Selective Memory of a Psychological Agent

Jeanne Hagenbach, Frédéric Koessler

Selective Memory of a Psychological Agent *

Jeanne Hagenbach† Frédéric Koessler‡

November 22, 2021

Abstract

We consider a single psychological agent whose utility depends on his action, the state of the world, and the belief he holds about that state. The agent is initially informed about the state and decides whether to memorize it, otherwise he has no recall. We model the memorization process by a multi-self game in which the privately-informed first self voluntarily discloses information to the second self, who has identical preferences and acts upon the disclosed information. We show that, for broad categories of psychological utility functions, there exists an equilibrium in which every state is voluntarily memorized. In contrast, if there are exogenous failures in the memorization process, the agent always memorizes states selectively. In this case, we characterize the partially informative equilibria for common classes of psychological utilities.

Keywords: Multi-self games; disclosure games; imperfect recall; selective memory; motivated beliefs; psychological games; anticipatory utility.

JEL Classification: C72; D82

*We thank Martin Dufwenberg, Toomas Hinnosaar, Juan Ivars, Philippe Jehiel, Yves Le Yaouanq, Elliot Lipnowski, Peter Schwardmann and Ina Taneva for helpful comments. We also thank three anonymous referees and the associate editor for constructive advice. We are grateful to the (virtual) seminar participants at Sciences Po, Paris School of Economics, University of Munich, University of Nottingham, University of East Anglia, University of Arizona, WZB Berlin, Ben-Gurion University, LUISS Guido Carli University, University of Edinburgh, Dauphine University, Carnegie Mellon University and the Virtual Seminar in Economic Theory.

†Sciences Po Paris — CNRS. Jeanne Hagenbach thanks the European Research Council (grant 850996 – MOREV) for financial support. She is grateful to the WZB for having welcomed her in 2021 as the K.W.Deutsch Professor.

‡Paris School of Economics – CNRS. Frédéric Koessler acknowledges the support of the ANR (Investissements d'Avenir, ANR-17-EURE-001, and StratCom, ANR-19-CE26-0010-01).

1 Introduction

When evolving in an uncertain environment, individuals form beliefs about it with the primary objective to take the most appropriate decisions. When considering only this *instrumental* value of beliefs, and if the agent's preferences are time-consistent, accuracy is beneficial and more information is welcome. However, research in psychology and in behavioral economics suggests that beliefs formation is not uniquely driven by the desire for accuracy because individuals sometimes attribute an *intrinsic* value to what they believe. They may for example prefer holding a less true but more optimistic view of themselves or the world they live in. In this paper, we consider a single "psychological" agent whose utility is *directly* affected by his beliefs about the state of the world and, as is more standard, by his action and the true state.[1]

We develop a general game-theoretical framework to understand what information a psychological agent voluntarily keeps in mind, given that accurate beliefs allow him to take optimal actions but may be detrimental for his well-being. Precisely, we consider a multi-self disclosure game in which the agent is initially informed about the state of the world and selectively decides which states to disclose to his later self. We interpret Self 1's act of disclosing information to Self 2 as the act of memorizing the state, without which, Self 2 has no recall and is uninformed.[2] After seeing what Self 1 memorized, Self 2 forms a posterior belief about the state and takes an action. The two selves share a common utility function. Our objective is to understand how the agent's preferences over posterior beliefs and the form of his psychological utility affect how much and what kind of information is voluntarily memorized.

First, we show that for broad categories of psychological utilities, there always exists an equilibrium in which the agent voluntarily memorizes everything. In such an equilibrium, even if the agent intrinsically cares about the beliefs he finally holds and has the possibility to influence these beliefs by forgetting, he does not manipulate his beliefs. The intuition is as follows. In our game, the equilibrium beliefs are determined by Bayesian updating. Hence, the agent does not only learn the memorized information but also can make inferences from the absence of memory. In particular, he can skeptically attribute no memory to bad news. This skepticism sustains full memorization in equilibrium for various classes of psychological utilities such as state-independent utility, separable utility or anticipatory utility. An example of state-independent utility can be found in Hestermann, Le Yaouanq, and Treich (2020) in which the agent's well-being is directly affected by his meat consumption and by his beliefs about the level of animals suffering, but not by the true level. The psychological utility is separable when it is the sum of a standard material utility (that depends on the action and state) and of a utility function that depends only on the posterior beliefs about the state. An example of such a form can be found in Bénabou, Falk, and Tirole (2019) in which the optimal action of the

[1] An alternative term for the utility we consider is "belief-based". We use "psychological" in reference to the literature on psychological game theory pioneered by Geanakoplos, Pearce, and Stacchetti (1989).

[2] While the process of memorizing information can be a cognitive process, it can also take the form of tangible actions that can well be interpreted as disclosure to a future self: make reminding notes, repeat statements aloud, put some event in context, let some papers in evidence, etc.

agent depends on his moral type. The agent additionally cares about the perception of his own morality. A third class of psychological utility is the more common anticipatory utility: the agent is affected both by his standard utility and by the expectation of this utility in the future.

Next, we extend the model by assuming that, with a positive probability, the first self may be unable to memorize the state for exogenous reasons. In that case, we show that, if the marginal effect of the agent's belief on his utility is not negligible, there is no equilibrium in which the agent voluntarily memorizes every state when able to do so. To understand this impossibility result, note that, with memorization failures, even a sophisticated agent cannot be fully skeptical: the agent must attribute, at least partially, the lack of memory to his inability to memorize in general. It follows that there is always some type of Self 1 who, by forgetting, can manipulate Self 2's beliefs in a beneficial way. This beliefs manipulation has a first-order positive effect on his utility but a second-order negative effect of taking a suboptimal action. In short, we show that the presence of memorization failures, which is hardly debatable, offers the agent some wiggle room to manipulate his beliefs and that he always steps into it.

To illustrate the previous result and understand what kind of information is selectively memorized, we fully characterize the equilibria in the following class of problems. The type space is an interval of the real line. The agent takes a binary action, with the high action being adapted to high types and the low action to low types. The agent's utility increases with his expectation of his type, as would for instance be the case if the type were the agent's own morality or ability. We show that an equilibrium is characterized by either (i) a unique threshold such that the agent voluntarily memorizes his type if and only if it is above that threshold, or (ii) a unique pair of thresholds such that the agent voluntarily memorizes his type if and only if it is below the lower threshold or above the higher one. Said differently, the agent either remembers when he is of a good enough type, or when his type is extremely good or extremely bad. In both cases, when the type is high enough, the agent has an interest in memorizing it because the truth is favorable both for taking the right action and for the belief-based part of his utility. In contrast, when the type is low, forgetting may induce a higher expectation of the state but a suboptimal action. If the material cost of forgetting is larger than its psychological benefit, low types are memorized. We then have an equilibrium with two thresholds in which only intermediary types are forgotten.

In the next section we describe how our results relate to the theoretical literature on strategic information disclosure. We also compare our contribution to some theoretical and empirical findings in the behavioral literature on motivated beliefs and memory management. The general model is presented in Section 3. We establish the full memorization result in Section 4 and introduce exogenous memorization failures in Section 5. Section 6 is devoted to a discussion of possible extensions of the model.

2 Related Literature

Our multi-self memorization game has the same structure as the sender-receiver games from the literature on inter-personal and strategic information disclosure. In the classical disclosure games, starting with Milgrom (1981) and Grossman (1981), the sender only cares about the receiver's action while the receiver would like to match his action to the state. A fully revealing equilibrium is constructed by considering skeptical beliefs, i.e., by assigning deviations from full disclosure to a state that induces the worst action for the sender.[3] In contrast, in our paper there is common interest between Self 1 (the sender) and Self 2 (the receiver). In addition, the agent cares not only about the action taken by Self 2, but also about the actual state and the beliefs of Self 2.[4] For some classes of psychological utility functions, for example when the agent cares only about Self 2's beliefs, we construct a fully revealing equilibrium using skeptical beliefs exactly as in Milgrom (1981). For more general utility functions, for example when beliefs enter directly into the agent's utility but also affect his action, we adapt some technics from generalized disclosure games in which the sender cares both about the receiver's action and the actual state (Seidmann and Winter 1997 and Hagenbach, Koessler, and Perez-Richet 2014). Hence, while the literature on disclosure examines the link between the players' conflits of interests and the possibility of full information disclosure, our paper examines the link between the form of the agent's psychological utility function and the possibility of full memorization. In both cases, weak assumptions are needed to get existence of a fully revealing equilibrium if no further constraint, such as bounded rationality, noisy or costly communication, is added.

Our memorization game with exogenous memory failures is closely related to some models of strategic disclosure that incorporate communication frictions. Specifically, it has the same structure as the disclosure game of Dye (1985), which extends the analysis of Milgrom (1981) by assuming that the sender is not always informed about the state. We show that, as in Dye (1985), a one-threshold equilibrium strategy arises when the agent has state-independent preferences and his utility is monotonically affected by Self 2's beliefs about the state. However, when the optimal action of the agent is state-dependent, there is a trade-off for the agent because beliefs have both an affective and functional role, and the equilibrium disclosure strategy is usually different from a one-threshold strategy. Kőszegi (2006) also sheds light on this trade-off in a model of information disclosure by a sender who is uninformed about the state with some exogenous probability. In his work, the sender and the receiver share a common anticipatory utility and there exist equilibria with a two-threshold structure as in our Proposition 7.[5]

Our paper also contributes to the literature on motivated beliefs, recently growing in behavioral economics and surveyed in Bénabou and Tirole (2016). We provide new elements to

[3]Under further assumptions on players' utility functions, and assuming that actions and states are unidimensional, the fully revealing equilibrium outcome is unique.

[4]We model an agent whose beliefs enter directly his utility function, as in the literature on psychological games recently surveyed in Battigalli and Dufwenberg (2020).

[5]Lipnowski and Mathevet (2018) consider a benevolent sender who designs ex-ante the information to be transmitted to a psychological receiver. We discuss the link to this work in Section 6.

the study of both the demand and supply of such beliefs. On the side of the motivation behind internal beliefs manipulation, we consider a general form of belief-based utility which includes as particular cases the functions used in Kőszegi (2006), Hestermann et al. (2020), as well as in other papers mentioned in Section 3. On the side of the means by which the agent forms self-serving beliefs, our memorization game is inspired by the memory-management model of Bénabou and Tirole (2002) but allows for richer state spaces and different memorization options. In Bénabou and Tirole (2002), as well as in Hestermann et al. (2020), the state space is binary and the first self has asymmetric memorization options in the two states: bad news can be memorized or suppressed, whereas no memorization is the only option when there is no news. Chew, Huang, and Zhao (2020) additionally integrates the idea of delusion, the act of fabricating an event that did not occur. They do so by considering three states and allowing the no news state to be transmitted as good news. We consider a symmetric memorization technology, that is, every state can be either memorized or not memorized. In all the games mentioned above, beliefs are determined in equilibrium by Bayes' rule. This is in contrast to Brunnermeier and Parker (2005) and Caplin and Leahy (2019) who propose beliefs formation models in which beliefs are chosen freely by trading-off the costs and benefits of distorting beliefs away from the Bayesian framework.

Several experiments have been run recently that provide evidence of selective forgetting in the lab. In Zimmermann (2020) or Chew et al. (2020), experimental subjects forget negative feedback about their performance in an IQ test or forget their mistakes in past intelligence tests. In the context of dictator games, Saucet and Villeval (2019) similarly point at an asymmetric recall between past altruistic and selfish decisions.[6] One can interpret these observations as stemming from the 1-threshold equilibrium we characterize: the agent remembers news only if they are good or comforting enough. In addition, we show existence of a 2-threshold equilibrium in which the agent memorizes good news but also bad news which are important to avoid wrong decisions. In the above-mentioned experiments, individuals do not take decisions based on the beliefs about their performance but if they would, remembering failures could be important.

3 Model

3.1 The psychological agent

There is a single agent who has two selves, Self 1 and Self 2, modeled as two different players with the same preferences. There is a non-empty set of states Θ, where Θ is a compact subset of an Euclidean space, with a full-support prior probability distribution $\mu \in \Delta(\Theta)$.[7] Self 1 is privately informed about the realization of the state $\theta \in \Theta$. Self 1 acts in period 1 by disclosing information about the state to Self 2. Self 2 is a priori uninformed about the state, acts in

[6] In the field, Huffman, Raymond, and Shvets (2019) document that managers hold overly-positive memories from their past performance in the workplace. In the psychology literature, Kunda (1990) and Baumeister (2010) document that individuals selectively remember and interpret information in motivated directions.

[7] For every compact set S, $\Delta(S)$ denotes the set of Borel probability measures over S.

period 2 and has a non-empty set of actions A, where A is a compact subset of an Euclidean space. When the state is θ, the action is $a \in A$ and the posterior belief of Self 2 is $\nu \in \Delta(\Theta)$, the common psychological utility of Self 1 and Self 2 is equal to

$$u(a, \theta, \nu).$$

In the standard expected utility framework, $u(a, \theta, \nu)$ does not depend on ν. We assume that the utility function $u : A \times \Theta \times \Delta(\Theta) \to \mathbb{R}$ is continuous.

We let

$$U(a, \nu) := E_{\theta \sim \nu}[u(a, \theta, \nu)] = \int_{\Theta} u(a, \theta, \nu) d\nu(\theta),$$

be the expected utility of the agent when his belief is ν and he chooses action a. Self 2 who holds posterior belief ν chooses an optimal action $a \in A$ by maximizing $U(a, \nu)$.[8] The expected utility of the agent when his belief is ν and he chooses an optimal action given ν is denoted $U^*(\nu) = \max_{a \in A} U(a, \nu)$.

We interpret the model as (a single-agent decision problem represented by) a multi-self game in which the agent has imperfect recall, i.e., the agent is initially informed about a state that he later forgets, and selectively decides which information to memorize for his later self, who acts upon this information. Our focus is on selective recall and we do not allow the agent to make up memories that are initially untrue. Alternatively, the model can be interpreted as a sender-receiver game in which the first player is a privately informed benevolent sender who voluntarily discloses hard information to an uninformed decision-maker.

3.2 Examples of Psychological Utility

In this section we present three broad classes of utility functions of a psychological agent along with more specific examples studied recently in the economic literature.

3.2.1 State-Independent Utility

The utility is state-independent if it does not depend on θ, i.e., it can be written as

$$u(a, \theta, \nu) = u(a, \nu).$$

Example 1 (Intrinsic preference for information) A first example of state-independent utility is one in which the agent does not even take an action (A is a singleton) but is only and intrinsically affected by his beliefs about the state: $u(a, \nu) = u(\nu)$. Masatlioglu, Orhun, and Raymond (2019) present an experiment which focuses on agents' intrinsic preference for information not only considering by how much the information reduces uncertainty about the

[8] Note that, in contrast to the situation studied in Bénabou and Tirole (2002), our agent's preferences are consistent over time in the sense that Self 1 and Self 2 would choose the same action if they held the same belief about θ.

state (the informativeness level) but also considering the kind of uncertainty it eliminates (the skewness of information). ◇

Example 2 (Guilt from consumption) A second example is taken from Hestermann et al. (2020) who propose a model to investigate the "meat paradox", namely the fact that agents consume meat but dislike animal suffering. There is a binary set of states in $(0, 1]$ where the low state corresponds to bad conditions for animals raised to produce meat. The agent chooses the quantity $a \geq 0$ of meat to consume. He incurs a moral cost of guilt when he consumes meat and his perception of θ is low. Precisely, the state-independent utility of the agent is given by: $u(a, \nu) = r(a) - ca - waE_{\theta\sim\nu}(1 - \theta)$, where $r(a)$ is the valuation for meat defined over the consumption level a, $c \geq 0$ is the unit price of meat, and $w \geq 0$ parametrizes the individual level of morality. ◇

3.2.2 Separable Utility

The second category of utility functions is additively separable and can be written as

$$u(a, \theta, \nu) = u_M(a, \theta) + \psi(\nu),$$

where $u_M(a, \theta)$ is a standard material utility, and $\psi(\nu)$ is derived uniquely from the posterior belief ν and is independent of the real state θ and the action a.

Note first that the optimal action of the agent only depends on his material utility because

$$\arg\max_{a\in A} U(a, \nu) = \arg\max_{a\in A} E_{\theta\sim\nu}[u(a, \theta, \nu)] = \arg\max_{a\in A} E_{\theta\sim\nu}[u_M(a, \theta)].$$

The commonly-studied case in which beliefs do not enter into the utility function is a particular case of this functional form by letting $\psi(\nu) = 0$ for every ν. Example 1 (Intrinsic preference for information) is a particular case too, but Example 2 (Guilt from consumption) is not because the effect of the beliefs on the agent's utility cannot be separated from the effect of his action.

Example 3 (Moral self-image) An example of separable psychological utility can be found in Bénabou et al. (2019) (whose basic model builds on Bénabou and Tirole, 2006 and Bénabou and Tirole, 2011). An agent decides whether to act morally or not, which respectively corresponds to actions $a = 1$ or $a = 0$. The (positive) states correspond to the agent's intrinsic motivation to act morally, and can be high or low. Acting morally induces a personal cost $c > 0$ but yields benefits to society, in the form of a positive externality $r \geq 0$. In addition to the material utility derived from the action, the agent derives utility from the image he has about his own morality. His utility function is given by: $u(a, \theta, \nu) = \theta ra - ca + wE_{\tilde{\theta}\sim\nu}(\tilde{\theta})$, where $w \geq 0$ measures the strength of self-image concerns. ◇

Example 4 (Stubbornness) Another example of separable psychological utility was proposed by Lipnowski and Mathevet (2018) to represent an agent who takes actions which

maximize his standard material utility but dislikes changing his prior beliefs: $u(a, \theta, \nu) = u_M(a, \theta) - w|\nu - \mu|$, with $w > 0$. $\diamond$

3.2.3 Anticipatory Utility

The third category of utility functions correspond to anticipatory utilities:

$$u(a, \theta, \nu) = (1 - w)h(a, \theta) + wE_{\tilde{\theta} \sim \nu}[h(a, \tilde{\theta})].$$

An agent with such a utility derives physical utility $h(a, \theta)$ from his action and the state, but also derives utility from the expectation, given his belief ν, of his future physical utility. The parameter $w \in (-1, 1)$ weights the physical and anticipatory utility.

When the agent has anticipatory utility, his optimal action only depends on his physical utility $h(a, \theta)$ because

$$\arg\max_{a \in A} U(a, \nu) = \arg\max_{a \in A} E_{\theta \sim \nu}[u(a, \theta, \nu)] = \arg\max_{a \in A} E_{\theta \sim \nu}[h(a, \theta)].$$

Example 5 (Emotional agency) A natural interpretation of the anticipatory utility corresponds to the case in which $w > 0$ and the agent's well-being increases with the anticipation of his future utility. This is the case studied in Kőszegi (2006) except that the context is that of an informed sender who transmits information to a receiver taking a binary action. The two players share a common utility function which is a particular case of the form given above with a positive parameter w. The author gives the example of a caring doctor who forms a diagnose and transmits information about it to his patient. The doctor is aware that this information will affect not only the treatment that the patient will take but also the patient's emotions regarding his future health. Similarly, parents may have information about their child's prospects in some educational area and know that this information will affect both the child's action and his anticipatory feelings. $\diamond$

Example 6 (Disappointment and elation) Battigalli and Dufwenberg (2020) let the agent's well-being depend on the difference between the expectation of future utility and the realized utility. When the agent was expecting a given utility level and gets a lower one, he experiences disappointment. In the opposite case, he experiences some form of elation. The following utility function permits to incorporate this idea: $u(a, \theta, \nu) = h(a, \theta) + w\big(E_{\tilde{\theta} \sim \nu}[h(a, \tilde{\theta})] - h(a, \theta)\big)$, with $w < 0$ the parameter of sensitivity to the emotions of disappointment and elation. Disappointment corresponds to the case in which $E_{\tilde{\theta} \sim \nu}[h(a, \tilde{\theta})] > h(a, \theta)$ and decreases utility. Elation corresponds to the case in which $E_{\tilde{\theta} \sim \nu}[h(a, \tilde{\theta})] < h(a, \theta)$ and increases utility. This function can be rewritten $u(a, \theta, \nu) = (1 - w)h(a, \theta) + wE_{\tilde{\theta} \sim \nu}[h(a, \tilde{\theta})]$. $\diamond$

3.3 Memorization Game and Equilibrium

The game begins with the realization of the state θ, which is drawn according to the prior μ and which Self 1 observes. After observing the state θ, Self 1 sends a message $m \in \{m_\theta, m_\emptyset\}$ to Self 2. Self 1 of type θ can either disclose his type by sending m_θ, a message that no other type of Self 1 can send, or stay silent and send $m_\emptyset$, a message available to every type of Self 1.[9] Let $M = \{m_\emptyset\} \cup \{m_\theta : \theta \in \Theta\}$ be the set of all available messages in the game. Self 2 observes a message $m \in M$ (but not θ) and chooses an action $a \in A$.

Internal information processing is modeled as an intra-personal disclosure game. Before choosing an action, Self 1 decides for each realization of the state whether to memorize the state (by sending m_θ to his future self) or to forget the state (by sending $m_\emptyset$ to his future self). When the agent decides to memorize the state, it means that he actively decides to incorporate this piece of information into his beliefs. This can take the form of a cognitive process or the form of a tangible move that helps remember the state. Some examples of the latter are making a reminding note, repeating the state aloud, trying to put it in context etc.

A strategy for Self 1 is a function $\sigma_1 : \Theta \to [0,1]$, where for every $\theta \in \Theta$, $\sigma_1(\theta)$ is the probability that Self 1 sends message m_θ and $1-\sigma_1(\theta)$ is the probability that he sends message $m_\emptyset$. A strategy for Self 2 is a function $\sigma_2 : M \to A$.[10] After message m_θ, the belief of Self 2 is simply δ_θ, the probability distribution which assigns probability 1 to the state θ, because the information set of Self 2 after such a message is reduced to a singleton. Hence, a belief system for Self 2 is simply characterized by his belief $\nu \in \Delta(\Theta)$ when he receives message $m_\emptyset$.

Psychological Perfect Bayesian Equilibrium A (psychological perfect Bayesian) equilibrium is a profile of strategies and beliefs $(\sigma_1, \sigma_2, \nu)$ such that: 1. Beliefs are computed by Bayes rule whenever possible (otherwise beliefs are arbitrary).[11] 2. After every message, Self 2 chooses on optimal action given his beliefs after this message. 3. For every state, Self 1 sends a message with strictly positive probability only if this message maximizes his expected utility. Formally:

1. ν is obtained from μ and σ_1 by Bayes' rule, i.e.,

$$\nu(\theta)\left(1 - \int_\Theta \sigma_1(\theta)\, d\mu(\theta)\right) = (1-\sigma_1(\theta))\mu(\theta), \quad \text{for all } \theta \in \Theta;$$

2. $\sigma_2(m_\emptyset) \in \arg\max_{a\in A} U(a,\nu)$ and $\sigma_2(m_\theta) \in \arg\max_{a\in A} U(a,\delta_\theta)$ for every $\theta \in \Theta$;

3. $\sigma_1(\theta) > 0$ implies $u(\sigma_2(m_\emptyset),\theta,\nu) \leq U^*(\delta_\theta)$, and $\sigma_1(\theta) < 1$ implies $u(\sigma_2(m_\emptyset),\theta,\nu) \geq U^*(\delta_\theta)$.

[9] Our general results (before Section 5.2) are unchanged if Self 1 is also able to partially disclose his type (i.e., disclose of subset of types including his actual type) to Self 2; see Remark 2 for more details.

[10] The set A can be replaced by $\Delta(A)$ to allow for mixed strategies.

[11] Hence, beliefs are arbitrary only when the message $m_\emptyset$ is sent with probability zero.

4 When is Perfect Memory Voluntary?

Perfect memory is voluntary if Self 1 memorizes every state: $\sigma_1(\theta) = 1$ for every $\theta \in \Theta$. In this case, whatever the state θ, the equilibrium payoff is $U^*(\delta_\theta)$, which corresponds to the payoff that the agent would get under complete information. We say that an equilibrium is fully revealing if it is payoff-equivalent to an equilibrium with voluntary perfect memory. By definition, a fully revealing equilibrium exists iff there exists $\nu \in \Delta(\Theta)$ and $\tilde{a} \in \arg\max_{a \in A} U(a, \nu)$ such that

$$U^*(\delta_\theta) := \max_{a \in A} u(a, \theta, \delta_\theta) \geq u(\tilde{a}, \theta, \nu), \text{ for every } \theta \in \Theta.$$

Said differently, Self 1 memorizes every state iff, for every θ, the message $m_\emptyset$ induces a belief ν for Self 2 that leads to a lower payoff than believing the true θ.

Of course, in the standard expected utility framework there is always a fully revealing equilibrium, with any belief ν, because when the function u does not depend of ν we have $U^*(\delta_\theta) = \max_{a \in A} u(a, \theta) \geq u(\tilde{a}, \theta)$ for every $\tilde{a} \in A$. In addition, for every θ, the utility of the agent is the complete information utility $U^*(\delta_\theta)$ in every equilibrium, because otherwise Self 1 would deviate to message m_θ to get his first best $U^*(\delta_\theta)$. In the next section we provide an example of psychological utility for which there is no fully revealing equilibrium. Then, we give sufficient conditions on the agent's utility for the existence of a fully revealing equilibrium.

4.1 An Example without a Fully Revealing Equilibrium

In the next example, there is no fully revealing equilibrium. The psychological utility of the agent for belief ν is negatively correlated with the true state, meaning that this agent dislikes believing the truth whatever it is. It follows that, for any belief ν following $m_\emptyset$, Self 1 deviates from full revelation.

Example 7 Let $\Theta = \{0, 1\}$, identify ν with the belief on $\theta = 1$, and assume that

$$u(a, \theta, \nu) = u(\theta, \nu) = \begin{cases} -\nu & \text{if } \theta = 1 \\ -(1 - \nu) & \text{if } \theta = 0. \end{cases}$$

There is a fully revealing equilibrium iff there exists ν such that $u(1, 1) = -1 \geq u(1, \nu) = -\nu$ and $u(0, 0) = -1 \geq u(0, \nu) = -(1 - \nu)$, which is impossible. $\diamond$

In the appendix we provide an additional example (Example 9) in which there is no fully revealing equilibrium but the utility of the agent takes the following form: $u(a, \theta, \nu) = u_M(a, \theta) + \psi(a, \nu)$.

4.2 Existence of a Fully Revealing Equilibrium

As in the literature on disclosure games, we construct a fully revealing equilibrium by identifying a type $\hat{\theta}$ such that, if it is believed by Self 2 after $m_\emptyset$, every type of Self 1 gets a payoff which is not higher than his full information payoff. $\hat{\theta}$ is called a worst-case type. Regarding our memory interpretation, a Self 2 who believes $\hat{\theta}$ after $m_\emptyset$ is skeptical about his own lack of memory. We have in mind an agent who realizes that, since he is able to voluntarily memorize every state, no memory must be a sign of bad news. Bénabou and Tirole (2002) talk about *metacognition* when the agent is able to make such inferences. We go back to the issue of skepticism later in Remark 1 and Section 6.1.

In the following propositions we provide sufficient conditions on the agent's utility function for the existence of a fully revealing equilibrium. All proofs can be found in the appendix. The first proposition applies to all state-independent utility functions (Section 3.2.1), and therefore applies to Examples 1 (Intrinsic preference for information) and 2 (Guilt from consumption). For state-independent utilities, the worst-case type is simply the type $\hat{\theta}$ that minimizes $u(a^*(\theta), \delta_\theta)$, with $a^*(\theta)$ the optimal action given θ. For instance, it corresponds to the bad conditions for animals in Example 2.

Proposition 1 (State-Independent Utilities) *If $u(a, \theta, \nu) = u(a, \nu)$, then there exists a fully revealing equilibrium.*

The next proposition applies whenever there is a belief for the agent which is worst for him whatever the true state and the action. The worst case type is $\hat{\theta} \in \arg\min_{\tilde{\theta}} u(a, \theta, \delta_{\tilde{\theta}})$. In particular, it applies to all separable psychological utility functions (Section 3.2.2), and therefore applies to Examples 3 (Moral self-image) and 4 (Stubborness). It also covers applications in which the belief-based utility has a clear direction regardless of the structure (e.g., separability) of the utility function. In particular, it is consistent with a model where the intensity of the preference for high beliefs depends on the true state and/or on the action, as long as the agent always prefers high beliefs. This is the case in Example 8 presented after the proposition. This example is a modified version of Examples 2 (Guilt from consumption) and 3 (Moral self-image) in which the agent's utility is state-dependent and the cost of guilt associated to the belief of a lower type is larger when his action is high.

Proposition 2 *If there exists $\hat{\theta}$ such that $\min_{\tilde{\theta}} u(a, \theta, \delta_{\tilde{\theta}}) = u(a, \theta, \delta_{\hat{\theta}})$ for all a, θ, then there exists a fully revealing equilibrium. In particular, there is a fully revealing equilibrium in the following cases:*

(Separable Utility) $u(a, \theta, \nu) = u_M(a, \theta) + \psi(\nu)$.

(Directional belief-based utility) $u(a, \theta, \delta_{\tilde{\theta}})$ *is nondecreasing in $\tilde{\theta}$ for all a, θ.*

Example 8 (State-dependence and guilt from consumption) The states are linearly ordered and a low state corresponds to a type of product which is less good from, say, an ethical

or environmental point of view. The agent decides to consume or not, which respectively corresponds to actions $a = 1$ or $a = 0$. A lower θ induces a higher moral cost of guilt for the agent when he consumes. In contrast to Hestermann et al. (2020), the material benefit of consuming the good increases with θ, capturing the idea that the ethical or environmental dimension of a good can impact directly the material benefit derived from consuming it. For example, while organic vegetables or eggs make the agent feel less guilty about consumption, they may also have a better taste. The state-dependent utility of the agent is now given by

$$u(a, \theta, \nu) = \theta r a - ca - waE_{\tilde{\theta}\sim\nu}(1 - \tilde{\theta}),$$

with $r \geq 0$, $c > 0$ the unit price of meat, and $w \geq 0$ the individual level of morality. $\diamond$

The next proposition applies to all anticipatory utility functions (Section 3.2.3). For this class of utility functions, a worst-case type is a type that minimizes (maximizes, resp.) the material utility derived from the state and the optimal action in that state if $w \geq 0$ ($w \leq 0$, resp.).

Proposition 3 (Anticipatory Utility) *If $u(a, \theta, \nu) = (1 - w)h(a, \theta) + wE_{\tilde{\theta}\sim\nu}[h(a, \tilde{\theta})]$, then there exists a fully revealing equilibrium.*

Finally, the next proposition provides a sufficient condition for the existence of a fully revealing equilibrium in the class of utility functions that can be written as $u(a, \theta, \nu) = u_M(a, \theta) + \psi(a, \nu)$, under a monotonicity condition on the optimal action $a^*(\theta)$ (or any selection $a^*(\theta) \in \arg\max_{a\in A} U(a, \delta_\theta)$ in case of multiple optimal actions) and an increasing difference condition on the material utility function. These conditions are satisfied in Examples 1, 2, 3 and 8.

Proposition 4 *Assume that A and Θ are (possibly finite) compact subsets of $\mathbb{R}$ endowed with their natural order, $u(a, \theta, \nu) = u_M(a, \theta) + \psi(a, \nu)$, $a^*(\theta)$ is continuous and increasing in θ, and $u_M(a, \theta)$ has increasing differences in (a, θ) (i.e., for every $a' \geq a$, $u_M(a', \theta) - u_M(a, \theta)$ is increasing in θ). Then, there exists a fully revealing equilibrium.*

In Example 9 in the appendix, the agent's utility function takes the form $u(a, \theta, \nu) = u_M(a, \theta) + \psi(a, \nu)$ but there is no fully revealing equilibrium. In this example, the increasing difference assumption of Proposition 4 does not hold (but the other assumptions of the proposition are satisfied).

This first set of four propositions establishes that, for broad categories of psychological utilities, there exists an equilibrium in which every state is memorized by the agent. In such an equilibrium, there is no self-manipulation and the agent acts under complete information even if he could selectively forget the state and thereby influence his beliefs. We close this section on the existence of fully revealing equilibria with three remarks.

Remark 1 (Off-path beliefs) The fully revealing equilibria constructed in this section are such that every type θ discloses m_θ, so the message $m_\emptyset$ is off path. Hence, the construction of these equilibria relies on Self 2 being skeptical about the lack of information, and assigning probability one to a worst-case type $\hat{\theta}$ when he faces $m_\emptyset$. However, there is an outcome-equivalent equilibrium with no message off path, in which every type $\theta \neq \hat{\theta}$ discloses m_θ and type $\hat{\theta}$ sends message $m_\emptyset$. It follows that the fully revealing equilibrium outcome could be sustained without specific off-path beliefs for Self 2, he is just required to apply Bayes' rule given Self 1's strategy and the prior probability distribution of the state. This observation also implies that the fully revealing equilibrium does not rely on unreasonable off-path beliefs, and therefore satisfies standard refinement criteria for signaling games.

Remark 2 (Partial disclosure) The existence results of this section all extend to the case in which Self 1 can partially disclose information to Self 2 as long as every type has access to at least one message that no other type has access to (the condition of *own type certifiability* in Hagenbach et al., 2014). An interpretation of partial disclosure is that the agent would, for example, memorize that the state is in a low range but would not memorize the state more precisely. Formally, Self 1 of type θ could send a message $m \in M(\theta)$, i.e., $M(\theta)$ is the set of messages available to type θ. Message m then corresponds to the memorization of the event $M^{-1}(m) := \{\theta \in \Theta : m \in M(\theta)\}$. Own type certifiability requires that for every θ, there exists a message $m \in M(\theta)$ such that θ is the only type able to send message m, i.e., $M^{-1}(m) = \{\theta\}$.[12] The proofs of Propositions 1, 2 and 3 extend by considering for every off-equilibrium-path message m (not only $m_\emptyset$) a worst-case type $\hat{\theta}$ satisfying the constraint that m is a feasible message for type $\hat{\theta}$, i.e., $m \in M(\hat{\theta})$ or, equivalently, $\hat{\theta} \in M^{-1}(m)$. In Proposition 1 for example, for every off-path message m we would let $\hat{\theta} \in \arg\min_{\theta \in M^{-1}(m)} u(a^*(\theta), \delta_\theta)$. The proof of Proposition 4 also extends because the binary relation (defined by whether or not θ wants to induce belief θ') has a minimal element on every non-empty subset of Θ.

Remark 3 (Unicity of the fully revealing equilibrium) Without putting more structure on the forms of the utility functions, there may exist other equilibria that are not fully revealing.[13] Our focus is to contrast the possibility of voluntary perfect memory in equilibrium under weak assumptions on the agent's utility function with its general impossibility in the presence of exogenous memorization failures introduced in the next section.

[12] The memorization technologies considered in Bénabou and Tirole (2002) and Chew et al. (2020) do not satisfy own type certifiability. In Bénabou and Tirole (2002), the type is either low, θ_L, or high, θ_H. The memorization technology is given by $M(\theta_L) = \{m_L, m_\emptyset\}$, $M(\theta_H) = \{m_\emptyset\}$. There is no message available to θ_H only. In Chew et al. (2020), there are three possible states: $\theta_L, \theta_M, \theta_H$. The memorization technology is given by $M(\theta_L) = \{m_L, m_\emptyset\}, M(\theta_M) = \{m_\emptyset, m_H\}, M(\theta_H) = \{m_H\}$, so there is no message available to θ_M only or to θ_H only.

[13] For the class of anticipatory utilities for instance, Kőszegi (2006) combines a binary set of actions and a particular form of utility and demonstrates that the fully revealing equilibrium is unique in his Proposition 8.

5 Exogenous Memorization Failures

In this section we extend the model by assuming that there are exogenous (non-strategic) memorization failures against which the agent cannot do anything. Precisely, for each $\theta \in \Theta$ and whatever his strategy, Self 1 is able to memorize the state with probability $\alpha \in (0,1)$. With the complementary probability $1-\alpha$, Self 1 is unable to memorize the state even if he wants to. We are back to the previous model in the limit case in which $\alpha = 1$.[14] Equivalently, with probability $1-\alpha$, the only message available to Self 1 is $m_\emptyset$.[15]

In this model, we say that an equilibrium is fully revealing if every type θ memorizes the state with probability 1 when able to do so. We first observe that if the utility of the agent is standard (it does not directly depend on his beliefs), then there is a fully revealing equilibrium, and all equilibria are payoff-equivalent.

Observation 1 Let $\alpha \in (0,1)$ be the probability that Self 1 is able to memorize and assume that the agent has a standard utility function $u(a,\theta,\nu) = u(a,\theta)$. Then, there is a fully revealing equilibrium. In addition, in every equilibrium, the equilibrium payoff of the agent type θ who is able to memorize is the complete information payoff $U^*(\delta_\theta)$.

Indeed, full revelation constitutes an equilibrium because for every $\theta \in \Theta$ the induced payoff of the agent who is able to memorize is $U^*(\delta_\theta) \geq u(a,\theta)$ for every $a \in A$. All equilibria induce such a payoff for every type θ who is able to memorize because otherwise Self 1 would deviate and send message m_θ, and would get his first best $U^*(\delta_\theta)$.

5.1 Voluntary Selective Memory

The next proposition shows that if there are exogenous memorization failures and the marginal effect of the agent's beliefs on his utility is not negligible, then there is no fully revealing equilibrium. Said differently, the agent always voluntarily forgets some information. To illustrate this point, consider Example 3 with $\Theta = [0,1]$ and $u(a,\theta,\nu) = a(r\theta - c) + E_{\tilde{\theta}\sim\nu}(\tilde{\theta})$, where $c \neq E_{\theta\sim\mu}(\theta)$ and $r = 1$. The expected utility of Self 2 with belief ν when he takes action a is $U(a,\nu) = a(E_{\theta\sim\nu}(\theta) - c) + E_{\theta\sim\nu}(\theta)$. His optimal action is $a = 1$ if $E_{\theta\sim\nu}(\theta) > c$ and $a = 0$ if $E_{\theta\sim\nu}(\theta) < c$.

Assume that $\alpha \in (0,1)$ and consider a fully revealing strategy for Self 1. Then, the belief $\nu \in \Delta(\Theta)$ of Self 2 when he receives message $m_\emptyset$ (i.e., when he has no memory) is the prior, $\nu = \mu$, so $E_{\theta\sim\nu}(\theta) = E_{\theta\sim\mu}(\theta)$ and the agent takes action $a = 0$ if $E_{\theta\sim\mu}(\theta) < c$ and $a = 1$ if $E_{\theta\sim\mu}(\theta) > c$. It is immediate that there is no fully revealing equilibrium: every type θ slightly below $E_{\theta\sim\mu}(\theta)$ is better-off by deviating to message $m_\emptyset$ because he induces the same action

[14] For expositional simplicity, we assume that Self 1's ability to memorize the state is independent of the state, i.e., α is independent of θ. However, the model can be extended by considering a state-dependent probability to memorize. In this case, Observation 1 and Proposition 5 continue to apply by replacing the prior probability distribution μ by the appropriate conditional probability distribution.

[15] Another equivalent interpretation is that α is the probability that Self 1 is initially informed about the state.

but increases the conditional expectation of the state (from θ to $E_{\theta\sim\mu}(\theta)$). Note that a fully revealing equilibrium exists only in the non-generic case in which $c = E_{\theta\sim\mu}(\theta)$.

We show that the impossibility to have an equilibrium in which the agent voluntarily memorizes all the information applies much more generally under the following assumption.

Assumption 1

1. The utility of the agent only depends on his belief $\nu \in \Delta(\Theta)$ through the expected state given ν, i.e., it can be written as

$$u(a, \theta, e), \text{ where } e = E_{\theta\sim\nu}(\theta);$$

2. $\Theta \subset \mathbb{R}$ and A are convex,[16] and $u(a, \theta, e)$ is continuously differentiable on $A \times \Theta \times \Theta$;

3. For every $\nu \in \Delta(\Theta)$, the set of optimal actions of the agent as a function of his belief ν only depends on the expected state given ν, $e = E_{\theta\sim\nu}(\theta)$, and is denoted by $A^*(e) = \arg\max_{a\in A} U(a, \nu)$;

4. The optimal action of the agent is unique and given by $a^*(e)$ in a neighborhood of $\overline{e}$, and $a^*(e)$ is differentiable at $e = \overline{e}$, where $\overline{e} = E_{\theta\sim\mu}(\theta)$;[17]

5. *Locally non-satiated psychological utility.* The derivative of $u(a, \theta, e)$ with respect to e at $(a, \theta, e) = (a^*(\overline{e}), \overline{e}, \overline{e})$ is non-zero.

Proposition 5 *Under Assumption 1 there is no fully revealing equilibrium.*

The intuition of this result is as follows. Consider a fully revealing strategy. Any agent type θ can induce a conditional expected valuation equal to $\overline{e}$ by forgetting the information. Forgetting the information changes the second-period action in a sub-optimal way but, if θ is close enough to $\overline{e}$, this has a second-order effect on the agent's utility. In contrast, the modification of the second-period belief has a first-order effect on the utility by Assumption 1.[18] Hence, some types close enough to $\overline{e}$ have an incentive to deviate from full revelation.

Note that Proposition 5 extends to the case in which Self 1 can partially memorize his type as described in Remark 2. Indeed, the proof relies on showing deviation from a fully revealing strategy to a strategy such that $m_\emptyset$ is sent, so it works as long as message $m_\emptyset$ is available to Self 1. It is also clear from the proof of the proposition that, under Assumption 1, there is no equilibrium which is almost perfectly revealing, i.e., such that $\sigma_1(\theta) = 1$ for almost all $\theta \in \Theta$. Finally, note that every worst-case type $\hat{\theta}$ identified in the previous section corresponds to a type for which the condition "locally non-satiated psychological utility" is not satisfied at $\hat{\theta}$.

[16] The assumption can be generalized to multidimensional state spaces by applying directional derivatives. If the set of actions is finite, it can be made convex by replacing it by the set of mixed actions.

[17] In the previous example, this assumption is satisfied if $c \neq \overline{e}$.

[18] The idea that beliefs distortions can have a first-order benefit but a second-order cost can also be found in other behavioral models such as Compte and Postelwaite (2004) or Gottlieb (2010).

5.2 Equilibrium Characterization

In the previous section we have shown that, in general, the agent voluntarily forgets some information as long as there are some exogenous memorization failures. In this section we characterize partially revealing equilibria for some examples and classes of utility functions studied earlier, under the following assumptions. We assume that $\Theta = [0, 1]$ and the cumulative distribution function $F(\theta)$ is continuous and strictly increasing. The agent's utility can be written as $u(a, \theta, e)$, where e is the conditional expected state and $u(a, \theta, e)$ is strictly increasing in e. Finally, the agent's optimal decision (or, in case of multiplicity, any selection) only depends on his conditional expected state, and is denoted by $a^*(e)$. The ex-ante expected value of the state is denoted $\overline{e} = E_{\theta\sim\mu}(\theta)$.

5.2.1 State-Independent Utility

Assume that the utility function of the agent is state-independent, i.e., it can be written as $u(a, e)$. The assumption that u is strictly increasing in e implies that $u(a^*(e), e)$ is also strictly increasing in e. For instance, Example 2 (Guilt from consumption) satisfies these assumptions. The next proposition shows that the equilibrium is unique: the agent voluntarily memorizes news iff they are good enough.

Proposition 6 *Consider the class of state-independent utility of Section 5.2.1. There exists a unique equilibrium, characterized by the following 1-threshold memorization strategy:*

$$\sigma_1(\theta) = \begin{cases} m_\theta & \text{if } \theta > \theta_D \\ m_\emptyset & \text{if } \theta < \theta_D, \end{cases}$$

where θ_D is the unique solution in $(0, \overline{e})$ of the following equation:

$$\theta_D = \frac{\alpha F(\theta_D) E[\theta|\theta < \theta_D] + (1-\alpha)\overline{e}}{\alpha F(\theta_D) + (1-\alpha)}. \tag{1}$$

When the agent has state-independent preferences and his utility is positively affected by his belief about the state (i.e., u is strictly increasing in e), the equilibrium is exactly the same as in the disclosure model of Dye (1985) and Jung and Kwon (1988) where the sender always wants the receiver to believe that the state is high. The equilibrium threshold θ_D only depends on α and on the distribution F of the state, not on the parameters of the agent's utility function because the agent's utility is monotonic in Self 2's belief. Hence, whatever the state, Self 1 wants Self 2 to have the highest belief. Note that the equilibrium threshold θ_D is strictly increasing in α: if $\alpha \to 0$, then Self 1 memorizes the state iff it is higher than the ex-ante expected value of the state ($\theta_D = \overline{e}$); if $\alpha \to 1$, then the unique equilibrium is the fully revealing equilibrium ($\theta_D = 0$) obtained in Proposition 1.

Compared to the situation in which Self 2 acts while being fully informed, for states above the threshold, Self 2's beliefs are distorted downwards on average, but they are not distorted

when the agent is able to memorize. For states below the threshold, Self 2's beliefs and actions are always distorted upwards. In the context of Example 2 (extended to consider a continuous set of types Θ), when animals conditions are too bad, the psychological agent forms a biased optimistic expectation about animals suffering and consumes more meat than he would do if he were to face the truth. The lower is α, that is, the more important are the exogenous memorization failures, the larger is the wiggle room that this agent can use to self deceive in such an optimistic way.

5.2.2 Separable Utility

We consider the case where $u(a, \theta, e) = u_M(a, \theta) + \psi(e)$ and $A = \{0, 1\}$. We assume that $u_M(1, \theta) - u_M(0, \theta)$ is strictly increasing in θ. For the analysis to be interesting we assume that $u_M(1, 0) - u_M(0, 0) < 0$ and $u_M(1, 1) - u_M(0, 1) > 0$. Hence, there exists $\overline{\theta} \in (0, 1)$ such that

$$a^*(\theta) = \begin{cases} 0 & \text{if } \theta < \overline{\theta} \\ 1 & \text{if } \theta > \overline{\theta}. \end{cases}$$

Finally, we assume that $\psi(\theta)$ is strictly increasing in θ and $u_M(1, \theta) - u_M(0, \theta) - \psi(\theta)$ is strictly quasi-concave in θ.[19] These assumptions are satisfied in Example 3 (Moral self-image). The next proposition characterizes all equilibria. An equilibrium is either characterized by (i) a unique threshold such that the agent voluntarily memorizes his information iff the state is above the threshold, or (ii) a unique pair of thresholds such that the agent voluntarily memorizes his information iff the state is below the lower threshold or above the higher threshold.

Proposition 7 *Consider the class of separable utility of Section 5.2.2. An equilibrium is either characterized by a 1-threshold or by a 2-threshold memorization strategy:*

1. *There exists an equilibrium characterized by a 1-threshold memorization strategy iff* $u_M(0, 0) - u_M(a^*(\theta_D), 0) \leq \psi(\theta_D) - \psi(0)$. *The 1-threshold memorization strategy is unique and given by:*

$$\sigma_1(\theta) = \begin{cases} m_\theta & \text{if } \theta > \theta_D \\ m_\emptyset & \text{if } \theta < \theta_D, \end{cases}$$

   where θ_D *is the unique solution in* $(0, \overline{e})$ *of Equation (1).*

2. *If* $u_M(0, 0) - u_M(a^*(\theta_D), 0) > \psi(\theta_D) - \psi(0)$, *then there exists a unique equilibrium, char-*

[19] A function $f : \Theta \to \mathbb{R}$ is strictly quasi-concave if for all $\theta_1 \neq \theta_2$ and $\lambda \in (0, 1)$ we have $f(\lambda\theta_1 + (1 - \lambda)\theta_2) > \min\{f(\theta_1), f(\theta_2)\}$. That is, there exists $\theta^p \in \Theta$ such that f is strictly increasing for $\theta < \theta^p$ and strictly decreasing for $\theta > \theta^p$

acterized by a 2-threshold memorization strategy:

$$\sigma_1(\theta) = \begin{cases} m_\theta & \text{if } \theta > \overline{\theta}^*, \\ m_\emptyset & \text{if } \underline{\theta}^* < \theta < \overline{\theta}^*, \\ m_\theta & \text{if } \theta < \underline{\theta}^*, \end{cases}$$

where $(\underline{\theta}^*, \overline{\theta}^*)$ *is the unique solution such that* $0 < \underline{\theta}^* < \overline{\theta} < \overline{\theta}^* < \overline{e}$ *solving the following equations:*

$$\alpha(F(\overline{\theta}^*) - F(\underline{\theta}^*))\big(\overline{\theta}^* - E[\theta|\underline{\theta}^* < \theta < \overline{\theta}^*]\big) = (1-\alpha)(\overline{e} - \overline{\theta}^*); \tag{2}$$

$$u_M(0, \underline{\theta}^*) - u_M(1, \underline{\theta}^*) + \psi(\underline{\theta}^*) = \psi(\overline{\theta}^*). \tag{3}$$

To get an intuition of the proposition, first consider the case of intrinsic preference for information, i.e., $u_M(a, \theta) = 0$. Then, when $\psi(e)$ is strictly increasing in e, we are always in the case 1 of the previous proposition and in the case of state-independent utility studied in Proposition 6: the unique equilibrium is a 1-threshold equilibrium.

When the agent takes a payoff-relevant action which affects his (state-dependent) material utility, there might be a tradeoff between memorizing a low state or not memorizing it because when the agent does not memorize information he might take a suboptimal action. For low states, if the material cost of forgetting information is small compared to the psychological benefit of forgetting it, then the agent prefers to forget the information, as in the case of state-independent preferences. This corresponds to the case in which $u_M(0,0) - u_M(a^*(\theta_D), 0) \leq \psi(\theta_D) - \psi(0)$ as established in the proposition. However, if this inequality is not satisfied, the material cost of forgetting information is high compared to the psychological benefit and we are in the case 2 of the proposition. The agent prefers to memorize information when the state is low ($\theta < \underline{\theta}^*$). For intermediate states, in particular for states around $\overline{\theta}$, the action has a small effect on the material utility, so the psychological benefit dominates the material cost, and therefore the agent prefers to forget information. Finally, when the state is high ($\theta > \overline{\theta}^*$), forgetting information has both a material and psychological cost for the agent, so he always memorizes information.

In a 2-threshold equilibrium, when the agent is able to memorize, only intermediate states are distorted upwards compared to the complete information case: when the state θ is such that $\underline{\theta}^* < \theta < \overline{\theta}^*$, Self 2 forms an expectation of the state that equals $\overline{\theta}^*$ and takes action $a = 1$. When the state θ is above $\underline{\theta}^*$ but below $\overline{\theta}$, this action is different from the action $a = 0$ that would be taken under complete information.

Contrary to the 1-threshold equilibrium, the interval of states in which the state is memorized in a 2-threshold equilibrium also depends on the agent's utility function: the solution $(\underline{\theta}^*, \overline{\theta}^*)$ of Equations (2) and (3) depends on u_M and ψ. As an illustration, consider Example 3, i.e., $u(a, \theta, \nu) = \theta r a - ca + wE_{\tilde{\theta}\sim\nu}(\tilde{\theta})$, and assume the prior probability distribution of the state

is uniform. Then, we get $\overline{\theta} = \frac{c}{r}$, $\overline{e} = \frac{1}{2}$ and

$$\theta_D = \frac{\alpha F(\theta_D) E[\theta|\theta < \theta_D] + (1-\alpha)\overline{e}}{\alpha F(\theta_D) + (1-\alpha)} = \frac{\alpha\theta_D \frac{\theta_D}{2} + \frac{(1-\alpha)}{2}}{\alpha\theta_D + (1-\alpha)},$$

i.e., $\theta_D = \frac{\sqrt{1-\alpha}}{1+\sqrt{1-\alpha}}$. There is a 1-threshold equilibrium iff condition 1 of the proposition is satisfied, i.e., if $\theta_D \leq \frac{c}{r}$ or $\theta_D \geq \max\{\frac{c}{r}, \frac{c}{w}\}$. In particular, if the weight w on the belief-dependent part of the utility is high ($w > r$), then there is always a 1-threshold equilibrium. Likewise, if $c \to 0$ or $c \to r$ then the optimal action of the agent is state-independent, and there is a 1-threshold equilibrium.

Otherwise, if $\frac{c}{r} < \theta_D < \frac{c}{w}$, then the unique equilibrium is a 2-threshold equilibrium, and the pair of thresholds $(\underline{\theta}^*, \overline{\theta}^*)$ is the unique solution of Equations (2) and (3), which simplify to:

$$\alpha \left(\overline{\theta}^* + \frac{w\overline{\theta}^* - c}{r - w} \right)^2 = 2(1-\alpha) \left(\frac{1}{2} - \overline{\theta}^* \right); \tag{4}$$

$$\underline{\theta}^* = \frac{c - w\overline{\theta}^*}{r - w}. \tag{5}$$

When $\alpha \to 0$ we get $\underline{\theta}^* = \frac{c - \frac{w}{2}}{r-w}$ and $\overline{\theta}^* = 1/2$. If α increases, then $\underline{\theta}^*$ increases and $\overline{\theta}^*$ decreases: as in a 1-threshold equilibrium, more information is memorized when the exogenous probability of memorization failures $(1-\alpha)$ decreases. It is also immediate to show from the equations above that if w increases, then $\overline{\theta}^*$ and $\underline{\theta}^*$ decrease and $\overline{\theta}^* - \underline{\theta}^*$ increases. In particular, when w increases, the psychological gain of not memorizing information increases for low types and therefore low types have less incentives to memorize information. Symmetrically, if c increases, then $\overline{\theta}^*$ and $\underline{\theta}^*$ increase and $\overline{\theta}^* - \underline{\theta}^*$ decreases. When c increases, the material cost of not memorizing information increases for low types and therefore low types have more incentives to memorize information in order to implement the appropriate decision.

In Example 3, the 2-threshold equilibrium exists if and only if the solution of Equations (4) and (5) is such that $\frac{c}{r} \leq \overline{\theta}^* \leq \frac{c}{w}$. While this condition is satisfied when $\frac{c}{r} < \theta_D < \frac{c}{w}$ (i.e., when there is no 1-threshold equilibrium), it can also be that $\frac{c}{r} \leq \overline{\theta}^* \leq \frac{c}{w}$ when $\theta_D < \frac{c}{r}$, implying that the two types of equilibria can co-exist. As an illustration, we represent the equilibrium thresholds, θ_D, $\overline{\theta}^*$ and $\underline{\theta}^*$ as a function of α on Figures 1 and 2.

The kind of 2-threshold equilibrium exhibited in Proposition 7 can also be obtained in the class of anticipatory utilities and in Example 8, two cases that are not covered by the previous proposition. For a particular class of anticipatory utilities, the last proposition of Kőszegi (2006) shows that there can exist an equilibrium with three zones similar to ours and potentially other equilibria including a fully revealing one. For Example 8, we can perform a similar exercise as for Example 3 and show that a 2-threshold equilibrium exists if $w < r - 2c$.

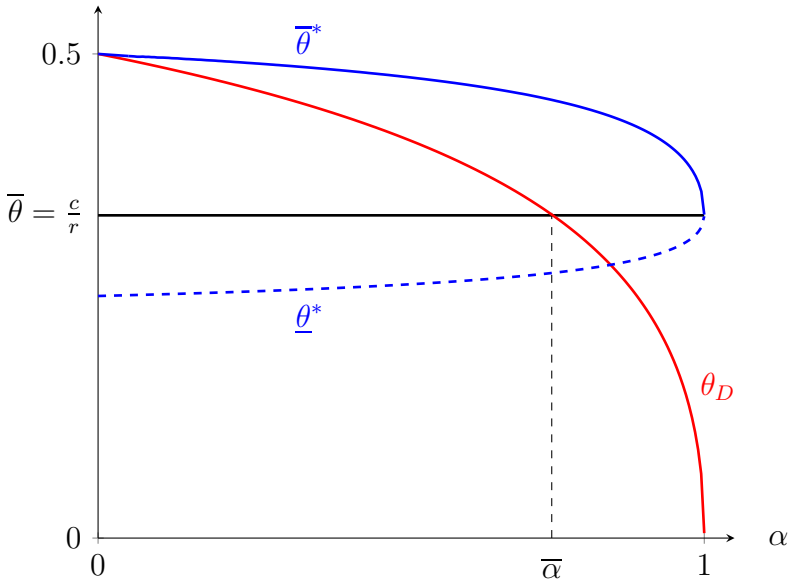

Figure 1: Equilibrium thresholds as a function of α in Example 3 with $r = 1$ and $c = w = \frac{1}{3}$. A 1-threshold equilibrium (θ_D) exists iff $\alpha \geq \overline{\alpha} = \frac{3}{4}$. A 2-threshold equilibrium $(\underline{\theta}^*, \overline{\theta}^*)$ always exists.

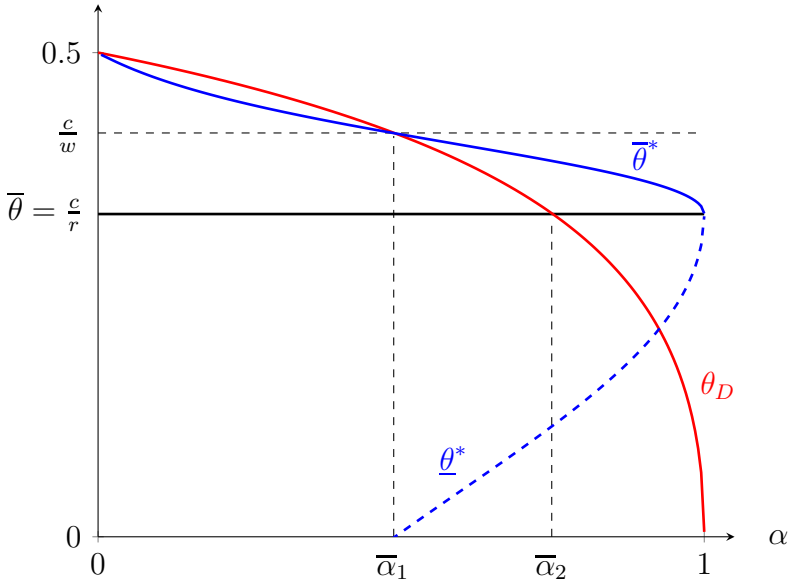

Figure 2: Equilibrium thresholds as a function of α in Example 3 with $r = 1$, $c = \frac{1}{3}$ and $w = \frac{4}{5}$. A 1-threshold equilibrium (θ_D) exists iff $\alpha \leq \overline{\alpha}_1 = \frac{24}{49}$ or $\alpha \geq \overline{\alpha}_2 = \frac{3}{4}$. A 2-threshold equilibrium $(\underline{\theta}^*, \overline{\theta}^*)$ exists iff $\alpha \geq \overline{\alpha}_1$.

6 Discussion and Extensions

6.1 Naive Agent

In the first part of the paper, the existence of fully revealing equilibria relies on Self 2's skepticism on or off the equilibrium path as explained by Remark 1. In contrast, a naive Self 2 takes every message at face value, even along the equilibrium path. That is, his belief after message m_θ is δ_θ, like a sophisticated agent, but his belief is the prior μ when the message is $m_\emptyset$. Such a naive agent has been considered by Milgrom and Roberts (1986). He corresponds to a Self 2 who is "fully cursed" in the sense of Eyster and Rabin (2005), or who is simplifying the memorization strategy of Self 1 by coarsely grouping all states in the same analogy class as in the analogy-based expectation equilibrium of Jehiel (2005): Self 2 only knows the probability that Self 1 memorizes information, but he does not know the probability that Self 1 memorizes information conditional on the state.

When Self 2 is naive, a fully revealing equilibrium exists iff there exists $\tilde{a} \in \arg\max_{a \in A} U(a, \mu)$ such that

$$U^*(\delta_\theta) := \max_{a \in A} u(a, \theta, \delta_\theta) \geq u(\tilde{a}, \theta, \mu), \text{ for every } \theta \in \Theta.$$

This condition is always satisfied in the standard case (when the agent's utility does not directly depend on his belief), and every equilibrium with a naive agent is payoff-equivalent to a full information outcome. However, the full revelation results of Section 4.2 do not apply anymore. Indeed, observe that a profile of equilibrium strategies with a naive agent is equivalent to an equilibrium profile under exogenous memory failures when $\alpha \to 0$. Then, when the conditions of Assumption 1 are satisfied, there is no fully revealing equilibrium. A naive agent always selectively forgets some information. The equilibria with a naive agent are exactly the same as those characterized in Section 5.2 when $\alpha \to 0$, i.e., with $\theta_D = E_{\theta \sim \mu}(\theta)$.

6.2 Present-Biased Preferences

In this paper, we have assumed that Self 1 and Self 2 share a common utility function, so that incentives to distort beliefs are not due to divergence of interests between the selves. While our setting allows to take into account some self-control problems,[20] it does not cover self-control problems that arise when the agent's preferences exhibit time-inconsistency due to quasi-hyperbolic discounting. For example, in the memory management model of Bénabou and Tirole (2002), the utility function of Self 1 is $u_1(a, \theta) = a(-c + \delta\theta V)$ but the utility function of Self 2 is $u_2(a, \theta) = a(-c + \delta\beta\theta V)$, where $a = 0$ corresponds to exerting no effort, $a = 1$ corresponds to exerting effort in a project which has benefit V with probability θ and a cost

[20] For example, temptation and self-control problems à la Gul and Pesendorfer (2001) can be represented by a (common) psychological utility which has a functional form as described in Proposition 4: $u(a, \theta, \nu) = u_M(a, \theta) + \psi(a, \nu)$, where $u_M(a, \theta)$ is the utility experienced by the non-tempted side, and $\psi(a, \nu) = \max_{b \in A} E_{\tilde{\theta} \sim \nu} \left[u_T(b, \tilde{\theta}) - u_T(a, \tilde{\theta}) \right]$ is the cost of self-control faced by the tempted side (see, Lipnowski and Mathevet, 2018).

$c > 0$, δ is a standard discount factor, and $\beta < 1$ captures the salience of the present. There is a conflict of interest because Self 1 would like to exert effort if $\theta \geq \frac{c}{\delta V}$, whereas Self 2 would like to exert effort only if $\theta \geq \frac{c}{\beta\delta V}$. Therefore, when the state is in the interval $[\frac{c}{\delta V}, \frac{c}{\beta\delta V}]$, Self 1 would like Self 2 to have overoptimistic beliefs about θ.

Bénabou and Tirole (2002) study the equilibria of the memorization game between Self 1 and Self 2 assuming there are two states θ_L and $\theta_H > \theta_L$, and only the low state θ_L can be memorized. That is, following the notation of the current article, type θ_L can send messages m_L or $m_\emptyset$ but type θ_H can only send message $m_\emptyset$. Under these assumptions, a fully revealing equilibrium may not exist when β is small, i.e., when there is a strong conflit of interest between the two selves. This non-existence is due to the asymmetry of memorization abilities: if both types are able to memorize the state at zero cost, i.e., type θ_H is also able to send message m_H,[21] then, even if β is very small, there exists a fully revealing equilibrium with a belief after message $m_\emptyset$ that puts sufficient weight on the low type. Actually, because the utility functions are not belief-dependent, this memorization game is equivalent to a standard inter-personal disclosure game: whatever the number of states and the degree of conflict of interest between the two selves, the existence of a fully revealing equilibrium follows from the same argument as in Milgrom (1981), Seidmann and Winter (1997) or Hagenbach et al. (2014).[22]

The model of Bénabou and Tirole (2002) has been recently extended by Chew et al. (2020) by considering three states, a different memorization technology, and adding self image concerns. Now, the agent would like to hold optimistic beliefs both to keep self control and a good self image. Using the same argument as above, i.e., by considering skeptical beliefs whenever the agent has no memory, it is immediate to check that a fully revealing equilibrium exists if every state can be memorized and memorization is costless. To conclude, while our model focuses on belief-dependent utility functions by abstracting away from conflicts of interests between the two selves, some of our results easily extend to standard classes of time-inconsistent preferences. The analysis of intra-personal memorization or inter-personal disclosure with belief-based utilities and general forms of conflicts of interests is left for future research.

6.3 Memorization Costs

We have assumed that there is no direct memorization cost in order to focus on the incentives to memorize when it only affects beliefs and decisions. If memorization is costly, the agent strictly prefers not to memorize in states in which the equilibrium benefits of memorizing are not high enough. Because the benefits of memorizing are endogenous, memorization costs can affect the structure and properties of equilibrium outcomes. While the characterization of equilibria with memorization costs is beyond the scope of this paper, we discuss below the effect of such costs on equilibria assuming they replace the memorization failures studied above. In the classes of utility functions studied in Section 5.2, we illustrate that equilibrium outcomes would be similar

[21] This is the own-type certifiability condition discussed in Remark 2.

[22] Following the terminology of Hagenbach et al. (2014), this disclosure game satisfies "directional masquerade": if Self 1 would like to masquerade as type θ' when his actual type is θ, then $\theta' > \theta$.

to those obtained with exogenous memorization failures. We consider the same memorization game as in Section 3.3 but assume that for each state θ, the cost of sending message m_θ is $c > 0$. The cost of no memorization (sending message $m_\emptyset$) is zero.

Consider first the class of games studied in Section 5.2.1 with state-independent utility functions. The utility of the agent is $u(a^*(e), e)$ when the expected value of the state for Self 2 is e and he chooses the optimal action $a^*(e)$. Let $\theta_\emptyset$ be the equilibrium expected value of the state when no information is memorized. If Self 1 memorizes the state θ, then his utility is $u(a^*(\theta), \theta) - c$. If Self 1 does not memorize the state, then his utility is $u(a^*(\theta_\emptyset), \theta_\emptyset)$. Since $u(a^*(e), e)$ is increasing in e, the best response of Self 1 is a 1-threshold strategy: he memorizes the state iff the state is high enough, as in Proposition 6. In addition, the threshold is strictly above 0, i.e., there is no fully revealing equilibrium, because for every $\theta_\emptyset$ and $c > 0$, there exists $\theta > 0$ such that $u(a^*(\theta), \theta) - c < u(a^*(\theta_\emptyset), \theta_\emptyset)$.[23]

Next, assume that preferences are state-dependent as in Section 5.2.2. As in the proof of Proposition 7, if θ is small, it is possible that the material gain of memorizing is high compared to the psychological loss of doing so, so the agent strictly prefers to memorize even if it is costly. When θ is high, if the psychological benefit of holding high beliefs is large and the cost c small, the agent memorizes. For some intermediate states, memorizing has no impact on the decision and has a small psychological benefit, so the agent strictly prefers not to memorize if the cost is strictly positive. Hence, as in the second part of Proposition 7, extreme states are memorized but some intermediate states are not.

6.4 Comparison with Optimal Information Acquisition

In this paper we have assumed that disclosure of information occurs while Self 1 is already informed about the state. If instead he can commit to a disclosure strategy before learning the state, i.e., Self 1 chooses an information structure for Self 2, then the timing of the game would go as follows. First, Self 1 chooses his disclosure strategy σ_1. Second, the state $\theta \in \Theta$ is drawn according to the prior μ. Third, a message m is drawn according to $\sigma_1(\cdot \mid \theta)$ and revealed to Self 2. Finally, Self 2 chooses an action. When the disclosure strategy σ_1 is unconstrained (i.e., it is any function $\sigma_1 : \Theta \to \Delta(M)$), the timing above corresponds to a Bayesian persuasion problem (see Kamenica and Gentzkow, 2011). If in addition the agent is psychological, we are in the model studied in Lipnowski and Mathevet (2018). A natural interpretation of this timing in a multi-self game is that of an agent whose first self strategically decides, without being himself informed, which information to freely acquire for Self 2. Note that in the standard case in which the utility function u of the agent does not depend on the belief ν, the function $\max_{a \in A} U(a, \nu)$ is convex (it is the maximum of convex – linear – functions) and acquiring full information is the optimal ex-ante choice.

The optimal strategies of information acquisition ex-ante and of information disclosure interim are usually different (except in the standard case, in which full disclosure is optimal in

[23] Introducing costs in a disclosure game à la Milgrom (1981), Verrecchia (1983) also demonstrates the existence of such a 1-threshold equilibrium.

both settings). In particular, acquiring full information is ex-ante optimal if $U^*(\nu)$ is convex, and acquiring no information is ex-ante optimal if $U^*(\nu)$ is concave. Clearly, in our setting, it could be the case that the unique equilibrium is fully revealing or non-revealing independently of the concavity of $U^*(\nu)$. For instance, consider Example 1 with two states, identify ν to the probability of one of the state, and assume that $u(a, \theta, \nu) = u(\nu) = U^*(\nu)$ is strictly increasing. If u is strictly convexe, i.e., the agent is "psychologically information-loving" in the sense of Lipnowski and Mathevet (2018), then he acquires full information. In contrast, if u is strictly concave, the unique ex-ante optimal policy is to acquire no-information. In both cases, the unique equilibrium of the interim disclosure game is fully revealing.[24]

7 Conclusion

In this paper, we consider a psychological agent who has no recall of past information unless he actively decides to memorize it. When the memorization process never fails and the agent is sophisticated, it is possible for him to interpret no memory as bad news. Our results show that, for general classes of psychological preferences, this skepticism leads to voluntary memorization of all the information. In contrast, if the memorization process sometimes fails for exogenous reasons or if the agent has a form of naivety with respect to his own incentives to selectively memorize, then there is room for him to internally manipulate his beliefs. When the agent has self image concerns for example, we show that the agent memorizes only the good or the extreme news about himself.

There is clear evidence of selective memory in the psychological and economic literature (see for instance Baumeister, 2010 or Zimmermann, 2020).[25] However, little is known about the extent to which individuals have some form of metacognition as defined in Bénabou and Tirole (2002), that is, are aware that their partial memory may be the result of an internal protection strategy. The agents' degree of naivety in this respect seems hard to control in the lab but it seems possible to vary exogenously the possibility to memorize. In particular, in the lab, the states could be made more or less complex or there could be more or less disturbance around the subjects. If a subject is naive, our results establish that perfect memory is not an equilibrium whatever such variations. If a subject is sophisticated, the amount and kind of information he voluntarily memorizes depends on the difficulty to memorize.

[24]Gentzkow and Kamenica (2017) and Escudé (2020) combine strategic information acquisition and information disclosure in the standard sender-receiver context in which agents are not psychological, but they have different preferences. The combinaison of information acquisition and information disclosure in psychological games is left for future research.

[25]Recently, several papers have studied more generally the link between the functioning of human memory and various biases in decision making and beliefs formation: Mullainathan (2002), Gennaioli and Shleifer (2010), Baliga and Ely (2011), Bordalo, Gennaioli, and Shleifer (2020) and Enke, Schwerter, and Zimmermann (2020).

A Appendix

To prove the results of Section 4.2 we construct fully revealing equilibria with extremal beliefs in the sense that the agent's belief ν off the equilibrium path (when he receives message $m_\emptyset$) is degenerate, i.e., $\nu = \delta_{\hat{\theta}}$ for some $\hat{\theta} \in \Theta$. In addition, we select an optimal action $a^*(\theta) \in \arg\max_{a \in A} U(a, \delta_\theta)$ for the agent when his belief is δ_θ. The following lemma provides a necessary and sufficient condition for the existence of a fully revealing equilibrium with such extremal beliefs and selection $a^*(\cdot)$.

Lemma 1 *Under the selection $a^*(\cdot)$, there exists a fully revealing equilibrium with extremal beliefs iff there exists $\hat{\theta} \in \Theta$ such that*

$$U^*(\delta_\theta) := u(a^*(\theta), \theta, \delta_\theta) \geq u(a^*(\hat{\theta}), \theta, \delta_{\hat{\theta}}), \text{ for every } \theta \in \Theta.$$

Proof. The proof directly follows from the definition of an equilibrium, the selection $a^*(\theta) \in \arg\max_{a \in A} U(a, \delta_\theta)$, $\theta \in \Theta$, and the restriction to extremal beliefs. ∎

Proof of Proposition 1. Let $\hat{\theta} \in \arg\min_{\theta \in \Theta} u(a^*(\theta), \delta_\theta)$. Then, $U^*(\delta_\theta) \geq u(a^*(\hat{\theta}), \delta_{\hat{\theta}})$ for every $\theta \in \Theta$, so there exists a fully revealing equilibrium by Lemma 1. ∎

Proof of Proposition 2. Let $\hat{\theta}$ be such that $\min_{\tilde{\theta}} u(a, \theta, \delta_{\tilde{\theta}}) = u(a, \theta, \delta_{\hat{\theta}})$ for all a, θ. Then, $U^*(\delta_\theta) = u(a^*(\theta), \theta, \delta_\theta) \geq u(a^*(\hat{\theta}), \theta, \delta_\theta) \geq u(a^*(\hat{\theta}), \theta, \delta_{\hat{\theta}})$, where the first inequality comes from the fact that $a^*(\theta) \in \arg\max_{a \in A} u(a, \theta, \delta_\theta)$, and the second inequality from the assumption of the proposition and the definition of $\hat{\theta}$. Hence, there exists a fully revealing equilibrium by Lemma 1. It is immediate to check that the result applies to separable utilities and directional utilities. In the separable case we have $\hat{\theta} \in \arg\min_{\theta \in \Theta} \psi(\delta_\theta)$. In the directional case we have $\hat{\theta} = \min\{\theta : \theta \in \Theta\}$. ∎

Proof of Proposition 3. Consider first the case where $w \in [0, 1)$. Let $\hat{\theta} \in \arg\min_{\theta \in \Theta} h(a^*(\theta), \theta)$. Then, $U^*(\delta_\theta) = (1 - w)h(a^*(\theta), \theta) + wh(a^*(\theta), \theta) \geq (1 - w)h(a^*(\theta), \theta) + wh(a^*(\hat{\theta}), \hat{\theta}) \geq (1 - w)h(a^*(\hat{\theta}), \theta) + wh(a^*(\hat{\theta}), \hat{\theta})$, where the first inequality comes from the fact that $h(a^*(\theta), \theta) \geq h(a^*(\hat{\theta}), \hat{\theta})$ for every $\theta \in \Theta$, and the second inequality from the fact that $a^*(\theta) \in \arg\max_{a \in A} h(a, \theta)$. Hence, there exists a fully revealing equilibrium by Lemma 1. The proof is similar for the case where $w \in (-1, 0)$ letting $\hat{\theta} \in \arg\max_{\theta \in \Theta} h(a^*(\theta), \theta)$. ∎

Proof of Proposition 4. Define the following continuous function $v : \Theta \times \Theta \to \mathbb{R}$:

$$v(\theta', \theta) = u_M(a^*(\theta'), \theta) + \psi(a^*(\theta'), \delta_{\theta'}).$$

Self 1 of type θ wants to induce beliefs θ' iff $v(\theta', \theta) > v(\theta, \theta)$. This defines a binary relation on Θ. We want to show that this relation has a minimal element, that is, that there exists $\hat{\theta} \in \Theta$ such that $v(\theta, \theta) \geq v(\hat{\theta}, \theta)$ for every θ, and hence from Lemma 1 there exists a fully revealing equilibrium. From Lemma 2 and Theorem 2 in Hagenbach et al. (2014), to show that

this minimal element exists it suffices to show that $v(\theta', \theta)$ has increasing differences in (θ', θ). For every $\theta'' \geq \theta'$ we have

$$v(\theta'', \theta) - v(\theta', \theta) = u_M(a^*(\theta''), \theta) - u_M(a^*(\theta'), \theta) + \psi(a^*(\theta''), \delta_{\theta''}) - \psi(a^*(\theta'), \delta_{\theta'}).$$

From the assumption that a^* is increasing in θ and $u_M(a, \theta)$ has increasing differences in (a, θ), we get that $u_M(a^*(\theta''), \theta) - u_M(a^*(\theta'), \theta)$ is increasing in θ. Hence, $v(\theta'', \theta) - v(\theta', \theta)$ is increasing in θ, i.e., $v(\theta', \theta)$ has increasing differences in (θ', θ). ∎

Proof of Proposition 5. Consider a fully revealing strategy. By Assumptions 1.1, 1.3 and 1.4, $a^*(\overline{e})$ is the unique sequentially rational action of the agent when he receives message $m_\emptyset$. Full revelation constitutes an equilibrium iff

$$u(a^*(\overline{e}), \theta, \overline{e}) \leq U^*(\delta_\theta) = \max_{a \in A} u(a, \theta, \theta), \text{ for all } \theta \in \Theta.$$

By Assumption 1.4, this condition implies that for every θ in a small enough neighborhood of $\overline{e}$ we have

$$u(a^*(\overline{e}), \theta, \overline{e}) \leq u(a^*(\theta), \theta, \theta). \tag{6}$$

Under Assumptions 1.2 and 1.4 we can apply the envelop theorem and get:

$$\frac{du(a^*(e), \overline{e}, e)}{de}\Big|_{e=\overline{e}} = \frac{\partial u(a^*(\overline{e}), \overline{e}, e)}{\partial e}\Big|_{e=\overline{e}},$$

which is non-zero by Assumption 1.5. Since u is continuously differentiable (Assumption 1.2) we also have

$$\frac{du(a^*(e), \theta, e)}{de}\Big|_{e=\overline{e}} \neq 0,$$

for every θ close enough to $\overline{e}$. For such $\theta < \overline{e}$ (if the derivative is strictly positive) or $\theta > \overline{e}$ (if the derivative is strictly negative), we then have $u(a^*(\overline{e}), \theta, \overline{e}) > u(a^*(\theta), \theta, \theta)$, which contradicts the equilibrium condition (6). ∎

Proof of Proposition 6. Consider any memorization strategy for Self 1, and let $\theta_\emptyset = E[\theta | m = m_\emptyset]$ be the expected value of the state when no information is memorized. Note that $\theta_\emptyset \in (0, 1)$. For every $\theta \in [0, 1]$, the utility of the agent is $u(a^*(\theta_\emptyset), \theta_\emptyset)$ when the state is not memorized, and $u(a^*(\theta), \theta)$ when the state is memorized. Let $\Delta(\theta \mid \theta_\emptyset) = u(a^*(\theta_\emptyset), \theta_\emptyset) - u(a^*(\theta), \theta)$. For every θ, the unique best response of Self 1 is to memorize if $\Delta(\theta \mid \theta_\emptyset) < 0$ and not to memorize if $\Delta(\theta \mid \theta_\emptyset) > 0$. By assumption, $\Delta(\theta \mid \theta_\emptyset)$ is strictly decreasing. We also have $\Delta(0 \mid \theta_\emptyset) > 0 > \Delta(1 \mid \theta_\emptyset)$. Hence, an equilibrium consists in a 1-threshold strategy for Self 1: $\sigma_1(\theta) = m_\emptyset$ if $\theta < \theta^*$ and $\sigma_1(\theta) = m_\theta$ if $\theta > \theta^*$, where $\theta^* \in (0, 1)$ is the unique (interior) solution in θ of $\Delta(\theta \mid \theta_\emptyset) = 0$, i.e., $\theta^* = \theta_\emptyset$. We also have

$$\theta_\emptyset = E[\theta | m = m_\emptyset] = \frac{\alpha F(\theta^*) E[\theta | \theta < \theta^*] + (1 - \alpha)\overline{e}}{\alpha F(\theta^*) + (1 - \alpha)},$$

where $\overline{e} = E_{\theta\sim\mu}(\theta)$ is the prior expected value of θ. Hence, $\theta^* = \theta_D$, where $\theta_D < \overline{e}$ solves Equation (1). The existence and uniqueness of a solution $\theta_D \in (0, \overline{e})$ to this equation follows from the proof of Proposition 1 in Jung and Kwon (1988), based on the model of Dye (1985). We rewrite the previous equation as follows:

$$\alpha F(\theta_D)\big(\theta_D - E[\theta|\theta < \theta_D]\big) = (1-\alpha)(\overline{e} - \theta_D). \tag{7}$$

At $\theta_D = 0$, the RHS of (7) is $(1-\alpha)\overline{e} > 0$ while the LHS is zero. At $\theta_D = \overline{e}$, the RHS is zero and the LHS is strictly positive. Both sides of the equality are continuous in θ_D, the RHS is decreasing in θ_D and the LHS can be rewritten as $\alpha\Big(\int_0^{\theta_D} F(x)dx\Big)$, which is increasing in θ_D. Therefore, there exists a unique solution of Equation (7), with $0 < \theta_D < \overline{e}$. ∎

Proof of Proposition 7. Consider any memorization strategy for Self 1, and let $\theta_\emptyset = E[\theta|m = m_\emptyset]$ be the expected value of the state when no information is memorized. Note that $\theta_\emptyset \in (0,1)$. For every $\theta \in [0,1]$, the utility of the agent is $u_M(a^*(\theta_\emptyset), \theta) + \psi(\theta_\emptyset)$ when the state is not memorized, and $u_M(a^*(\theta), \theta) + \psi(\theta)$ when the state is memorized. Denote the difference by

$$\Delta(\theta \mid \theta_\emptyset) = u_M(a^*(\theta_\emptyset), \theta) - u_M(a^*(\theta), \theta) - \psi(\theta) + \psi(\theta_\emptyset).$$

Given the assumptions above, we have:

- $\Delta(1 \mid \theta_\emptyset) < 0$;
- If $a^*(\theta_\emptyset) = 0$ (i.e., $\theta_\emptyset < \overline{\theta}$) then $\Delta(\theta \mid \theta_\emptyset)$ is strictly decreasing in θ;
- If $a^*(\theta_\emptyset) = 1$ (i.e., $\theta_\emptyset > \overline{\theta}$) then $\Delta(\theta \mid \theta_\emptyset)$ is strictly quasi-concave in θ.

Hence, the equilibrium memorization strategy is either

(i) a 1-threshold equilibrium θ^*, with $\theta^* \in (0,1)$, $\sigma_1(\theta) = m_\emptyset$ if $\theta < \theta^*$ and $\sigma_1(\theta) = m_\theta$ if $\theta > \theta^*$, or,

(ii) a 2-threshold equilibrium $(\underline{\theta}^*, \overline{\theta}^*)$, with $0 < \underline{\theta}^* < \overline{\theta}^* < 1$, $\sigma_1(\theta) = m_\emptyset$ if $\theta \in (\underline{\theta}^*, \overline{\theta}^*)$ and $\sigma_1(\theta) = m_\theta$ if $\theta \notin (\underline{\theta}^*, \overline{\theta}^*)$.

In case (i) we have $\Delta(\theta \mid \theta_\emptyset) > 0$ for all $\theta \in (0, \theta^*)$, so we must have $\Delta(0 \mid \theta_\emptyset) \geq 0$. In addition, $\Delta(\theta \mid \theta_\emptyset) < 0$ for all $\theta > \theta^*$, so θ^* is the unique (interior) solution θ of $\Delta(\theta \mid \theta_\emptyset) = 0$, i.e., $\theta^* = \theta_\emptyset$. As in the Proof of Proposition 6, $\theta^* = \theta_D$ is the unique solution in $(0, \overline{e})$ of Equation (1). We conclude that there exists is a 1-threshold equilibrium iff $\Delta(0 \mid \theta_D) \geq 0$, i.e., $u_M(0,0) - u_M(a^*(\theta_D), 0) \leq \psi(\theta_D) - \psi(0)$. This inequality is always satisfied if $a^*(\theta_D) = 0$, i.e., $\theta_D < \overline{\theta}$. Otherwise, if $\theta_D > \overline{\theta}$ it is satisfied iff $u_M(0,0) - u_M(1,0) < \psi(\theta_D) - \psi(0)$.

Consider now case (ii), with $u_M(0,0) - u_M(a^*(\theta_D), 0) > \psi(\theta_D) - \psi(0)$. We have $\theta_\emptyset \geq \overline{\theta}$, $\Delta(\theta \mid \theta_\emptyset) > 0$ for all $\theta \in (\underline{\theta}^*, \overline{\theta}^*)$ and $\Delta(\theta \mid \theta_\emptyset) < 0$ for all $\theta \notin [\underline{\theta}^*, \overline{\theta}^*]$. So we must have $\Delta(0 \mid \theta_\emptyset) < 0$, $\Delta(\underline{\theta}^* \mid \theta_\emptyset) = \Delta(\overline{\theta}^* \mid \theta_\emptyset) = 0$ and $\theta_\emptyset \in \{\underline{\theta}^*, \overline{\theta}^*\}$.

Note that $\Delta(\overline{\theta} \mid \theta_\emptyset) = -\psi(\overline{\theta}) + \psi(\theta_\emptyset) > 0$, so we have $0 < \underline{\theta}^* < \overline{\theta} < \overline{\theta}^* = \theta_\emptyset < 1$. We also have $\theta_\emptyset = E[\theta | m = m_\emptyset]$, so

$$\overline{\theta}^* = \frac{\alpha\big(F(\overline{\theta}^*) - F(\underline{\theta}^*)\big)E[\theta|\underline{\theta}^* < \theta < \overline{\theta}^*] + (1-\alpha)\overline{e}}{\alpha\big(F(\overline{\theta}^*) - F(\underline{\theta}^*)\big) + (1-\alpha)} < \overline{e}.$$

We rewrite this equality and, together with the condition $\Delta(\underline{\theta}^* \mid \overline{\theta}^*) = 0$, the couple $(\underline{\theta}^*, \overline{\theta}^*)$ should solve Equations (2) and (3).

Let's first examine Equation (2). The LHS of Equation (2) can be rewritten as follows:

$$\alpha\Big(\big(F(\overline{\theta}^*)-F(\underline{\theta}^*)\big)(\overline{\theta}^*-\underline{\theta}^*)-\int_{\underline{\theta}^*}^{\overline{\theta}^*} x dF(x)+\big(F(\overline{\theta}^*)-F(\underline{\theta}^*)\big)\underline{\theta}^*\Big) = \alpha\Big(\int_{\underline{\theta}^*}^{\overline{\theta}^*} F(x)dx+\big(F(\overline{\theta}^*)-F(\underline{\theta}^*)\big)\underline{\theta}^*\Big),$$

which is strictly increasing in $\overline{\theta}^*$. At $\overline{\theta}^* = \overline{\theta}$ we have $\underline{\theta}^* = \overline{\theta}$ from (3), so the LHS of Equation (2) is zero. The RHS of Equation (2) is strictly decreasing in $\overline{\theta}^*$, strictly positive at $\overline{\theta}^* = \overline{\theta}$ and zero at $\overline{\theta}^* = \overline{e}$. Hence, we conclude that for every admissible value of $\underline{\theta}^*$, Equation (2) has a unique solution $\overline{\theta}^* \in (\overline{\theta}, \overline{e})$.

The LHS of Equation (2) is strictly decreasing in $\underline{\theta}^*$, while the RHS is constant in $\underline{\theta}^*$. Hence, from the observations of the previous paragraph, the solution $\overline{\theta}^*$ of Equation (2) is strictly increasing in $\underline{\theta}^*$, and is equal to θ_D at $\underline{\theta}^* = 0$.

Consider now Equation (3). For every $\underline{\theta}^*$, the solution $\overline{\theta}^*$ is unique and strictly quasi convexe in $\underline{\theta}^*$ because the LHS of Equation (3) is strictly quasi convexe in $\underline{\theta}^*$ and ψ is strictly increasing. In addition, from $u_M(0,0) - u_M(a^*(\theta_D), 0) > \psi(\theta_D) - \psi(0)$, we deduce that the LHS of Equation (3) is strictly below $\psi(\theta_D)$ at $\underline{\theta}^* = 0$, which implies, together with the fact that ψ is strictly increasing, that the solution $\overline{\theta}^*$ of Equation (3) is strictly above θ_D at $\underline{\theta}^* = 0$. Finally, at $\underline{\theta}^* = \overline{\theta}$, the solution is $\overline{\theta}^* = \overline{\theta}$.

We conclude from the properties of Equations (2) and (3) above that there exists a unique solution $(\underline{\theta}^*, \overline{\theta}^*)$ satisfying the properties required for the existence of a 2-threshold equilibrium, i.e., such that $0 < \underline{\theta}^* < \overline{\theta} < \overline{\theta}^* < \overline{e} < 1$. ■

Example 9 Let $\Theta = A = \{0, 1\}$, and $u(a, \theta, \nu) = u_M(a, \theta) + \psi(a, \nu)$, where $u_M(a, \theta)$ is given by the following table

	$\theta = 0$	$\theta = 1$
$a = 0$	0	0
$a = 1$	1	-1

and

$$\psi(a, \nu) = \begin{cases} -w\nu^2 & \text{if } a = 0 \\ -w(1-\nu)^2 & \text{if } a = 1, \end{cases}$$

where $w \in (1,2)$. We have $U(0,\nu) = -w\nu^2$ and $U(1,\nu) = 1 - 2\nu - w(1-\nu)^2$, so, since $w > 1$,

$$\arg\max_{a \in A} U(a,\nu) = \begin{cases} \{0\} & \text{if } \nu < 1/2 \\ \{1\} & \text{if } \nu > 1/2, \end{cases}$$

and any action (or mixed action) is optimal for $\nu = 1/2$. Assume by way of contradiction that there is a fully revealing equilibrium with belief $\nu \in [0,1]$ off the equilibrium path.

(i) If $\nu < 1/2$, then Self 2 plays action $a = 0$ when Self 1 deviates from full memorization to no memorization, and hence Self 1 with type $\theta = 1$ does not deviate iff $U^*(\delta_1) \geq u(0,1,\nu)$, i.e., $-1 \geq 0 - w\nu^2$, which is equivalent to $\nu \geq \frac{1}{\sqrt{w}}$, which is impossible because $w < 4$.

(ii) If $\nu > 1/2$, then Self 2 plays action $a = 1$ when Self 1 deviates from full memorization to no memorization, and hence Self 1 with type $\theta = 0$ does not deviate iff $U^*(\delta_0) \geq u(1,0,\nu)$, i.e., $0 \geq 1 - w(1-\nu)^2$, which is equivalent to $\nu \leq 1 - \frac{1}{\sqrt{w}}$, which is impossible because $w < 4$.

(iii) If $\nu = 1/2$, then Self 2 is indifferent between action $a = 0$ and $a = 1$ when Self 1 deviates from full memorization to no memorization. Let $\alpha \in [0,1]$ be the probability that Self 2 plays action $a = 1$ after no memorization. Self 1, with type $\theta = 1$ and $\theta = 0$ respectively, does not deviate iff $U^*(\delta_1) \geq \alpha u(1,1,1/2) + (1-\alpha)u(0,1,1/2)$ and $U^*(\delta_0) \geq \alpha u(1,0,1/2) + (1-\alpha)u(0,0,1/2)$, i.e., $1 - \frac{w}{4} \leq \alpha \leq \frac{w}{4}$, which is impossible because $w < 2$. Hence, there is no fully revealing equilibrium, even if we allow Self 2 to use mixed strategies. ◇

References

Sandeep Baliga and Jeffrey C. Ely. Mnemonomics: The sunk cost fallacy as a memory kludge. *American Economic Journal: Microeconomics*, 3:35–67, 2011.

Pierpaolo Battigalli and Martin Dufwenberg. Belief-dependent motivations and psychological game theory. *Journal of Economic Literature*, forthcoming, 2020.

Roy F. Baumeister. The self. In *Advanced Social Psychology: The State of Science*. Oxford University Press, 2010.

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

Roland Bénabou and Jean Tirole. Incentives and prosocial behavior. *American economic review*, 96(5):1652–1678, 2006.

Roland Bénabou and Jean Tirole. Identity, morals and taboos: Beliefs as assets. *Quarterly Journal of Economics*, 126(2):805–855, 2011.

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

Roland Bénabou, Armin Falk, and Jean Tirole. Narratives, imperatives, and moral reasoning. *mimeo*, 2019.

Pedro Bordalo, Nicola Gennaioli, and Andrei Shleifer. Memory, attention and choice. *The Quarterly Journal of Economics*, 135(3):1399–1442, 2020.

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

Andrew Caplin and John Leahy. Wishful thinking. *working paper*, 2019.

Soo Hong Chew, Wei Huang, and Xiaojian Zhao. Motivated false memory. *Journal of Political Economy*, 128(10):3913–3939, 2020.

Olivier Compte and Andrew Postelwaite. Confidence-enhanced performance. *American Economic Review*, 94(5):1536–57, 2004.

Ronald A Dye. Disclosure of nonproprietary information. *Journal of accounting research*, pages 123–145, 1985.

Benjamin Enke, Frederik Schwerter, and Florian Zimmermann. Associative memory and belief formation. *working paper*, 2020.

Matteo Escudé. Communication with partially verifiable endogenous information. *mimeo*, 2020.

Erik Eyster and Matthew Rabin. Cursed equilibrium. *Econometrica*, 73(5):1623–1672, 2005.

John Geanakoplos, David Pearce, and Ennio Stacchetti. Psychological games and sequential rationality. *Games and economic Behavior*, 1(1):60–79, 1989.

Nicola Gennaioli and Andrei Shleifer. What comes to mind. *Quarterly Journal of Economics*, 125(4):1399–1433, 2010.

Matthew Gentzkow and Emir Kamenica. Disclosure of endogenous information. *Economic Theory Bulletin*, 5(1):47–56, 2017.

Daniel Gottlieb. Will you never learn? self deception and biases in information processing. *working paper*, 2010.

Sanford J. Grossman. The informational role of warranties and private disclosure about product quality. *Journal of Law and Economics*, 24:461–483, 1981.

Faruk Gul and Wolfgang Pesendorfer. Temptation and self-control. *Econometrica*, 69(6):1403–35, 2001.

Jeanne Hagenbach, Frédéric Koessler, and Eduardo Perez-Richet. Certifiable pre-play communication: Full disclosure. *Econometrica*, 82(3):1093–1131, 2014.

Nina Hestermann, Yves Le Yaouanq, and Nicolas Treich. An economic model of the meat paradox. *European Economic Review*, 129:103569, 2020.

David Huffman, Collin Raymond, and Julia Shvets. Persistent overconfidence and biased memory: Evidence from managers. *working paper*, 2019.

Philippe Jehiel. Analogy-based expectation equilibrium. *Journal of Economic Theory*, 123(2): 81–104, 2005.

W. Jung and Y. Kwon. Disclosure when the market is unsure of information endowment of managers. *Journal of Accounting Research*, 26:146–153, 1988.

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

Botond Kőszegi. Emotional agency. *The Quarterly Journal of Economics*, 121(1):121–155, 2006.

Ziva Kunda. The case fo motivated reasoning. *Psychological Bulletin*, 108(3):480–490, 1990.

Elliot Lipnowski and Laurent Mathevet. Disclosure to a psychological audience. *American Economic Journal: Microeconomics*, 10(4):67–93, 2018.

Yusufcan Masatlioglu, A Yesim Orhun, and Collin Raymond. Intrinsic information preferences and skewness. *mimeo*, 2019.

Paul Milgrom. Good news and bad news: Representation theorems and applications. *Bell Journal of Economics*, 12:380–391, 1981.

Paul Milgrom and John Roberts. Relying on the information of interested parties. *Rand Journal of Economics*, 17(1):18–32, 1986.

Sendhil Mullainathan. A memory-based model of bounded rationality. *The Quarterly Journal of Economics*, 117(3):735–774, 2002.

Charlotte Saucet and Marie-Claire Villeval. Motivated memory in dictator games. *Games and economic Behavior*, 117:250–275, 2019.

Daniel J. Seidmann and Eyal Winter. Strategic information transmission with verifiable messages. *Econometrica*, 65(1):163–169, 1997.

Robert Verrecchia. Discretionary disclosure. *Journal of Accounting and Economics*, 5:179–194, 1983.

Florian Zimmermann. The dynamics of motivated beliefs. *American Economic Review*, 110(2): 337–361, 2020.