WZB

Wissenschaftszentrum Berlin
für Sozialforschung

Jeanne Hagenbach
Frédéric Koessler

Selective Memory of a Psychological Agent

Discussion Paper

SP II 2021–201

February 2021

WZB Berlin Social Science Center

Research Area
Markets and Choice

Research Unit
Market Behavior

Wissenschaftszentrum Berlin für Sozialforschung gGmbH
Reichpietschufer 50
10785 Berlin
Germany
www.wzb.eu

Jeanne Hagenbach, Frédéric Koessler
Selective Memory of a Psychological Agent

Affiliation of the authors:

Jeanne Hagenbach
Sciences Po Paris – CNRS

Frédéric Koessler
Paris School of Economics – CNRS

Wissenschaftszentrum Berlin für Sozialforschung gGmbH
Reichpietschufer 50
10785 Berlin
Germany
www.wzb.eu

Abstract

Selective Memory of a Psychological Agent

by Jeanne Hagenbach and Frédéric Koessler*

We consider a single psychological agent whose utility depends on his action, the state of the world, and the belief that 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 identify broad categories of psychological utility functions for which there exists an equilibrium in which every state is voluntarily memorized. In contrast, if there are exogenous failures in the memorization process, then the agent memorizes states selectively. In this case, we characterize the partially informative equilibria for common classes of psychological utilities. If the material cost of forgetting is low, then the agent only memorizes good enough news. Otherwise, only extreme news are voluntarily memorized.

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

JEL classification: C72; D82

* E-mail: hagenbach.jeanne@gmail.com, frederic.koessler@psemail.eu.
The authors thank Martin Dufwenberg, Toomas Hinnosaar, Juan Ivars, Philippe Jehiel, Yves Le Yaouanq, Elliot Lipnowsky, Peter Schwardmann and Ina Taneva for helpful comments. They are also grateful to the (virtual) seminar participants at Sciences Po, Paris School of Economics, the University of Munich, the University of Nottingham, the University of Arizona, the WZB and the Virtual Seminar in Economic Theory. Jeanne Hagenbach thanks the European Research Council (grant 850996 –MOREV) for financial support. She is very grateful to the WZB for welcoming her in 2021 as the K.W. Deutsch Professor. Frédéric Koessler acknowledges the support of the ANR (Investissements d'Avenir, ANR-17-EURE-001, and StratCom, ANR-19-CE26-0010-01).

Wissenschaftszentrum Berlin für Sozialforschung gGmbH
Reichpietschufer 50
10785 Berlin
Germany
www.wzb.eu

1 Introduction

When evolving in an uncertain environment, individuals form beliefs about it with the primary objective to take the most appropriate decisions. For example, before buying a product or applying to a job, agents try to get an idea of the product quality or of their fit to the job. When considering only this *instrumental* value of beliefs, accuracy is beneficial and more information is welcome. However, recent research in behavioral economics, earlier research in psychology, as well as introspection suggest that beliefs formation is not uniquely driven by the desire for accuracy because individuals may attribute an *intrinsic* value to what they believe. They may for example prefer holding a less true but more optimistic view of the world they are in. Customers may value per se the beliefs that their purchases are not detrimental for the environment. Job applicants may like to think, at least temporarily, that they will be hired and their future is bright. 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.

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 game in which the agent is initially informed about the state of the world and selectively decides which states to memorize for his later self, who otherwise has not recall. While the process of memorizing information can be a cognitive process, it can also take the form of tangible actions that individuals commonly use to remember things: make reminding notes, repeat statements loud, put some event in context, let some papers in evidence, etc. In all these cases, the memorization process may fail for exogenous reasons and we incorporate the possibility that the agent is sometimes unable to print information in his mind even if he would like to. Considering both perfect or imperfect memorization processes, we link the information that the agent voluntarily remembers to his preferences over beliefs and the form of his psychological utility.

In our game, memorization is strategic and the equilibrium beliefs of the second self are determined by Bayesian updating. It means that the second self does not only learn the information that the first self decides to memorize, but can also make inferences from the absence of memory. In particular, when the agent is practically able to memorize everything, he can skeptically attribute no memory to bad news. In a first set of results, we show that, because of this skepticism and for broad categories of psychological utilities, there exists an equilibrium in which every state (type of self 1) is fully memorized by the agent. In this equilibrium, there is no self-manipulation and the agent acts under complete information. In contrast, if we consider exogenous memorization failures, there is no such equilibrium anymore. Intuitively, some information is always voluntarily forgotten by the agent because he can now attribute, at least partially, this lack of memory to his inability to memorize in general and not only to bad news. Overall, one main message of our work is that memory is always selective when it is exogenously imperfect.

Formally, we consider two temporal selves of a single agent who are involved in a disclosure

game. In the first stage of the game, self 1 is privately informed about the state of the world and can decide either to disclose this state to self 2 or not to disclose it. After seeing the message, self 2 forms a posterior belief about the state and eventually takes an action. When a state is disclosed, self 2 learns it and we interpret disclosure as voluntary memorization of this state by the agent. The common psychological utility shared by the two selves is a function of the state and the action, but also of the posterior belief about the state. In this game, an equilibrium is fully revealing if self 1 discloses every state to self 2, that is, if every state is voluntarily memorized. Such an equilibrium exists if and only if there exists a belief that, if held by self 2 in the absence of disclosure, induces a payoff for self 1 that is lower than the payoff he would get under complete information.

We show that a fully revealing equilibrium exists for four broad categories of psychological utilities, namely *state-independent utility*, *separable utility*, *anticipatory utility*, and a class of utilities satisfying an increasing differences property. The psychological utility is state-independent when the agent's well-being is only affected by his action and his beliefs about the state, not by the actual value of the state. An example can be found in Hestermann, Le Yaouanq, and Treich (2020) in which the state corresponds to the level of suffering of animals used to produce meat. The agent is directly affected by how much meat he consumes and by his perception of 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 purely belief-based utility (i.e., 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 agent depends on his moral type. In addition, the agent intrinsically cares about the perception of his own morality. The third category 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. Finally, the fourth category of psychological utility can be written as the sum of a standard material utility and a function of the action and belief of the agent, under some monotonicity and increasing differences assumptions. For example, an extension of the above example by Hestermann et al. (2020) in which the agent's utility is directly affected by the state belongs to this class. In all these cases, the agent can have perfect memory despite the fact that he can selectively memorize and thereby influence the beliefs that he intrinsically values. The results rely on the fact that, for these categories of psychological utilities, self 2 can skeptically interpret no memory as coming from a type of self 1 that no other type would like to masquerade as.

Next, we extend the model by assuming that the first self may be, with a positive probability, unable to disclose the state for exogenous reasons. Going back to the memory interpretation, it means that the memorization process sometimes fails. In this model, an equilibrium is fully revealing when self 1 memorizes the state whenever he is practically able to do so. Assuming that the state space is continuous and that the marginal effect of the agent's belief on his utility is not negligible, we show that there is no fully revealing equilibrium anymore. To understand this impossibility result, note that a self 2 who has no memory now partially attributes this

fact to self 1 being potentially unable to memorize. It follows that there is always some type of self 1 who, by not memorizing, can manipulate his beliefs in a beneficial way. This belief manipulation has a first-order positive effect on his utility but a second-order negative effect of taking a suboptimal action. With exogenous memorization failures, the agent has wiggle room to manipulate his beliefs because even a sophisticated self 2 cannot be fully skeptical when having no memory.

Finally, we illustrate selective memory by characterizing the partially revealing equilibria in the following class of problems. The type space is an interval of the real line. The agent takes a binary action, with a high action being adapted to high types and the low action being adapted to low types. The agent's utility increases with his expectation of his type, as would for instance be the case if the type is 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 the threshold, or (ii) a unique pair of thresholds such that the agent voluntarily memorize 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 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 their 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 remembered. We then have an equilibrium with two thresholds in which only intermediary types are forgotten. If the reverse is true, low types are forgotten and the equilibrium has one threshold.

2 Related Literature

We contribute to the literature on motivated beliefs, recently growing in behavioral economics and surveyed in Bénabou and Tirole (2016). Our paper provides new elements to the study of both the demand and supply of such beliefs. On the side of the motivation behind internal beliefs manipulation, we model an agent whose beliefs enter directly his utility function.[1] We consider a general form of belief-based utility which includes as particular cases the functions used in two closely related papers, Kőszegi (2006) and 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, we use a model of memory-management inspired by Bénabou and Tirole (2002). We modify this model to allow for richer action and state spaces, general psychological utility functions, and a symmetric treatment of all states in terms of disclosure options. We also add the feature that the agent may, with some exogenous probability, be unable to disclose

[1] The idea of utilities depending directly on players' beliefs dates back to psychological game theory, pioneered by Geanakoplos, Pearce, and Stacchetti (1989) and recently surveyed in Battigalli and Dufwenberg (2020). Note however that we focus on the agent's beliefs about the state of the world and do not consider his beliefs about any action.

the state even if he wants to. By doing so, we get closer to the theoretical literature on strategic information disclosure developed by Milgrom (1981), Grossman (1981), Dye (1985), Seidmann and Winter (1997) and Hagenbach, Koessler, and Perez-Richet (2014) among others. This literature examines the link between the players' conflits of interests and the amount of information revealed in equilibrium with a focus on the possibility of unraveling.

Bénabou and Tirole (2002) and Hestermann et al. (2020) consider two different motivations for the agent to manipulate his beliefs, but a common multiself game of internal disclosure. Hestermann et al. (2020) investigates the paradox that agents eat meat while caring about the state of animal suffering. The agent's utility function is directly affected by his meat consumption and by his expectation of that state, which falls into our category of state-independent psychological utility. In Bénabou and Tirole (2002), the agent has self-control problems and may overcome them by forming motivating beliefs about his ability.[2] In both papers, the state space is binary and the first self can disclose information to the second self who otherwise has no recall.[3] This disclosure game has the property that the first self has asymmetric disclosure options in both states: a bad state can be disclosed or not disclosed, whereas no disclosure is the only option for the good state. In contrast, we consider that every state can be either disclosed or not disclosed.[4] In Bénabou and Tirole (2002) and Hestermann et al. (2020), if the good state could be separated from the bad state, the issue of self deception would be simpler because there would always exist a fully revealing equilibrium. These authors link the propensity to selectively remember to the parameters of the utility function, such as the cost and valuation of meat, or the degree of time inconsistency. We investigate whether perfect memory can be voluntary in broader categories of psychological utilities.

Kőszegi (2006) is another closely-related paper which models strategic communication between an informed sender and a decision-maker. The players can be interpreted at two selves of an agent because they share a common utility.[5] This utility belongs to our category of anticipatory utilities. Communication takes the form of cheap talk but also, in the last section of the paper, of hard information disclosure from a sender who is potentially uninformed about the state with some exogenous probability. This perturbation of the disclosure game on the sender side can initially be found in Dye (1985) and is theoretically equivalent to our exogenous memorization failures. Kőszegi (2006) characterizes the partially revealing equilibria for a particular form of anticipatory utility and sheds lights on the existence of equilibria which are similar to the two-thresholds equilibria derived in our Proposition 7. Lipnowski and Mathevet (2018) also consider a benevolent sender who discloses information to a psychological receiver,

[2]Carrillo and Mariotti (2000) also consider an agent with time-inconsistent preferences and show how that strategic ignorance can serve as a disciplining device.

[3]In the game theory literature, the issue of imperfect recall is studied in Piccione and Rubinstein (1997a) and Piccione and Rubinstein (1997b) for example.

[4]Chew, Huang, and Zhao (2020) extend Bénabou and Tirole (2002) to incorporate 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 a good news.

[5]The idea of a concerned expert who strategically reveals information to an agent can also be found in Caplin and Leahy (2004).

but the former designs ex-ante the information to be transmitted to the latter. We further discuss the link between their results and ours in Section 6.

More generally, the paper is related to various works which model the internal processes of beliefs formation. In all the strategic communication games mentioned above, beliefs are not chosen freely but determined in equilibrium. In contrast, in Brunnermeier and Parker (2005), the agent chooses his beliefs at the beginning of his life under no constraints. Once these beliefs are chosen, the agent behaves as a Bayesian in all subsequent periods. The agent has an anticipatory utility and trades off the benefit from anticipating a better future and the cost of taking suboptimal actions. In Caplin and Leahy (2019), the decision-maker is a wishful thinker who is free to choose his beliefs but incurs a cost from distorting beliefs away from the Bayesian framework. We do not consider any such costs or direct costs to disclose information.

The paper also relates to a recent literature in economics that tries to understand how the functioning of human memory impacts belief formation and decision making. On the one hand, the above-mentioned memory-management models study how selective memory can help form motivated beliefs. Some recent experiments, such as Zimmermann (2020) and Chew et al. (2020), provide empirical evidence of strategic forgetfulness when subjects receive negative feedbacks about their own performance or when they do mistakes in intelligence tests.[6] On the other hand, Mullainathan (2002), Gennaioli and Shleifer (2010), Baliga and Ely (2011) and Bordalo, Gennaioli, and Shleifer (2020) link exogenous memory limitations to various biases in decision-making. Bordalo et al. (2020) additionally models associative memory, the process by which an agent can remember past experiences when he faces similar new choices. They incorporate the fact that such a process is imperfect and depends on the particular features of the current choices, leading to an asymmetric retrieval of past signals. Enke, Schwerter, and Zimmermann (2020) provide experimental evidence on the role of associative memory for beliefs formation. We do not consider associative memory but show that the imperfectness of the memorization process affects individuals' ability to exercise selective recall.

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 period 2 and has a non-empty set of actions A, where A is a compact subset of an Euclidean

[6] In the context of dictator games, Saucet and Villeval (2019) similarly point at an asymmetric recall between past altruistic and selfish decisions. In the psychology literature, Kunda (1990) or 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.

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. 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. Equivalently, the model can be interpreted as a sender-receiver game in which the first player is a privately informed benevolent sender who voluntarily disclose 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

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

Note that in this case we have $u(a, \nu) = U(a, \nu)$.

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:

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

[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 when they hold the same belief about θ.

Masatlioglu, Orhun, and Raymond (2019) present an experiment which focuses on agents' intrinsic preference for information not only considering how much the information reduces uncertainty about the 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 of meat to consume in $A = \mathbb{R}_+$. 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 beliefs 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

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

Of course, 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:

$$u(a, \theta, \nu) = \theta r a - ca + w E_{\tilde{\theta} \sim \nu}(\tilde{\theta}),$$

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 $w > 0$. ◇

3.2.3 Anticipatory Utility

The third category of utility functions correspond to anticipatory utilities:

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

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 beliefs 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. ◇

Example 6 (Disappointment and elation) Battigalli and Dufwenberg (2020) propose a way to model the fact the agent's well-being may 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 $u(a,\theta,\nu) = (1-w)h(a,\theta) + wE_{\tilde{\theta}\sim\nu}[h(a,\tilde{\theta})]$. ◇

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, where for every θ, θ' in Θ we have $m_\theta \neq m_\emptyset$ and $m_\theta \neq m_{\theta'}$. Said differently, 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 $a \in A$.

In this game, 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 loud, trying to put it in context etc. Note that in the game we consider, self 1 is privately informed about the state when he acts and does *not* commit to an information disclosure strategy as in the Bayesian persuasion framework (Kamenica and Gentzkow, 2011, Lipnowski and Mathevet, 2018). In Section 6.2, we provide a more detailed discussion of the relation between these models.

A strategy for self 1 is a measurable 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 measurable 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 $m_\emptyset$.

[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 below for more details.

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

Psychological Perfect Bayesian Equilibrium A (psychological perfect Bayesian) equilibrium is a profile of strategies and a belief $(\sigma_1, \sigma_2, \nu)$ such that

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 $u(\sigma_2(m_\emptyset), \theta, \nu) \geq U^*(\delta_\theta)$.

4 When is Perfect Memory Voluntary?

Perfect memory is voluntary if self 1 discloses the state to self 2 whatever the 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 discloses 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 fonction 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. ◇

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

For expositional simplicity, we focus throughout the rest of this section on 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. Note that getting a fully revealing equilibrium under these assumptions is harder than without.

The following lemma provides a necessary and sufficient condition for the existence of a fully revealing equilibrium with extremal beliefs under the 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. ■

In the literature on disclosure games, the type $\hat{\theta}$ is typically called a worst-case type. This type, if believed by self 2 after $m_\emptyset$, gives any type of self 1 the worst payoff. 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. Proposition 1 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). Proposition 2 applies to all separable psychological utility functions (Section 3.2.2), and therefore applies to Examples 3 (Moral self-image)

and 4 (Stubborness). Proposition 3 applies to all anticipatory utility functions (Section 3.2.3). Finally, Proposition 4 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 $a^*(\theta)$ and an increasing difference condition on the material utility function. These conditions are satisfied in Examples 1, 2, 3, as well as in a modified version of former examples that does not fall into any of the category of utility functions presented in Section 3.2 (see Example 8).

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

Proof. 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. ■

Proposition 2 (Separable Utility) *If $u(a, \theta, \nu) = u_M(a, \theta) + \psi(\nu)$, then there exists a fully revealing equilibrium.*

Proof. Let $\hat{\theta} \in \arg\min_{\theta \in \Theta} \psi(\delta_\theta)$. Then, $U^*(\delta_\theta) = u_M(a^*(\theta), \theta) + \psi(\delta_\theta) \geq u_M(a^*(\theta), \theta) + \psi(\delta_{\hat{\theta}}) \geq u_M(a^*(\hat{\theta}), \theta) + \psi(\delta_{\hat{\theta}})$, where the first inequality comes from the fact that $\psi(\delta_\theta) \geq \psi(\delta_{\hat{\theta}})$ for every $\theta \in \Theta$, and the second inequality from the fact that $a^*(\theta) \in \arg\max_{a \in A} u_M(a, \theta)$. Hence, there exists a fully revealing equilibrium by Lemma 1. ■

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.*

Proof. 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 $\hat{\theta} \in \arg\max_{\theta \in \Theta} h(a^*(\theta), \theta)$. ■

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.*

Proof. 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 now 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). This completes the proof of the proposition. ■

In Example 9 in the appendix, the agent's utility function takes the form $u(a, \theta, \nu) = u_M(a, \theta) + \psi(a, \nu)$ and there is no fully revealing equilibrium. In this example, the increasing difference assumption of the proposition does not hold (but the other assumptions of Proposition 4 are satisfied). Proposition 4 however applies to the following example, which is a modified version of Examples 2 (Guilt from consumption) and 3 (Moral self-image).

Example 8 (Guilt from consumption and state-dependence) 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 correspond 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 however 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 also may 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. ◇

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 memorize the states 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 the worst-case type $\hat{\theta}$ when he faces $m_\emptyset$. However, there is an outcome-equivalent equilibrium with no message and belief 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.

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 a message that no other type has access to (the condition of *own type certifiability* in Hagenbach et al., 2014). An interpretation is that the agent would memorize that the state is in a low range but would not memorize the state more accurately. Formally, self 1 of type θ could disclose a message $m \in M(\theta)$ and a given message m would then provide evidence that the state is in $M^{-1}(m) := \{\theta \in \Theta : m \in M(\theta)\}$. The proofs of Propositions 1, 2 and 3 extend by considering the worst-case type $\hat{\theta}$ not only following the off path message $m_\emptyset$ but also following any off equilibrium path message m. In Proposition 1 for example, we would consider such a message m and 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 subset of Θ.

Remark 3 (Unicity of the fully revealing equilibrium) In this section, we have shown that, for the three broad classes of utility functions introduced in section 3.2 and for a category of utility functions of the form $u(a, \theta, \nu) = u_M(a, \theta) + \psi(a, \nu)$, there always exists a fully revealing equilibrium in which self 2 is informed about θ when acting. Without putting more structure on the forms of the utility functions we consider, some equilibria might not be fully revealing.[11] Instead, our focus is to contrast the possibility of voluntary perfect memory in equilibrium with its impossibility in the presence of exogenous memorization failures introduced in the next section. While the literature has focused on self-deception of agents with belief-based utility, our paper suggests that such agents' ability to self-deceive may rely on how good is their memory in the first place.

5 Selective Memory with 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 unable to memorize the state with probability $1 - \alpha(\theta) \in (0, 1)$. With the complementary probability $\alpha(\theta)$, self 1 is able to memorize the state if he wants to and we are back to the previous model in the limit case in which $\alpha(\theta) = 1$ for every $\theta \in \Theta$. Equivalently, with probability $1 - \alpha(\theta)$, the only message available to self 1 is $m_\emptyset$.[12] We assume

[11] 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.

[12] Another equivalent interpretation is that, for each $\theta \in \Theta$, $\alpha(\theta)$ is the probability that self 1 is initially informed about θ.

that the probability distribution on Θ conditional on self 1 being unable to memorize is well defined and is denoted by $\bar{\nu} \in \Delta(\Theta)$. Note that $\bar{\nu} = \mu$ if $\alpha(\cdot)$ is constant.

In this model, we say that an equilibrium is fully revealing if $\sigma_1(\theta) = 1$ for every type $\theta \in \Theta$ who is able to memorize. We first observe that if the utility of the agent is standard (it does not directly depend on his belief), then there is a fully revealing equilibrium, and all equilibria are payoff-equivalent.

Observation 1 Consider any function $\alpha(\cdot)$ 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 informed agent type θ is the complete information payoff $U^*(\delta_\theta)$.

Proof. Full revelation constitutes an equilibrium because for every $\theta \in \Theta$ the induced payoff of the agent is $U^*(\delta_\theta) \geq u(a, \theta)$ for every $a \in A$. All equilibria induce such a payoff for every type θ 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, the state space is "large", and the marginal effect of the agent's belief on his utility is not negligible, then there is no fully revealing equilibrium. Said differently, the agent always voluntarily forgets some information. To illustrate the idea, 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 the probability that self 1 is able to memorize is independent of θ: $\alpha(\theta) = \alpha \in (0, 1)$. 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 full disclosure is not an 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 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 memorize all the information applies much more generally under the following assumption.

Assumption 1 Let $\bar{e} = E_{\theta\sim\bar{\nu}}(\theta)$ be the expected state conditional on self 1 being unable to memorize.[13]

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,[14] 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 $\bar{e}$, and $a^*(e)$ is differentiable at $e = \bar{e}$;[15]

5. *Locally non-satiated psychological utility.* The derivative of $u(a, \theta, e)$ with respect to e at $(a, \theta, e) = (a^*(\bar{e}), \bar{e}, \bar{e})$ 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 $\bar{e}$ by forgetting the information. Forgetting the information changes the second-period action in a sub-optimal way but, if θ is close enough to $\bar{e}$, 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. Hence, some types close enough to $\bar{e}$ have an incentive to deviate from full disclosure. The formal proof is as follows.

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

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

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

$$u(a^*(\bar{e}), \theta, \bar{e}) \leq u(a^*(\theta), \theta, \theta). \tag{1}$$

[13] In particular, if $\alpha(\theta)$ does not depend on θ, then $\bar{e}$ is the prior expected state $E_{\theta\sim\mu}(\theta)$.

[14] 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.

[15] In the previous example, this assumption is satisfied if $c \neq E_{\theta\sim\bar{\nu}}(\theta)$.

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

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

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

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

for every θ close enough to $\bar{e}$. For such $\theta < \bar{e}$ (if the derivative is strictly positive) or $\theta > \bar{e}$ (if the derivative is strictly negative), we then have

$$u(a^*(\bar{e}),\theta,\bar{e}) > u(a^*(\theta),\theta,\theta),$$

which contradicts the equilibrium condition (1). ■

Note that Proposition 5 extends to the case in which self 1 can partially disclose 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 $\hat{\theta}$.

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. The agent's optimal decision (or, in case of multiplicity, any selection) only depends on his conditional expected state, and is denoted by $a^*(e)$. Finally, we assume that the memorization failures are state independent: $\alpha(\theta) = \alpha \in (0,1)$ for every θ.

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 of 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 disclosure 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 $(0, \bar{e})$ of the following equation:

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

Proof. Consider any disclosure strategy for self 1, and let $\theta_\emptyset = E[\theta | m = m_\emptyset]$ be the expected value of the state when no information is disclosed to self 2. 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 disclosed, and $u(a^*(\theta), \theta)$ when the state is disclosed. 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 disclose if $\Delta(\theta \mid \theta_\emptyset) < 0$ and not to disclose if $\Delta(\theta \mid \theta_\emptyset) > 0$. By assumption, $\Delta(\theta \mid \theta_\emptyset)$ is strictly increasing. 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

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

where $\bar{e} = E_{\theta \sim \mu}(\theta)$ is the prior expected value of θ. Hence, $\theta^* = \theta_D$, where $\theta_D < \bar{e}$ solves Equation (2). The existence and uniqueness of a solution $\theta_D \in (0, \bar{e})$ 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:

$$\alpha F(\theta_D)\big(\theta_D - E[\theta | \theta < \theta_D]\big) = (1-\alpha)(\bar{e} - \theta_D). \tag{3}$$

At $\theta_D = 0$, the RHS of (3) is $(1-\alpha)\bar{e} > 0$ while the LHS is zero. At $\theta_D = \bar{e}$, 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 (3), with $0 < \theta_D < \bar{e}$. ∎

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 always 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 discloses the state iff it is higher than the

ex-ante expected value of the state ($\theta_D = \bar{e}$); 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, self 2's beliefs are not distorted for states above the threshold. For states below the threshold, self 2's beliefs and actions are 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 $\bar{\theta} \in (0, 1)$ such that

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

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.[16] 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 disclosure strategy:*

1. *There exists an equilibrium characterized by a 1-threshold disclosure strategy iff* $u_M(0, 0) - u_M(a^*(\theta_D), 0) \leq \psi(\theta_D) - \psi(0)$. *The 1-threshold disclosure 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* $(0, \bar{e})$ *of the Equation (2).*

[16] 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 $\theta > \theta^p$

2. *If* $u_M(0,0) - u_M(a^*(\theta_D),0) > \psi(\theta_D) - \psi(0)$*, then there exists a unique equilibrium, characterized by a 2-threshold disclosure 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* $0 < \underline{\theta}^* < \bar{\theta} < \overline{\theta}^* < \bar{e}$ *solving the following equations:*

$$\alpha(F(\overline{\theta}^*) - F(\underline{\theta}^*))\big(\overline{\theta}^* - E[\theta|\underline{\theta}^* < \theta < \overline{\theta}^*]\big) = (1-\alpha)(\bar{e} - \overline{\theta}^*); \tag{4}$$

$$u_M(0, \underline{\theta}^*) - u_M(1, \underline{\theta}^*) + \psi(\underline{\theta}^*) = \psi(\overline{\theta}^*). \tag{5}$$

Proof. Consider any disclosure strategy for self 1, and let $\theta_\emptyset = E[\theta|m = m_\emptyset]$ be the expected value of the state when no information is disclosed to self 2. 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 disclosed, and $u_M(a^*(\theta), \theta) + \psi(\theta)$ when the state is disclosed. 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., $\theta_\emptyset < \bar{\theta}$) then $\Delta(\theta \mid \theta_\emptyset)$ is strictly decreasing in θ;
- If $a^*(\theta_\emptyset) = 1$ (i.e., $\theta_\emptyset > \bar{\theta}$) then $\Delta(\theta \mid \theta_\emptyset)$ is strictly quasi-concave in θ.

Hence, the equilibrium disclosure 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 $(0, \bar{e})$ of Equation (2). 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., $\theta_D < \bar{\theta}$. Otherwise, if $\theta_D > \bar{\theta}$ 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 $\theta_\emptyset \geq \bar{\theta}$, $\Delta(\theta \mid \theta_\emptyset) > 0$ 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 $\Delta(\bar{\theta} \mid \theta_\emptyset) = -\psi(\bar{\theta}) + \psi(\theta_\emptyset) > 0$, so we have

$$0 < \underline{\theta}^* < \bar{\theta} < \overline{\theta}^* = \theta_\emptyset < 1.$$

We also have $\theta_\emptyset = E[\theta | m = m_\emptyset]$, so

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

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 (4) and (5).

Let's first examine Equation (4). The LHS of Equation (4) 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 $\bar{\theta}^*$. At $\bar{\theta}^* = \bar{\theta}$ we have $\underline{\theta}^* = \bar{\theta}$ from (5), so the LHS of Equation (4) is zero. The RHS of Equation (4) is strictly decreasing in $\bar{\theta}^*$, strictly positive at $\bar{\theta}^* = \bar{\theta}$ and zero at $\bar{\theta}^* = \bar{e}$. Hence, we conclude that for every admissible value of $\underline{\theta}^*$, Equation (4) has a unique solution $\bar{\theta}^* \in (\bar{\theta}, \bar{e})$.

The LHS of Equation (4) is strictly decreasing in $\underline{\theta}^*$, while the RHS is constant in $\underline{\theta}^*$. Hence, from the observations of the previous paragraph, the solution $\bar{\theta}^*$ of Equation (4) is strictly increasing in $\underline{\theta}^*$, and is equal to θ_D at $\underline{\theta}^* = 0$.

Consider now Equation (5). For every $\underline{\theta}^*$, the solution $\bar{\theta}^*$ is unique and strictly quasi convexe in $\underline{\theta}^*$ because the LHS of Equation (5) 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 (5) is strictly below $\psi(\theta_D)$ at $\underline{\theta}^* = 0$, which implies, together with the fact that ψ is strictly increasing, that the solution $\bar{\theta}^*$ of Equation (5) is strictly above θ_D at $\underline{\theta}^* = 0$. Finally, at $\underline{\theta}^* = \bar{\theta}$, the solution is $\bar{\theta}^* = \bar{\theta}$.

We conclude from the properties of Equations (4) and (5) above that there exists a unique solution $(\underline{\theta}^*, \bar{\theta}^*)$ satisfying the properties required for the existence of a 2-threshold equilibrium, i.e., such that $0 < \underline{\theta}^* < \bar{\theta} < \overline{\theta}^* < \bar{e} < 1$. ■

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, meaning that it is always

the case that some low states are distorted upwards compared to the complete information case.

When the agent takes a payoff-relevant action which affects his (state-dependent) material utility, there might be a tradeoff between disclosing a low state or not disclosing it because when the agent does not disclose information he might take a suboptimal action. Note first that this tradeoff is necessarily associated to a suboptimal action for low states, that is, to a case in which no disclosure induces action $a = 1$. For low states, if the material cost from hiding information is small compared to the psychological benefit from hiding 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 from hiding information is high compared to the psychological benefit and we are in the case 2 of the proposition. The agent prefers to disclose information when the state is low ($\theta < \underline{\theta}^*$). For intermediate states, in particular for states around $\bar{\theta}$, the action has a small effect on the material utility, so the psychological benefit always dominates the material cost, and therefore the agent prefers to hide information. Finally, when the state is high ($\theta > \underline{\theta}^*$), hiding information has both a material and psychological cost for the agent, so he always disclose information.

In a 2-threshold equilibrium, 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 $\bar{\theta}$, 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 disclosed in a 2-threshold equilibrium does not only depend on α and on the distribution F of the states, but it also depends on the agent's utility function: the solution $(\underline{\theta}^*, \overline{\theta}^*)$ of Equations (4) and (5) 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 $\bar{\theta} = \frac{c}{r}$, $\bar{e} = \frac{1}{2}$ and

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

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 (4) and (5), which simplify to:

$$\alpha\left(\overline{\theta}^{*}+\frac{w\overline{\theta}^{*}-c}{r-w}\right)^{2}=2(1-\alpha)\left(\frac{1}{2}-\overline{\theta}^{*}\right); \tag{6}$$

$$\underline{\theta}^{*}=\frac{c-w\overline{\theta}^{*}}{r-w}. \tag{7}$$

When $\alpha \to 0$ we get $\underline{\theta}^{*}=\frac{c-\frac{w}{2}}{r-w}$ and $\bar{\theta}^{*}=1/2$. If α increases, then $\underline{\theta}^{*}$ increases and $\overline{\theta}^{*}$ decreases: as in a 1-threshold equilibrium, more information is disclosed 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 disclosing information increases for low types and therefore low types have less incentives to disclose 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 disclosing information increases for low types and therefore low types have more incentives to disclose information in order to implement the appropriate decision.

In Proposition 7, we provide a necessary and sufficient condition for the existence of a 1-threshold equilibrium and show that, if this condition is not satisfied, then there exists a 2-threshold equilibrium. In Example 3, we can further show that the 2-threshold equilibrium exists if and only if the solution of Equations (6) and (7) is such that $\frac{c}{r}\leq\overline{\theta}^{*}\leq\frac{c}{w}$. While this condition is satisfied when $\frac{c}{r}<\theta_{D}<\frac{c}{w}$ (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 the two figures below. On Figure 1, the parameters of Example 3 are $r=1$, $c=w=\frac{1}{3}$. A 1-threshold equilibrium exists iff $\alpha\leq\bar{\alpha}=\frac{3}{4}$, and a 2-threshold equilibrium always exists. On Figure 2, the parameters are $r=1$, $c=\frac{1}{3}$ and $w=\frac{4}{5}$. A 1-threshold equilibrium exists iff $\alpha\leq\bar{\alpha}_{1}=\frac{24}{49}$ or $\alpha\geq\bar{\alpha}_{2}=\frac{3}{4}$, and 2-threshold equilibrium exists iff $\alpha\geq\bar{\alpha}_{1}$. The two equilibria therefore co-exist for $\alpha\geq\bar{\alpha}_{2}$.

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) states that, when the informed agent observes the state with probability $\alpha<1$, there can exist an equilibrium with three zones similar to ours[17] 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 $w<r-2c$.

[17]Going back to our interpretation of Proposition 7 and to the patient-doctor relationship mentioned in Kőszegi (2006), the doctor discloses low states to the agent if concealing the information prevents the patient from realizing the gravity of his problem and from doing the right thing.

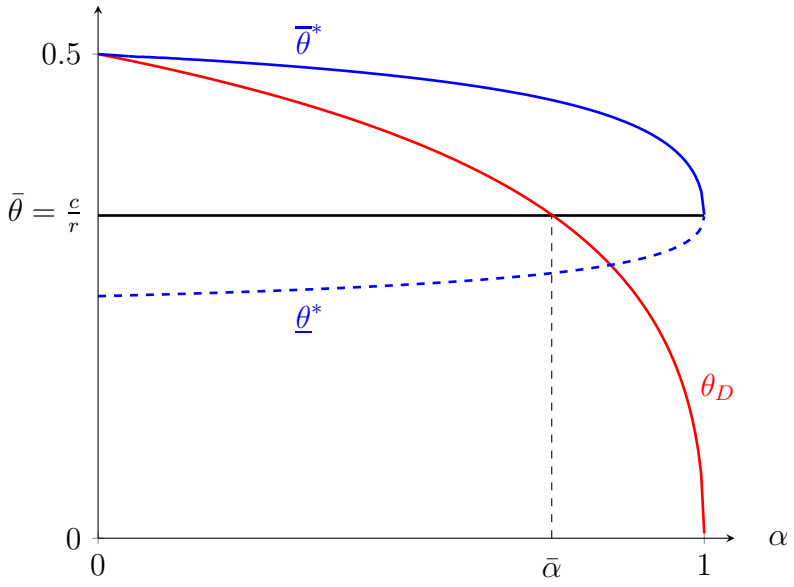

Figure 1: Equilibrium thresholds as a function of α for $r = 1$ and $c = w = \frac{1}{3}$.

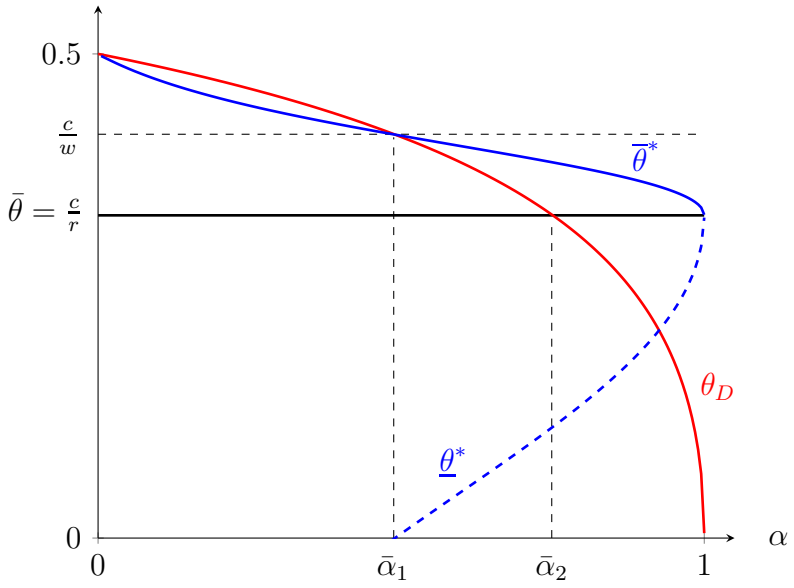

Figure 2: Equilibrium thresholds as a function of α for $r = 1$, $c = \frac{1}{3}$ and $w = \frac{4}{5}$.

6 Discussion

6.1 Naive Agent

In the first part of the paper, we have shown that there exist fully revealing equilibria for relatively broad categories of utility functions. These equilibria rely 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 always the prior μ when the message is $m_\emptyset$, even when message $m_\emptyset$ is sent by a strict subset of self 1's types. Such a naive agent has been considered by Milgrom and Roberts (1986). He also corresponds to a self 2 who is "fully cursed" in the sense of Eyster and Rabin (2005), or who is simplifying the disclosure 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 discloses information, but he does not know the probability that self 1 discloses 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 $\alpha(\theta) = \alpha \to 0$. Then, when the conditions of Assumption 1 are satisfied for $\bar{\nu} = \mu$, 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 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 θ. 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 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, there is no fully revealing equilibrium in the interim disclosure game.[18]

7 Conclusion

In this paper, we consider a psychological agent who has no recall of past information unless he actively decides to memorize it. In case 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). 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, we show that equilibrium memory depends on the difficulty to memorize. In particular, voluntary perfect memory is only possible when he is practically able to memorize any state.

Our model considers only one aspect of the overall functioning of memory, namely how information is memorized in the first place. When a state is memorized, we assume that it comes automatically to the mind of the agent later on. As mentioned in the introduction, some papers focus instead on the functioning of associative memory, namely how information memorized in the past comes back to mind in the present. We do not incorporate this process

[18] 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.

in our model. Note however that the imperfectness of the process by which information is passed on to the later self can be read as the first-self's inability to memorize as well as the second-self's inability to bring back past information to mind.

8 Appendix

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

$$U(0, \nu) = -w\nu^2 \text{ and } U(1, \nu) = 1 - 2\nu - w(1-\nu)^2,$$

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 disclosure to no disclosure, 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 disclosure to no disclosure, 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 disclosure to no disclosure. Let $\alpha \in [0, 1]$ be the probability that self 2 plays action $a = 1$ after no disclosure. Self 1, with type $\theta = 1$ and $\theta = 0$ respectively, does not deviate iff

$$U^*(\delta_1) \geq \alpha u(1, 1, 1/2) + (1 - \alpha)u(0, 1, 1/2) \text{ and } U^*(\delta_0) \geq \alpha u(1, 0, 1/2) + (1 - \alpha)u(0, 0, 1/2),$$

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. The supply of information by a concerned expert. *Economic Journal*, 114:487–505, 2004.

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

Juan D. Carrillo and Thomas Mariotti. Strategic ignorance as a self-disciplining device. *Review of Economic Studies*, 67(3):529–544, 2000.

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

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.

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

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.

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.

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

P. Milgrom and J. 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.

Michele Piccione and Ariel Rubinstein. On the interpretation of decision problems with imperfect recall. *Games and economic Behavior*, 20(1):3–24, 1997a.

Michele Piccione and Ariel Rubinstein. The absent-minded driver's paradox: Synthesis and responses. *Games and economic Behavior*, 20(1):121–130, 1997b.

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

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

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

Discussion Papers of the Research Area Markets and Choice 2021

Research Unit: **Market Behavior**

Jeanne Hagenbach and Frédéric Koessler SP II 2021-201
Selective Memory of a Psychological Agent

All discussion papers are downloadable:
http://www.wzb.eu/en/publications/discussion-papers/markets-and-choice