is good news, both about the profitability of the project π and about the DM's skills θ.

[51] Sunk cost effects are also called "irrational escalation of commitment," the "entrapment effect," or "too much invested to quit."
[52] See Brockner (1992) for a review of the literature.

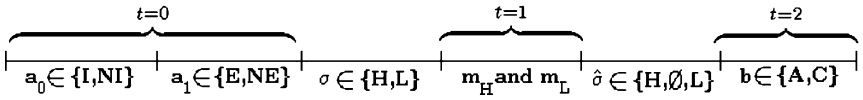

Fig. 7. Timing of the model.

After observing the signal σ, the DM may engage in memory manipulation m_L and m_H which leads to a recollection $\hat{\sigma} \in \{H, L, \varnothing\}$. Then, she chooses whether or not to abort the project. I write $b = A$ if the project is aborted and $b = C$ if it is continued. If the project is aborted, the DM obtains a monetary payoff of 0. If it is not aborted, the DM has an expected monetary payoff conditional on signal $\sigma \in \{H, L\}$ of π_σ.

I assume that the project is ex-ante efficient $E[\pi] > 0$.[53] As was shown in Proposition 4, the agent will prefer to observe the signal σ if the objective value of information, $V = -(1-q)\pi_L > 0$, is greater than the expected manipulation costs, $q\psi_H(m_H^*) + (1-q)\psi_L(m_L^*) > 0$. Hence, if the loss from not aborting an ex-post inefficient project is "not too large," the DM will prefer not to reevaluate the project:

Proposition 10 *(Sunk cost effect). There exist $\bar{\pi}_1 \leqslant \bar{\pi}_2 < 0$ such that*:

1. *the DM reevaluates the project in any PBE if $\pi_L \leqslant \bar{\pi}_1$,*
2. *the DM does not reevaluate the project in any PBE if $\pi_L \geqslant \bar{\pi}_2$, and*
3. *there exist PBE where the DM reevaluates the project and the DM doesn't reevaluate the project if $\bar{\pi}_1 < \pi_L < \bar{\pi}_2$.*

Since reevaluating one's previous decision is informative about the person's skills or knowledge, it leads to self-deception. Therefore, the DM will prefer not to reevaluate her initial choice if the monetary loss π_L from continuing an inefficient project is lower than the expected cost of memory manipulation. The key feature of the model is *not* the psychological cost from failure itself. The individual will eventually find out whether the project is successful or not. However, by not reevaluating a project, the individual avoids the psychological cost from *self-deception*.[54]

Appendix C. Finitely many outcomes

In the main text, each outcome σ_a could take two possible realizations: high or low. It is straightforward to generalize this framework to allow for any finite number of possible outcomes. Suppose that, given action $a \in A$, an outcome $s \in \{1, 2, \ldots, S_a\}$ is realized, $S_a \geqslant 2$. An outcome s is remembered with probability $\eta_{s,a} + m_s$, where $\eta_{s,a} \in [0, 1]$. Self 1 employs memory manipulation $m_s \in [-\eta_{s,a}, 1 - \eta_{s,a}]$, which costs $\psi_s(m_s) \geqslant 0$. Then, self 2 observes a recollection of the outcome, denoted by $\hat{\sigma}_a \in \{1, 2, \ldots, S_a, \varnothing\}$, and takes an action $b \in B$.

Preferences are represented by the von Neumann–Morgenstern utility function $u : \Theta \times A \times B \times \mathbb{R} \to \mathbb{R}$ which is strictly increasing in θ. When $u(\theta, a, b, x) = u(\theta, a, b, y)$ for all $x, y \in \mathbb{R}$, the model has *purely informative signals*. If signals are purely informative and A and B are singletons, we say that they have a *purely hedonic value*. When $u(\theta, a, b, x) \neq u(\theta, a, b, y)$ for some $x, y \in \mathbb{R}$ we say that the model has *monetary signals*.

The following example provides a more realistic formalization of the entrepreneurship model:

Entrepreneurship example. The performance $s \in \{S, F\}$ of an entrepreneur is affected by two independent variables: her attributes θ and the external conditions $r \in \{1, 2, 3, \ldots, R\}$. Attributes and external conditions are substitutes for the entrepreneur's performance. Therefore, given her performance $s \in \{S, F\}$, more favorable external conditions r provide bad news about the agent's attributes (in the sense of first-order stochastic dominance). The individual always recollects her performance s, but may manipulate her memory to change the rate at which she remembers the external conditions r.

This situation is modeled as follows. Let $\sigma \in \{S, F\} \times \{1, \ldots, R\}$ denote the outcome of the project. Outcomes are ordered by first-order stochastic dominance:

$$(S, 1) \succcurlyeq_{FOSD} (S, 2) \succcurlyeq_{FOSD} \cdots \succcurlyeq_{FOSD} (S, R) \succcurlyeq_{FOSD} (F, 1) \succcurlyeq_{FOSD} \cdots \succcurlyeq_{FOSD} (F, R),$$

where we write $x \succcurlyeq_{FOSD} y$ if x first-order stochastically dominates y. Given an outcome (s, r), self 1 chooses the probability at which the external conditions r are forgotten by exerting manipulation effort $m_{s,r}$. Then, self 2 applies Bayes' rule to the recollections $(s, \hat{r})$, where $\hat{r} \in \{r, \varnothing\}$. The agent's payoff net of manipulation costs given a recollection $(s, \hat{r})$ is

$$E\left[v(\theta, s) \middle| s, \hat{r}\right] + \tau(s),$$

where $s \in \{S, F\}$ and $\hat{r} \in \{1, 2, \ldots, R, \varnothing\}$. Fig. 8 presents the agent's decision tree.

[53] If the project is ex-ante inefficient, $E[\pi] \leqslant 0$, the DM would never invest.

[54] The argument above is related to agency explanations. For example, as argued by Li (2007), in environments with adverse selection, agents may prefer not to change their opinions if this publicly conceals bad news about their abilities. It is unclear, however, whether agency concerns would play an important role in contexts where the decisions are not publicly observed.

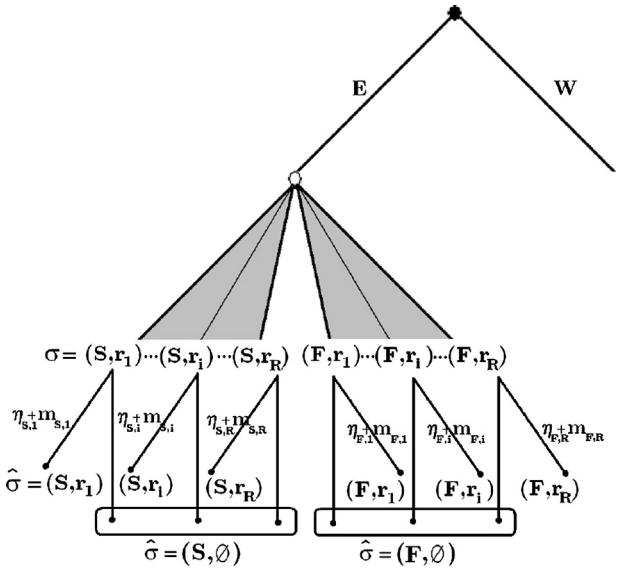

Fig. 8. Entrepreneur example.

It is straightforward to extend the results from the general framework to this environment. In particular, expected manipulation costs are always strictly positive. Therefore, the agent will require the expected monetary payoffs from starting a new company to be strictly higher than the payoff from the previous job to become an entrepreneur. Moreover, if all outcomes have the same natural rate of recollection (i.e., $\eta_{s,r} = \eta$ for all s, r), then $m_{s,1} \geqslant m_{s,2} \geqslant \cdots \geqslant m_{s,R}$ with at least one strict inequality, $s \in \{S, F\}$. Hence, the agent will remember negative external conditions more frequently than positive ones.

Most of the results presented in the text immediately generalize to this framework. The only result that is not immediate is the convergence result from Proposition 6. Let the vector p^a_θ denote the probabilities of each outcome s_a conditional on type θ. I will assume the following identification condition:

Identification Condition. (i) For all $\theta \in \Theta$ and all $\varepsilon > 0$, there exists a $\delta(\theta)$ such that

$$\left\| p^a_\theta - p^a_{\theta'} \right\| < \delta(\theta) \quad \Longrightarrow \quad \left|\theta - \theta'\right| < \varepsilon,$$

for all $a \in A$.

(ii) For all a and all $s \in \{1, \ldots, S_a\}$, there exists some $\tilde{m} > -\eta_s$ with $\psi_s(\tilde{m}) > \sup_{\theta,b,\sigma}\{u(\theta, a, b, \sigma)\}$.

The first assumption above states that types with similar conditional distributions should be close to each other. It is always satisfied in the two-outcome case considered in the text since $p^a_\theta = (q(\theta), 1 - q(\theta))$, and q is continuous and strictly increasing. The second assumption states that the DM is not able to forget outcomes with probability one. The following proposition generalizes Proposition 6 to the case where the DM takes actions ex-ante and there are finitely many outcomes:

Proposition 11. *Suppose the identification condition is satisfied. Let $N \to \infty$. Then, in any PBE, $\hat{\theta}_n \to \theta$ for almost all histories.*

Appendix D. Proofs

D.1. Proofs of propositions and lemmas

Proof of Proposition 1. The proof will use the following result:

Claim 1. $u_H(a, \phi_b^*(a, \varnothing), L) > u_L(a, \phi_b^*(a, L), L)$.

Proof. For notational simplicity, I will omit the term a from the terms in this proof. Therefore, I will write $u_s(\phi_b, \sigma)$ instead of $u_s(a, \phi_b, \sigma)$ and $\phi_b^*(\hat{\sigma})$ instead of $\phi_b^*(a, \hat{\sigma})$. I will also omit the manipulation efforts m_H and m_L from $\alpha(m_H, m_L)$ for notational clarity.

By revealed preference, $u_\varnothing(\phi_b^*(\varnothing), L) \geqslant u_\varnothing(\phi_b^*(L), L)$. From the definition of $u_\varnothing$, we have

$$\alpha^* u_H\big(\phi_b^*(\varnothing), L\big) + \big(1 - \alpha^*\big) u_L\big(\phi_b^*(\varnothing), L\big) \geqslant \alpha^* u_H\big(\phi_b^*(L), L\big) + \big(1 - \alpha^*\big) u_L\big(\phi_b^*(L), L\big),$$

which can be rearranged as:

$$\alpha^*\big[u_H\big(\phi_b^*(\varnothing), L\big) - u_H\big(\phi_b^*(L), L\big)\big] \geqslant \big(1-\alpha^*\big)\big[u_L\big(\phi_b^*(L), L\big) - u_L\big(\phi_b^*(\varnothing), L\big)\big].$$

Thus, by revealed preference, we have

$$u_H\big(\phi_b^*(\varnothing), L\big) - u_H\big(\phi_b^*(L), L\big) \geqslant \frac{1-\alpha^*}{\alpha^*}\big[u_L\big(\phi_b^*(L), L\big) - u_L\big(\phi_b^*(\varnothing), L\big)\big] \geqslant 0,$$

establishing that $u_H(\phi_b^*(\varnothing), L) \geqslant u_H(\phi_b^*(L), L)$. Since first-order stochastic dominance implies that $u_H(\phi_b^*(L), L) > u_L(\phi_b^*(L), L)$, it follows that $u_H(\phi_b^*(\varnothing), L) > u_L(\phi_b^*(L), L)$. □

Proof of the proposition. Because $\eta_H < 1$, the set of strictly positive efforts given a high signal $(0, 1-\eta_H]$ is non-empty. Recall that the PBE definition implies that self 1 takes the beliefs as well as the strategy of the other self as given when choosing how much memory manipulation to exert. More precisely, Condition (2) of Definition 1 states that $m_H^*(a)$ solves:

$$\begin{aligned}\max_{m_H}(\eta_H + m_H)\big\{&\big[1-\alpha_a\big(m_L^*(a), m_H^*(a)\big)\big]\big[u_H\big(a, \phi_b^*(a,\varnothing), H\big) - u_L\big(a, \phi_b^*(a,\varnothing), H\big)\big]\\ &+ u_H\big(a, \phi_b^*(a,H), H\big) - u_H\big(a, \phi_b^*(a,\varnothing), H\big)\big\} + u_\varnothing\big(a, \phi_b^*(a,\varnothing), H\big) - \psi_H(m_H). \qquad (4)\end{aligned}$$

Since the objective function above is a concave function of m_H, it suffices to show that its derivative evaluated at $m_H = 0$ is strictly positive:

$$u_H\big(a, \phi_b^*(a,H), H\big) - u_L\big(a, \phi_b^*(a,\varnothing), H\big) - \alpha_a\big(m_L^*(a), m_H^*(a)\big)\big[u_H\big(a, \phi_b^*(a,\varnothing), H\big) - u_L\big(a, \phi_b^*(a,\varnothing), H\big)\big] > 0, \qquad (17)$$

for all $m_L^*(a), m_H^*(a)$, where I have used the fact that $\psi_H'(0) = 0$.

Note that, by revealed preference, $u_H(a, \phi_b^*(a,H), H) \geqslant u_H(a, \phi_b^*(a,\varnothing), H)$. Hence,

$$\frac{u_H(a, \phi_b^*(a,H)) - u_L(a, \phi_b^*(a,\varnothing))}{u_H(a, \phi_b^*(a,\varnothing)) - u_L(a, \phi_b^*(a,\varnothing))} \geqslant 1.$$

Rearranging, we obtain:

$$\begin{aligned}&u_H\big(a, \phi_b^*(a,H), H\big) - u_L\big(a, \phi_b^*(a,\varnothing), H\big) - \alpha_a\big(m_L^*(a), m_H^*(a)\big)\big[u_H\big(a, \phi_b^*(a,\varnothing), H\big) - u_L\big(a, \phi_b^*(a,\varnothing), H\big)\big]\\ &\quad\geqslant u_H\big(a, \phi_b^*(a,H), H\big) - u_L\big(a, \phi_b^*(a,\varnothing), H\big) - \big[u_H\big(a, \phi_b^*(a,\varnothing), H\big) - u_L\big(a, \phi_b^*(a,\varnothing), H\big)\big] \geqslant 0,\end{aligned}$$

where the second inequality follows by revealed preference and first-order stochastic dominance. Hence, the expression on the left-hand side of (17) is non-negative. Suppose it is equal to zero. Then, by the previous inequality, it must be the case that $\alpha_a(m_L^*(a), m_H^*(a)) = 1$. But, since $\alpha_a(m_L^*(a), m_H^*(a)) = 1$ implies that $m_L^* = 1-\eta_L$, the Kuhn–Tucker condition of the maximization of (5) (which follows from Condition (2) of Definition 1 for signal $\sigma = L$), requires that

$$u_L\big(a, \phi_b^*(a,L), L\big) - u_H\big(a, \phi_b^*(a,\varnothing), L\big) \geqslant \psi_L'(1-\eta_L) \geqslant 0.$$

But, from Claim 1, $u_L(a, \phi_b^*(a,L), L) - u_H(a, \phi_b^*(a,\varnothing), L) < 0$, which contradicts the inequality above. Hence, we have established that $m_H^*(a) > 0$ for all $a \in A$. □

Proof of Proposition 2. Fix an action $a \in A$. I will omit a from the manipulation strategies for notational clarity. Moreover, since B is a singleton, I will omit b from the utility functions. From Proposition 1, it follows that $m_H^* > 0$. First, note that if $m_H^* = 1-\eta_H$, the ex-post self infers that a low signal was observed when she recollects $\hat{\sigma} = \emptyset$. Therefore, there is no benefit from memory manipulation after a low signal in that case: $m_L^* = 0$.

Next, we establish that $m_L^* \leqslant 0$. By the concavity of Eq. (5), it suffices to show that its derivative evaluated at $m_L = 0$ is weakly negative:

$$-\alpha\big(m_L^*, m_H^*\big)\big[u_H(a,L) - u_L(a,L)\big] - \underbrace{\psi_L'(0)}_{0} \leqslant 0,$$

which is true because $\alpha(m_L^*, m_H^*) \geqslant 0$ and $u_H(a,L) > u_L(a,L)$.

Let $u_H(a,H) - u_L(a,H) < \psi_H'(1-\eta_H)$ and, in order to obtain a contradiction, suppose that there exists a PBE with $m_L^* = 0$. Then, from Kuhn–Tucker's conditions of the maximization of (5),

$$\alpha\big(0, m_H^*\big)\big[u_H(a,L) - u_L(a,L)\big] = 0$$

for some m_H^* that maximizes (4) given $m_L^* = 0$. Because $u_H(a,L) > u_L(a,L)$, this is satisfied if and only if $\alpha(0, m_H^*) = 0$, which implies that $m_H^* = 1-\eta_H$. From Kuhn-Tucker's conditions of the maximization of (4), there exists a PBE with $m_H^* = 1-\eta_H$ if and only if

$$\big[u_H(a,H)-u_L(a,H)\big]\big[1-\alpha\big(m_L^*,m_H^*\big)\big]\geqslant\psi_H'(1-\eta_H),$$

for some m_L^* that maximizes (5). Substituting $\alpha(m_L^*(a),m_H^*(a))=0$, we obtain $u_H(a,H)-u_L(a,H)\geqslant\psi_H'(1-\eta_H)$, which is a contradiction. Therefore, $m_L^*<0$, which, as we have seen previously, requires $m_H^*<1-\eta_H^*$. □

Proof of Proposition 3. Fix $a\in A$. For notational clarity, I omit the term a from all the expressions in this proof. Fix a PBE. Define self-2's ex-post choice by $\phi_b(\hat{s},\alpha^*)$, $\hat{s}\in\{L,H,\varnothing\}$:

$$\big\{\phi_b(s)\big\}\in\arg\max_{\{\phi^b\}_{b\in B}}\sum_{b\in B}\phi^b u_H(b,s),\quad s=L,H,\quad\text{and}$$
$$\big\{\phi_b\big(\varnothing,\alpha^*\big)\big\}_{b\in B}\in\arg\max_{\{\phi^b\}_{b\in B}}\alpha^*\sum_{b\in B}\phi^b u_H(b,H)+\big(1-\alpha^*\big)\sum_{b\in B}\phi^b u_L(b,L)\tag{18}$$

(where I used the fact that $\phi_b(s)$ is not a function of α^* for $s=L,H$).

Define the function W_s as the expected utility of self 1 conditional on $\sigma=s$:

$$W_s\big(m_s,\alpha^*\big)=\max_{m_s}\Bigg\{(1-\eta_s-m_s)\Bigg[\alpha^*\sum_{b\in B}\phi_b\big(\varnothing,\alpha^*\big)u_H(b,H)+\big(1-\alpha^*\big)\sum_{b\in B}\phi_b\big(\varnothing,\alpha^*\big)u_L(b,L)\Bigg]$$
$$+(\eta_s+m_s)\sum_{b\in B}\phi_b(s)u_s(b,s)-\psi_s(m_s)\Bigg\}$$

Applying the envelope theorem to the maximization in (18), we obtain

$$\frac{d}{d\alpha^*}\Bigg(\alpha^*\sum_{b\in B}\phi_b\big(\varnothing,\alpha^*\big)u_H(b,H)+\big(1-\alpha^*\big)\sum_{b\in B}\phi_b\big(\varnothing,\alpha^*\big)u_L(b,L)\Bigg)$$
$$=\sum_{b\in B}\phi_b\big(\varnothing,\alpha^*\big)\big[u_H(b,H)-u_L(b,L)\big]$$
$$=u_H\big(\phi_b(\varnothing),H\big)-u_L\big(\phi_b(\varnothing),L\big).$$

Definition 1 implies that m_s^* maximizes W_s. Therefore, the envelope theorem gives

$$\frac{\partial W_s}{\partial\alpha^*}\bigg|_{m_s=m_s^*}=\big(1-\eta_s-m_s^*\big)\frac{d}{d\alpha^*}\Bigg(\alpha^*\sum_{b\in B}\phi_b\big(\varnothing,\alpha^*\big)u_H(b,H)+\big(1-\alpha^*\big)\sum_{b\in B}\phi_b\big(\varnothing,\alpha^*\big)u_L(b,L)\Bigg)$$
$$=\big(1-\eta_s-m_s^*\big)\big[u_H\big(\phi_b(\varnothing),H\big)-u_L\big(\phi_b(\varnothing),L\big)\big].$$

Note that the ex-ante expected utility is $qW_H+(1-q)W_L$ and that $\frac{\partial\alpha(m_H,m_L)}{\partial m_H}\leqslant 0\leqslant\frac{\partial\alpha(m_H,m_L)}{\partial m_L}$ with at least one inequality being strict. Therefore, we have

$$\frac{\partial\mathcal{U}(m_H,m_L,a,\{b_a\})}{\partial m_H}<0<\frac{\partial\mathcal{U}(m_H,m_L,a,\{b_a\})}{\partial m_L}.$$

The result then follows from the concavity of $\mathcal{U}$. □

Proof of Proposition 4. Follows from Eqs. (8) and (9). □

Proof of Proposition 5. First, consider the forgetfulness model. From Corollary 1,

$$E[u]-U(\Sigma_a)=(1-q_a)\psi_L\big(m_L^*(a)\big)\geqslant 0.$$

The amount of manipulation $|m_L^*|$ is increasing in Δu in the Forgetfulness Model with purely hedonic signals. Since, by Assumption 3, $\Delta u(\kappa,a)$ is decreasing in κ for any a, $E[u]-U(\Sigma_a)$ is decreasing in κ.

Consider the limited memory model. From Corollary 1,

$$E[u]-U(\Sigma_a)=q_a\psi_H\big(m_H^*(a)\big)\geqslant 0.$$

It can be shown that the set of equilibrium manipulations is increasing in the benefit of manipulation Δu (in the sense of strong set order). Assumption 3 and the monotonicity of $\psi_H(m_H)$ in $m_H\geqslant 0$ imply that the set of equilibrium premia $\{E[u]-U(\Sigma_a)\}$ is decreasing in κ (in the sense of strong set order). □

Definition of PBE for the repeated game. For notational clarity, I omit the arguments from the profiles of actions and manipulations.

Definition 2. A PBE of the repeated game is a strategy profile $(\phi_a^*, \phi_b^*, m_H^*, m_L^*)$ and posterior beliefs $\mu(.|.)$ such that:

1. $\phi_a^* \in \arg\max_{a\in A}\{E_{\hat{\sigma}_a}[E_\mu[u(a,\phi_b^*(a,\hat{\sigma}_a),\theta,\sigma_a)|\hat{\sigma}_a] + \delta V(a,\phi_b^*(a,\hat{\sigma}_a),\hat{\sigma}_a;h^n)|m_L^*(a),m_H^*(a)] - q\psi_H(m_H^*(a)) - (1-q)\times\psi_L(m_L^*(a))\}$;
2. $m_s^*(h^n,a)$ maximizes
$$(\eta_s + m_s)\{E_\mu[u(\theta,a,\phi_b^*(h^n,s),s)|h^n,s] + \delta V(a,\phi_b^*(a,\hat{\sigma}_a),s;h^n)\}$$
$$+ (1-\eta_s-m_s)\{E_\mu[u(\theta,\phi_b^*(h^n,\varnothing),\varnothing)|h^n,\varnothing] + \delta V(a,\phi_b^*(a,\varnothing),\varnothing;h^n)\} - \psi_s(m_s)$$
with respect to m_s, $s\in\{H,L\}$.
3. $\phi_b^*(a,\hat{\sigma}_a)\in\arg\max_{b\in B}\{E_\mu[u(a,b,\theta,s)+\delta V(a,b,\hat{s};h^n)|\hat{s}=\hat{\sigma}]\}$; for $s\in\{H,L\}$ and $\hat{s}\in\{H,L,\varnothing\}$.
4. $\mu(.|h)$ is obtained by Bayes' rule if $\Pr(h|m_{L,n}^*,m_{H,n}^*)>0$, for all $h\in\mathcal{H}^n\cup\{\varnothing,L,H\}\times\mathcal{H}^{n-1}$.
5. The continuation payoff V satisfies
$$V(a,b,\hat{s};h^n) = E_\mu\left\{\sum_{z=n+1}^{N}\delta^{z-n}[u(\theta,a,b,\sigma_z) - \Pr(\sigma_z=H)\psi_H(m_H^*(h^z)) - \Pr(\sigma_z=L)\psi_L(m_L^*(h^z))] \mid (\hat{s},b,h^{n-1})\right\},$$
for $(a,b,\hat{s};h^n)\in A\times B\times\{\varnothing,L,H\}\times\mathcal{H}^n$, and
$$V(\phi_a,\phi_b,\hat{s},h^n) = \sum_{a\in A}\sum_{b\in B}V(a,b,\hat{s};h^n)\phi_a\phi_b,$$
for $(\phi_a,\phi_b,\hat{s},h^n)\in\Delta(A)\times\Delta(B)\times\{\varnothing,L,H\}\times\mathcal{H}^n$.

Proof of Proposition 6. Although this is a special case of Proposition 11, I will present an alternative proof which emphasizes the intuition for the result. For notational simplicity, I will assume that there are no ex-post decisions (i.e., B is a singleton), although allowing for ex-post decisions would not affect any of the proofs.

If memory manipulation were constant, recollections would be i.i.d. and Doob's Consistency Theorem would imply that $\hat{\theta}_n(h^n)$ would converge to θ[55]:

Lemma 1. *Suppose $m_{H,n}(h^{n-1}) = \underline{m}_H$ and $m_{L,n}(h^{n-1}) = \underline{m}_L$ for all h^{n-1}, n and let $N\to\infty$, where either $\eta_H + \underline{m}_H > 0$ or $\underline{m}_L + \eta_L > 0$. Then $\hat{\theta}_n\to\theta$ for almost all histories.*

Proof. When $m_{H,n}(h^{n-1}) = \underline{m}_H$ and $m_{L,n}(h^{n-1}) = \underline{m}_L$ for all h^{n-1}, n, recollections are independent and identically distributed.

I claim that $\theta\mapsto\Pr_\theta$ is injective, that is, for all $\theta_1\neq\theta_2$, there exists $\hat{\sigma}\in\{\varnothing,L,H\}$ such that $\Pr_{\theta_1}(\hat{\sigma})\neq\Pr_{\theta_2}(\hat{\sigma})$. The probability of each recollection $\hat{\sigma}$ is:
$$\Pr_\theta(\hat{\sigma}=H) = q(\theta)(\eta_H+\underline{m}_H),$$
$$\Pr_\theta(\hat{\sigma}=L) = [1-q(\theta)](\eta_L+\underline{m}_L),$$
$$\Pr_\theta(\hat{\sigma}=\varnothing) = q(\theta)(1-\eta_H-\underline{m}_H) + [1-q(\theta)](1-\eta_L-\underline{m}_L),$$
where $q(\theta) := \Pr(\sigma=H|\theta)$ is strictly increasing. Thus, $\theta_1>\theta_2$ implies $\Pr_{\theta_1}(\hat{\sigma}=H)\geqslant\Pr_{\theta_2}(\hat{\sigma}=H)$ and $\Pr_{\theta_1}(\hat{\sigma}=L)\leqslant\Pr_{\theta_2}(\hat{\sigma}=L)$, with one of the inequalities being strict.

Since $\{\varnothing,L,H\}$ is finite, consistency follows from Theorem 1 in Freedman (1963). □

Although observed outcomes σ_n are i.i.d., recollections $\hat{\sigma}_n$ are generally neither independent nor identically distributed because of endogenous memory manipulation. Therefore, we cannot directly apply Doob's theorem to establish the consistency of beliefs.

In order to show that the individual eventually learns her true type, we will proceed in two steps. First, we show that beliefs conditional on each history on the equilibrium path under any manipulation strategy can be ranked according to first-order stochastic dominance. Thus, in any equilibrium, expected attributes are bounded below by their expectation in a

[55] The original theorem of Doob (1949) only established consistency for almost all θ. For finite-dimensional models such as the one here, consistency for all θ was established by Freedman (1963).

hypothetical scenario where the DM remembers high outcomes and forgets low outcomes the most. Conversely, expected attributes are bounded above by their expectation in a hypothetical scenario where the DM forgets high outcomes and remembers low outcomes the least. Second, we use the fact, that recollections are i.i.d. in these two hypothetical scenarios (because memory manipulation is constant), we can apply Lemma 1. Because expected attributes in any equilibrium is bounded by beliefs in these two scenarios, both of them converging to the true parameter, it follows that they must converge to the true parameter as well.

Lemma 2. *For any fixed history h^n, $F(\theta|h^n; m_H, m_L)$ is increasing in m_H and decreasing in m_L.*

Lemma 2 states that posterior beliefs θ can be ordered in terms of first-order stochastic dominance with respect to the amounts of memory manipulation. Therefore, conditional on reaching each history, the agent always prefers that she had forgotten high outcomes and remembered low outcomes.[56] Because the agent is ultimately concerned about the observed outcome σ_n and not its recollection $\hat{\sigma}_n$, $F(\theta|h^n; m_H, m_L)$ is not a function of m_H and m_L in all histories that do not contain any $\hat{\sigma}_n = \varnothing$. However, whenever the agent recollects $\hat{\sigma}_n = \varnothing$, she is always better off when she forgets high outcomes and remembers low outcomes (since it reduces the probability of arriving at $\hat{\sigma}_n = \varnothing$ after a low outcome $\sigma_n = L$). Hence, $(\theta|h^n; -\eta_H, 1-\eta_L)$ first-order stochastically dominates $(\theta|h^n; m_H, m_L)$ for all m_H, m_L.

Note that any distribution over attributes $\theta \in \Theta$ induces a distribution over signal structures $q(\theta) \in q(\Theta)$. To simplify the notation, we will work with the distribution over signal structures q rather than the distribution over types. This is without loss of generality here because, by assumption $q(\theta) \equiv \Pr(\sigma = H|\theta)$ is strictly increasing. With a slight abuse of notation, I will write $F(q|h^n)$ for the c.d.f. of q given history h^n.[57]

Note that actions $b_n \in B$ are functions of the sequence of recollections $\{\hat{\sigma}_1, \ldots, \hat{\sigma}_n\}$. Therefore, to simplify notation and with no loss of generality, I omit the actions $\{b_1, b_2, \ldots, b_n\}$ from histories. Thus, abusing notation again, I will refer to a history as a sequence of recollections $h^n = \{\hat{\sigma}_1, \ldots, \hat{\sigma}_n\}$ in all proofs.

Denote by $h^{n\backslash k}$ the history $\{\hat{\sigma}_1, \ldots, \hat{\sigma}_{k-1}, \hat{\sigma}_{k+1}, \ldots, \hat{\sigma}_n\}$. I will use the following result:

Claim 2. *For any history h^n, we have*:

$$F\left(q|h^{n\backslash k}, \hat{\sigma}_k = H\right) \leqslant F\left(q|h^{n\backslash k}, \hat{\sigma}_k = L\right).$$

This claim states that, for any history, a high signal is good news about q and a low signal is bad news about q in terms of first-order stochastic dominance. This proof uses tedious but long algebraic manipulations and is presented in the online appendix. □

We are now ready to prove the lemma:

Proof of Lemma 2. As shown previously, $F(x|h^n)$ is not a function of $m^*_{L,k}$ and $m^*_{H,k}$ for k such that $\hat{\sigma}_k \neq \varnothing$. Therefore, we only need to establish the results for k such that $\hat{\sigma}_k = \varnothing$.

Consider an arbitrary k such that $\hat{\sigma}_k = \varnothing$. Then, $F(x|h^n)$ is equal to

$$\frac{\begin{array}{l}(1-\eta_H-m^*_{H,k})\int_0^x q^{\#H+1}\times(1-q)^{\#L}\prod_{t\neq k:\sigma_t=\varnothing}[q(1-\eta_H-m^*_{H,t})+(1-q)(1-\eta_L-m^*_{L,t})]f(q)\,dq\\ \quad+(1-\eta_L-m^*_{L,k})\int_0^x q^{\#H}\times(1-q)^{\#L+1}\prod_{t\neq k:\sigma_t=\varnothing}[q(1-\eta_H-m^*_{H,t})+(1-q)(1-\eta_L-m^*_{L,t})]f(q)\,dq\end{array}}{\begin{array}{l}(1-\eta_H-m^*_{H,k})\int_0^1 q^{\#H+1}(1-q)^{\#L}\prod_{t\neq k:\sigma_t=\varnothing}[q(1-\eta_H-m^*_{H,t})+(1-q)(1-\eta_L-m^*_{L,t})]f(q)\,dq\\ \quad+(1-\eta_L-m^*_{L,k})\int_0^1 q^{\#H}\times(1-q)^{\#L+1}\prod_{t\neq k:\sigma_t=\varnothing}[q(1-\eta_H-m^*_{H,t})+(1-q)(1-\eta_L-m^*_{L,t})]f(q)\,dq\end{array}}.$$

With some algebraic manipulations, it follows that $\frac{dF}{dm_{H,k}}(x|h^n) > 0$ if and only if

$$\frac{\int_0^x q^{\#H}\times(1-q)^{\#L+1}\prod_{t\neq k:\sigma_t=\varnothing}[q(1-\eta_H-m^*_{H,t})+(1-q)(1-\eta_L-m^*_{L,t})]f(q)\,dq}{\int_0^1 q^{\#H}\times(1-q)^{\#L+1}\prod_{t\neq k:\sigma_t=\varnothing}[q(1-\eta_H-m^*_{H,t})+(1-q)(1-\eta_L-m^*_{L,t})]f(q)\,dq}$$
$$> \frac{\int_0^x q^{\#H+1}\times(1-q)^{\#L}\prod_{t\neq k:\sigma_t=\varnothing}[q(1-\eta_H-m^*_{H,t})+(1-q)(1-\eta_L-m^*_{L,t})]f(q)\,dq}{\int_0^1 q^{\#H+1}\times(1-q)^{\#L}\prod_{t\neq k:\sigma_t=\varnothing}[q(1-\eta_H-m^*_{H,t})+(1-q)(1-\eta_L-m^*_{L,t})]f(q)\,dq}.$$

[56] The first-order dominance (FOSD) is for fixed h^n. Since the probability of each history is itself a function of m_L and m_H, it does not follow that there is unconditional FOSD.

[57] If F_θ denotes the distribution over attributes, then the induced distribution over signals is determined by $F(q|h^n) = F_\theta(q^{-1}(q)|h^n)$. This is a standard maneuver in probability theory, known as the "change of variables" or the "push out" method. If q were not injective, the DM would only learn the true induced distribution over signals, but not the true type θ (i.e., parameter θ would be unidentified).

Note that the left-hand side is equal to $F(x|h^n, \hat{\sigma}_{n+1} = L)$, whereas the right-hand side is equal to $F(x|h^n, \hat{\sigma}_{n+1} = H)$. From the previous claim, it follows that $F(x|h^n, \hat{\sigma}_{n+1} = L) \geqslant F(x|h^n, \hat{\sigma}_{n+1} = H)$, which proves that the condition above is satisfied. Therefore, we have shown that $\frac{dF}{dm^*_{H,k}}(x|h^n) > 0$. The argument for $\frac{dF}{dm^*_{L,k}}(x|h^n) < 0$ is analogous. □

An implication of Lemma 2 is that:

$$E\big[\theta\big|h^n; \underline{m}_H, \underline{m}_L\big] \leqslant E\big[\theta\big|h^n; m_H, m_L\big] \leqslant E\big[\theta\big|h^n; -\eta_H, 1-\eta_L\big], \tag{19}$$

for all $m_H \geqslant \underline{m}_H$ and $m_L \leqslant \underline{m}_L$ and all histories h^n. Under Assumption 5, it is possible to select $\underline{m}_H$ and $\underline{m}_L$ such that: (i) in any PBE, $m_H(h^t) \geqslant \underline{m}_H$ and $m_L(h^t) \leqslant \underline{m}_L$ for all histories on the equilibrium path; and (ii) $E[\theta|h^n; \underline{m}_H, \underline{m}_L]$ converges to θ.[58] Moreover, because both extreme terms are not functions of endogenous variables, they provide bounds for the rate of convergence of expected types in any PBE. This establishes Proposition 6.

Proof of Proposition 7. In period N, conditions 1 and 2 from the definition of a PBE state that

$$m_{L,N}\big(L, h^{N-1}\big) \in \arg\max_{m_L}(\eta_L + m_L)\int u(\theta)\,dF\big(\theta\big|L, h^{N-1}\big) + (1-\eta_L - m_L)\int u(\theta)\,dF\big(\theta\big|\varnothing, h^{N-1}\big) - \psi_L(m_L),$$

and

$$m_{L,N}\big(H, h^{N-1}\big) \in \arg\max_{m_H}\Bigg\{(\eta_H + m_H)\Bigg[\int u(\theta)\,dF\big(\theta\big|H, h^{N-1}\big)\Bigg] + (1-\eta_H - m_H)\Bigg[\int u(\theta)\,dF\big(\theta\big|\varnothing, h^{N-1}\big)\Bigg] - \psi_H(m_H)\Bigg\}.$$

From Proposition 6, $\int u(\theta)\,dF(\theta|h^N)$ converges to $u(\theta)$ for almost all histories. But, when $\int u(\theta)\,dF(\theta|h^N) = u(\theta)$, it follows that $m_L(L, h^{N-1})$ maximizes

$$(\eta_L + m_L)u(\theta) + (1-\eta_L - m_L)u(\theta) - \psi_L(m_L) = u(\theta) - \psi_L(m_L),$$

which has a global maximum at $m_L = 0$. Hence, by continuity, $m_L(L, h^{N-1}) \to 0$ (a.s.). Similarly, when $\int u(\theta)\,dF(\theta|h^N) = u(\theta)$, $m_H(H, h^{n-1})$ maximizes $u(\theta) - \psi_H(m_H)$ so that $m_H(H, h^{n-1}) \to 0$ (a.s.). □

Proof of Proposition 8. This is a special case of Proposition 17, presented on Online Appendix C. □

Proof of Proposition 9. The result follows straight from the fact that CE_B and CE_S are implicitly defined by

$$u(CE_B) = q\big[u(H) - \psi_H\big(m^*_H\big)\big] + (1-q)\big[u(L) - \psi_L\big(m^*_L\big)\big],$$
$$u(CE_S) = qu(H) + (1-q)u(L),$$

and, $q\psi_H(m^*_H) + (1-q)\psi_L(m^*_L) > 0$. □

Proof of Proposition 10. The expected utility of self 1 if she chooses (I, NE) is $q(\theta_H + \pi_H) + (1-q)(\theta_L + \pi_L)$. Her expected utility if she chooses NI is $q\theta_H(1-q)\theta_L$. Because $q\pi_H + (1-q)\pi_L > 0$, NI is never chosen.

If self 1 chooses (I, E), she obtains:

$$q(\theta_H + \pi_H) + (1-q)\theta_L - q\psi_H\big(m^*_H\big) - (1-q)\psi_L\big(m^*_L\big).$$

Therefore, (I, E) is chosen if

$$q(\theta_H + \pi_H) + (1-q)\theta_L - q\psi_H\big(m^*_H\big) - (1-q)\psi_L\big(m^*_L\big) \geqslant q(\theta_H + \pi_H) + (1-q)(\theta_L + \pi_L).$$

Rearranging, we obtain

$$-(1-q)\pi_L \geqslant q\psi_H\big(m^*_H\big) + (1-q)\psi_L\big(m^*_L\big). \tag{20}$$

It can be shown that in any PBE the manipulation efforts $m^*_s(I, E)$ is not a function of π_L and π_H, $s \in \{H, L\}$. Then, the result follows immediately from Eq. (20). □

Proof of Proposition 11. The identification condition ensures that the probability of recalling a signal $\hat{s} = s$ is bounded away from zero. Then, we can apply the active supermartingale theorem from Fudenberg and Levine (1992, Theorem A.1), which implies that $\hat{\theta}_n \to \theta$ for almost all histories. □

[58] When $\eta_H > 0$, Lemma 1 implies that both extremes in the inequality (19) converge to θ for any $\underline{m}_L$. When there exists $\tilde{m} > -\eta_L$ with $\psi_L(\tilde{m}) \geqslant \sup_\theta\{u(a, \theta, \sigma)\} - \inf_\theta\{u(a, \theta, \sigma)\}$, Lemma 1 implies that both extremes in the inequality (19) converge to θ for $\underline{m}_L = \tilde{m}$ for any η_H.

D.2. Remarks and examples

Proof of the claim in Remark 1. Let $\hat{\mu}$ and μ denote the cumulative distribution functions of $\hat{\theta}_{\hat{\sigma}} \in \{\hat{\theta}_L, \hat{\theta}_\varnothing, \hat{\theta}_H\}$ and $\theta_\sigma \in \{\theta_L, \theta_H\}$, respectively. $\hat{\theta}_{\hat{\sigma}}$ second-order stochastically dominates θ_σ if, for any concave function $g : \Theta \to \mathbb{R}$,

$$\int g(\hat{\theta}_{\hat{\sigma}})\, d\mu(\hat{\theta}_{\hat{\sigma}}) \geqslant \int g(\theta_\sigma)\, d\mu(\theta_\sigma). \tag{21}$$

But

$$\int g(\theta_\sigma)\, d\mu(\theta_\sigma) = qg(\theta_H) + (1-q)g(\theta_L), \quad \text{and}$$

$$\int g(\hat{\theta}_{\hat{\sigma}})\, d\mu(\hat{\theta}_{\hat{\sigma}}) = q(m_H + \eta_H)g(\theta_H) + \big[q(1 - m_H - \eta_H) + (1-q)(1 - m_L - \eta_L)\big]g(\hat{\theta}_\varnothing) + (1-q)(\eta_L + m_L)g(\theta_L).$$

Substituting in inequality (21) and dividing by $q(1 - m_H - \eta_H) + (1-q)(1 - m_L - \eta_L)$, we obtain:

$$g\big(\alpha(m_L, m_H)\theta_H + \big[1 - \alpha(m_L, m_H)\big]\theta_L\big) \geqslant \alpha(m_L, m_H)g(\theta_H) + \big[1 - \alpha(m_L, m_H)\big]g(\theta_L),$$

which is true because g is concave. □

Forgetfulness Model with Purely Hedonic Signals. Existence follows from Proposition 1. For a fixed m_L^*, self 1 solves:

$$\max_{-1 \leqslant m_L \leqslant 0} u_L - m_L \alpha\big(m_L^*, 0\big)\Delta u - \psi_L(m_L).$$

The Kuhn–Tucker conditions are:

$$\alpha\big(m_L^*, 0\big)\Delta u \geqslant -\psi_L'(-1) \quad \Longrightarrow \quad m_L = -1,$$
$$\alpha\big(m_L^*, 0\big)\Delta u \leqslant 0 \quad \Longrightarrow \quad m_L = 0, \quad \text{and}$$
$$0 < \alpha\big(m_L^*, 0\big)\Delta u < -\psi_L'(-1) \quad \Longrightarrow \quad \alpha\big(m_L^*, 0\big)\Delta u = -\psi_L'(m_L).$$

In the PBE, $m_L = m_L^*$. Substituting $\alpha(m_L, 0) = \frac{q}{q-(1-q)m_L}$ and using the implicit function theorem, it follows that the unique PBE has manipulation efforts:

$$\frac{q}{q-(1-q)m_L^*}\Delta u = -\psi_L'\big(m_L^*\big) \quad \text{if } \Delta u < -\frac{\psi_L'(-1)}{q}, \quad \text{and}$$
$$m_L^* = -1 \quad \text{if } \Delta u \geqslant -\frac{\psi_L'(-1)}{q}. \tag{22}$$

The first comparative static follows by inspection. Let the cost of manipulation be $\psi_L(m_L, \kappa)$, where κ parametrizes the marginal cost of memory manipulation: $\frac{\partial^2 \psi}{\partial m_L \partial \kappa} < 0$. Therefore, higher κ's lead to a higher marginal cost of memory manipulation ($-m_L \geqslant 0$). Then, differentiation of the first equation from (22) and an inspection of the condition for boundary equilibria establishes the second and third. The first comparative statics claims. □

Example 6. It is helpful to separate the analysis in 2 cases: (i) $q \geqslant \frac{2}{5}$, and (ii) $q < \frac{2}{5}$. In case (i), self 2 chooses a high ex-post action, $b(\varnothing) = b_H$ when she expects self 1 to manipulate her memory, $m_L = -\frac{2}{3}$. In case (ii), she chooses a low ex-post action, $b(\varnothing) = b_L$ when she expects $m_L = -\frac{2}{3}$.

Case (i): The DM chooses to manipulate her memory if

$$\left(1 - \frac{2}{3}\right)\big[v_L(b_L) + \tau(L)\big] + \frac{2}{3}\big\{\alpha\big[v_H(b_L) + \tau(L)\big] + (1-\alpha)\big[v_L(b_L) + \tau(L)\big]\big\} - \psi_L\left(-\frac{2}{3}\right) > v_L(b_L) + \tau(L),$$

where α denotes the weight implies by Bayes' rule. This inequality is satisfied if and only if $\alpha > \frac{3}{32}$. Substituting the definition of α, we obtain $q > \frac{2}{31}$, which is satisfied since $q \geqslant \frac{2}{5} > \frac{2}{31}$.

The ex-ante expected utility from the signal is thus

$$q\big[v_H(b_H) + \tau(H)\big] + \frac{2}{3}(1-q)\big[v_L(b_H) + \tau(L)\big] + (1-q)\left(1 - \frac{2}{3}\right)\big[v_L(b_L) + \tau(L)\big] - (1-q)\psi_L\left(-\frac{2}{3}\right),$$

which is equal to $\frac{83q+1}{12}$. If the DM makes an uninformed decision, she obtains an ex-ante utility of

$$q\big[v_H(b_H) + \tau(H)\big] + (1-q)\big[v_L(b_H) + \tau(L)\big] \quad \text{if } b = b_H, \quad \text{and}$$
$$q\big[v_H(b_L) + \tau(H)\big] + (1-q)\big[v_L(b_L) + \tau(L)\big] \quad \text{if } b = b_L.$$

Thus, her utility is $7q$ if $q \geqslant \frac{1}{2}$, and $5q+1$ if $q < \frac{1}{2}$. The surplus from observing the signal is then

$$\frac{83q+1}{12} - \max\{7q, 5q+1\} = \begin{cases} \frac{1-q}{12} & \text{if } q \geqslant \frac{1}{2}, \\ \frac{23q-11}{12} & \text{if } q < \frac{1}{2}, \end{cases}$$

which is positive if and only if $q > \frac{11}{23}$.

Case (ii): In this case, because $b(\varnothing) = b_L$, the signal has no value. Therefore, the DM is always (weakly) better off by not observing the signal. In particular, since she exerts memory manipulation if $q \geqslant \frac{2}{31}$, the surplus is strictly negative for $q > \frac{2}{31}$ and it is equal to zero if $q < \frac{2}{31}$.

Appendix E. Supplementary material

Supplementary material related to this article can be found online at http://dx.doi.org/10.1016/j.geb.2013.11.013.

References

Akerlof, G.A., Dickens, W.T., 1982. The economic consequences of cognitive dissonance. Amer. Econ. Rev. 72 (3), 307–319.
Allport, G.W., 1943. The ego in contemporary psychology. Psychol. Rev. 50, 451–478.
Anderson, J.R., 1995. Learning and Memory: An Integrated Approach. John Wiley & Sons, New York.
Baron, R.A., 1999. Counterfactual thinking and venture formation: The potential effects of thinking about what might have been. J. Bus. Venturing 15, 79–91.
Bell, D.E., 1982. Regret in decision making under uncertainty. Operations Res. 30, 961–981.
Benabou, R., 2013. Groupthink: Collective delusions in organizations and markets. Rev. Econ. Stud. 80, 429–462.
Benabou, R., Tirole, J., 2002. Self-confidence and personal motivation. Quart. J. Econ. 117 (3), 871–915.
Benabou, R., Tirole, J., 2004. Willpower and personal rules. J. Polit. Economy 112 (4), 848–886.
Benabou, R., Tirole, J., 2006a. Belief in a Just World and redistributive politics. Quart. J. Econ. 121 (2), 699–746.
Benabou, R., Tirole, J., 2006b. Incentives and prosocial behavior. Amer. Econ. Rev. 96 (5), 1652–1678.
Benabou, R., Tirole, J., 2006c. Identity, dignity and taboos: beliefs as assets. Princeton University and Université de Toulouse. Mimeo.
Berglas, S., Baumeister, R., 1993. Your Own Worst Enemy: Understanding the Paradox of Self-Defeating Behavior. BasicBooks, New York.
Bernheim, B.D., Thomadsen, R., 2005. Memory and anticipation. Econ. J. 115, 271–304.
Bodner, R., Prelec, D., 2002. Self-signaling in a neo-Calvinist model of everyday decision making. In: Brocas, I., Carillo, J. (Eds.), Psychol. Econ., vol. II. Oxford University Press.
Brocas, I., Carrillo, J.D., 2008. The brain as a hierarchical organization. Amer. Econ. Rev. 98 (4), 1312–1346.
Brockner, J., 1992. The escalation of commitment to a failing course of action: toward theoretical progress. Acad. Manage. Rev. 17, 39–61.
Brockner, J., Houser, R., Birnbaum, G., Lloyd, K., Deitcher, J., Nathanson, S., Rubin, J.Z., 1986. Escalation of commitment to an ineffective course of action: The effect of feedback having negative implications for self-identity. Adm. Sci. Q. 31, 109–126.
Brunnermeier, M.K., Parker, J.A., 2005. Optimal expectations. Amer. Econ. Rev. 95, 1092–1118.
Budescu, D.V., Weinberg, S., Wallsten, T.S., 1988. Decisions based on numerically and verbally expressed uncertainties. J. Exp. Psychol. Hum. Percept. Perform. 14, 281–294.
Budescu, D.V., Kuhn, K.M., Kramer, K.M., Johnson, T.R., 2002. Modeling certainty equivalents for imprecise gambles. Org. Behav. Hum. Decis. Process. 88, 748–768.
Byrne, C.C., Kurland, J.A., 2001. Self-deception in an evolutionary game. J. Theor. Biol. 212 (4), 457–480.
Caplin, A., Eliaz, K., 2003. AIDS policy and psychology: A mechanism-design approach. RAND J. Econ. 34 (4), 631–646.
Caplin, A., Leahy, J., 2001. Psychological expected utility theory and anticipatory feelings. Quart. J. Econ. 116 (1), 55–79.
Caplin, A., Leahy, J., 2004. The supply of information by a concerned expert. Econ. J. 114, 487–505.
Compte, O., Postlewaite, Andrew, 2004. Confidence-enhanced performance. Amer. Econ. Rev. 94, 1536–1557.
Chau, A.W., Phillips, J.G., 1995. Effects of perceived control upon wagering and attributions in computer blackjack. J. Gen. Psychol. 122, 253–269.
Chow, C.C., Sarin, R.K., 2001. Comparative ignorance and the Ellsberg paradox. J. Risk Uncertainty 22, 129–139.
Dawson, E., Savitsky, K., Dunning, D., 2006. Don't tell me, I don't want to know: Understanding people's reluctance to obtain medical diagnostic information. J. Appl. Soc. Psychol. 36, 751–768.
Dewatripont, M., Jewitt, I., Tirole, J., 1999. The economics of career concerns, Part I: Comparing information structures. Rev. Econ. Stud. 66, 183–198.
Di Mauro, C., 2008. Uncertainty aversion vs. competence: An experimental market study. Theory Dec. 64, 301–331.
Doob, J.L., 1949. Application of the theory of martingales. In: Le Calcul des Probabilités et ses Applications. In: Colloques Internationaux du Centre National de la Recherche Scientifique, vol. 13. CNRS, Paris, pp. 23–27.
Dow, J., 1991. Search decisions with limited memory. Rev. Econ. Stud. 58, 1–14.
Dunning, D., 1995. Trait importance and modifiability as factors influencing self-assessment and self-enhancement motives. Pers. Soc. Psychol. Bull. 21, 1297–1306.
Edwards, W., 1968. Conservatism in human information processing. In: Kleinmutz, B. (Ed.), Formal Representation of Human Judgment. John Wiley and Sons, New York.
Edwards, W., 1982. Conservatism in human information processing. In: Kahneman, D., Slovic, P., Tversky, A. (Eds.), Judgment Under Uncertainty: Heuristics and Biases. Cambridge University Press.
Eliaz, K., Spiegler, R., 2006. Can anticipatory feelings explain anomalous choices of information sources?. Games Econ. Behav. 56, 87–104.
Festinger, L., 1957. A Theory of Cognitive Dissonance. Stanford University Press, Stanford, CA.
Fox, C.R., Tversky, A., 1995. Ambiguity aversion and comparative ignorance. Quart. J. Econ. 110, 585–603.
Fox, C.R., Weber, M., 2002. Ambiguity aversion, comparative ignorance, and decision context. Org. Behav. Hum. Decis. Process. 88, 476–498.
Freedman, D.A., 1963. On the asymptotic behavior of Bayes' estimates in the discrete case. Ann. Math. Stat. 34 (4), 1386–1403.
Fudenberg, D., Levine, D.K., 1992. Maintaining a reputation when strategies are imperfectly observed. Rev. Econ. Stud. 59 (3), 561–579.
Fudenberg, D., Levine, D.K., 2006. A dual self model of impulse control. Amer. Econ. Rev. 96, 1449–1476.
Fudenberg, D., Tirole, J., 1991. Game Theory. MIT Press, Cambridge, MA.
Genesove, D., Mayer, C., 2001. Loss aversion and seller behavior: Evidence from the housing market. Quart. J. Econ. 116, 1233–1260.
Gollwitzer, P.M., Earle, W.B., Stephan, W.G., 1982. Affect as a determinant of egotism: Residual excitation and performance attributions. J. Pers. Soc. Psychol. 43, 702–709.

Goodie, A.S., 2003. The effects of control on betting: Paradoxical betting on items of high confidence with low value. J. Exper. Psychol., Learn., Mem., Cogn. 29, 598–610.
Goodie, A.S., Young, D.L., 2007. The skill element in decision making under uncertainty: Control or competence?. Judgm. Decis. Mak. 2, 189–203.
Greenwald, A.G., 1980. The totalitarian ego: Fabrication and revision of personal history. Amer. Psychol. 35, 603–618.
Heath, C., Tversky, A., 1991. Preference and belief: Ambiguity and competence in choice under uncertainty. J. Risk Uncertainty 4 (1), 5–28.
Hellman, M.E., Cover, T.M., 1973. A review of recent results on learning with finite memory. Probl. Control Inf. Theory, 221–227.
Hilgard, E.R., 1949. Human motives and the concept of self. Amer. Psychol. 4, 374–382.
Hirshleifer, D., Welch, I., 2002. An economic approach to the psychology of change: Amnesia, inertia, and impulsiveness. J. Econ. Manag. Strategy 11, 379–421.
Holmstrom, B., 1999. Managerial incentive problems: A dynamic perspective. Rev. Econ. Stud. 66, 169–182.
Horswill, M.S., McKenna, F.P., 1999. The effect of perceived control on risk taking. J. Appl. Soc. Psychol. 29, 378–392.
Hvide, H., 2002. Pragmatic beliefs and overconfidence. J. Econ. Behav. Organ. 2002 (48), 15–28.
Johnston, W.A., 1967. Individual performance and self-evaluation in a simulated team. Organ. Behav. Hum. Perform. 2, 309–328.
Josephs, R.A., Larrick, R.P., Steele, C.M., Nisbett, R.E., 1992. Protecting the self from the negative consequences of risky decisions. J. Pers. Soc. Psychol. 62, 26–37.
Kahneman, D., Knetsch, J.L., Thaler, R.H., 1990. Experimental tests of the endowment effect and the coase theorem. J. Polit. Economy 98 (6), 1325–1348.
Keppe, H., Weber, M., 1995. Judged knowledge and ambiguity aversion. Theory Dec. 39, 51–77.
Kihlstrom, J.F., Beer, J.S., Klein, S.B., 2003. Self and identity as memory. In: Leary, M.R., Tangney, J.P. (Eds.), Handbook of Self and Identity. Guilford Press, New York, NY.
Kilka, M., Weber, M., 2000. Home bias in international stock return expectation. J. Psychol. Financ. Mark. 1, 176–192.
Kopczuk, W., Slemrod, J., 2005. Denial of death and economic behavior. Adv. Theor. Econ. 5 (1), 5.
Korner, I., 1950. Experimental Investigation of Some Aspects of the Problem of Repression: Repressive Forgetting. Contributions to Education, vol. 970. Bureau of Publications, Teachers' College, Columbia University, New York, NY.
Kőszegi, B., 2003. Health anxiety and patient behavior. J. Health Econ. 22, 1073–1084.
Kőszegi, B., 2006. Ego utility, overconfidence, and task choice. J. Eur. Econ. Assoc. 4, 673–707.
Kuehberger, A., Perner, J., 2003. The role of competition and knowledge in the Ellsberg task. J. Behav. Decis. Mak. 16, 181–191.
Kunda, Z., Sanitioso, R., 1989. Motivated changes in the self-concept. J. Pers. Soc. Psychol. 61, 884–897.
Larrick, R.P., 1993. Motivational factors in decision theories: The role of self-protection. Psychol. Bull. 113, 440–450.
Li, W., 2007. Changing one's mind when the facts change: Incentives of experts and the design of reporting protocols. Rev. Econ. Stud. 74, 1175–1194.
List, J., 2003. Does market experience eliminate market anomalies?. Quart. J. Econ. 118 (1), 41–71.
List, J., Haigh, M.S., 2005. A simple test of expected utility theory using professional traders. Proc. Natl. Acad. Sci. 102, 945–948.
Loomes, G., Sugden, R., 1982. Regret theory: An alternative theory of rational choice under uncertainty. Econ. J. 92, 805–824.
Lowenstein, G.A., 1987. Anticipation and the valuation of delayed consumption. Econ. J. 97, 666–684.
Mather, M., Shafir, E., Johnson, M.K., 2003. Memory and remembering chosen and assigned options. Mem. Cogn. 31 (3), 422–433.
Mijovic-Prelec, D., Prelec, D., 2010. Self-deception as self-signaling: A model and experimental evidence. Philos. Trans. R. Soc. Biol. Sci. 365, 227–240.
Mill, J., 1829. Analysis of the Phenomena of the Human Mind. Baldwin and Craddock, London.
Mullainathan, S., 2002. A memory-based model of bounded rationality. Quart. J. Econ. 117 (3), 735–774.
Philipson, T., Posner, R., 1995. A theoretical and empirical investigation of the effects of public health subsidies for STD testing. Quart. J. Econ. 110 (2), 445–474.
Piccione, M., Rubinstein, A., 1997. On the interpretation of decision problems with imperfect recall. Games Econ. Behav. 20, 3–24.
Prelec, D., 2008. Self-delusion: A neuroeconomics model and fMRI evidence. Seminar on the Foundations of Human Social Behavior. University of Zurich.
Quattrone, G.A., Tversky, A., 1984. Causal versus diagnostic contingencies: On self-deception and the Voter's illusion. J. Pers. Soc. Psychol. 46, 237–248.
Rabin, M., 1994. Cognitive dissonance and social change. J. Econ. Behav. Organ. 23, 177–194.
Rabin, M., 2000. Risk aversion and expected-utility theory: A calibration exercise. Econometrica 68 (5), 1281–1292.
Rapaport, D., 1961. Emotions and Memory. Science Editions, New York.
Reb, J., Connolly, T., 2007. Possession, feelings of ownership and the endowment effect. Judgm. Decis. Mak. 2, 107–114.
Robbins, H., 1956. A sequential decision problem with a finite memory. Proc. Natl. Acad. Sci. 42, 920–923.
Samuelson, P.A., 1963. Risk and uncertainty: A fallacy of large numbers. Scientia 57 (6), 1–6.
Schelling, T., 1985. The mind as a consuming organ. In: Elster, J. (Ed.), The Multiple Self. Cambridge University Press, New York.
Segal, U., Spivak, A., 1990. First order versus second order risk aversion. J. Econ. Theory 51, 111–125.
Sedikides, C., Green, J.D., Pinter, B.T., 2004. Self-protective memory. In: Beike, D., Lampinen, J., Behrend, D. (Eds.), The Self and Memory. Psychology Press, Philadelphia.
Staw, B.M., 1976. Knee deep in the Big Muddy. Org. Behav. Hum. Decis. Process. 35, 124–140.
Staw, B.M., 1981. The escalation of commitment to a course of action. Acad. Manage. Rev. 6, 577–587.
Steele, C.M., 1988. The psychology of self-affirmation: Sustaining the integrity of the self. In: Berkowitz, L. (Ed.), Adv. Exp. Soc. Psychol., vol. 21. Academic Press, New York, pp. 261–302.
Strahilevitz, A.M., Loewenstein, G., 1998. The effect of ownership history on the valuation of objects. J. Cons. Res. 25, 276–289.
Taylor, K., 1995. Testing credit and blame attributions as explanation for choices under ambiguity. Org. Behav. Hum. Decis. Process. 64, 128–137.
Taylor, S.E., Gollwitzer, P.M., 1995. The effects of mindset on positive illusions. J. Pers. Soc. Psychol. 69, 213–226.
Thaler, R.H., 1980. Toward a positive theory of consumer choice. J. Econ. Behav. Organ. 1, 39–60.
Thaler, R.H., Shefrin, H.M., 1981. An economic theory of self-control. J. Polit. Economy 89, 392–406.
Trivers, R., 2000. The elements of a scientific theory of self-deception. Ann. N.Y. Acad. Sci. 907, 114–131.
Wilson, A., 2003. Bounded memory and biases in information processing. Princeton University. Mimeo.
Wolf, J.R., Arkes, H.R., Muhanna, W.A., 2008. The power of touch: An examination of the effect of duration of physical contact on the valuation of objects. Judgm. Decis. Mak. 3 (6), 476–482.
Zuckerman, M., 1979. Attibution of success and failure revisited, or: The motivational bias is alive and well in attribution theory. J. Pers. 47, 245–287.