Value computation and value modulation: a dual-process theory of self-control *

Isabelle Brocas
University of Southern California and CEPR

Juan D. Carrillo
University of Southern California and CEPR

December 2013

Abstract

We develop a theory of self-control based on the dual process approach of Hare, Camerer and Rangel (2009). One brain system computes *the goal value of consumption and another brain system can* modulate *this value by transmitting information regarding the high-order considerations (e.g., healthiness or long term payoffs). We determine the optimal modulation and consumption strategy of the individual as a function of the cost of information transmission. The model has testable implications that depart form traditional analyses of temptation, self-control, hyperbolic discounting and cue triggered behavior. First, choices can be affected by distractors or other tasks orthogonal to the decision. Second, the likelihood of consuming a tempting good endogenously depends on environmental cues and, in particular, on how tempting the good is expected to be. And third, costly modulation can cause under-regulation (self-indulgence) but also over-regulation (self-restraint). The model can also shed light on some documented behavioral anomalies in the context of eating disorders, present-biased preferences and habit formation.*

*Address for correspondence: Department of Economics, University of Southern California, 3620 S. Vermont Ave., Los Angeles, CA 90089, USA, emails: <brocas@usc.edu> and <juandc@usc.edu>. We are grateful to seminar participants at the Society for Neuroeconomics Conference, Claremont Graduate University, Loyola Marymount University, Stanford University, the Toulouse Economics and Biology Workshop, the Stanford SITE workshop, the Amsterdam Symposium on Behavioral and Experimental Economics and the Behavioral Social Neuroscience Seminar Series at the California Institute of Technology for useful comments.

1 Introduction

Decision making of goods with short term and long term consequences has been an important research topic in economics and psychology. In recent years, scholars have proposed formal models to better understand time-inconsistencies, self-control problems and other forms of preference reversals frequently observed in the data.[1]

In parallel, neuroscientists and neuroeconomists have studied the neural correlates of temptation and short- vs. long-term choices. McClure et al. (2004, 2007) proposed a dual-valuation theory roughly consistent with the quasi-hyperbolic model, where one system weights immediate rewards and another weights all rewards. Kable and Glimcher (2007) challenged that approach and, instead, argued in favor of a single-valuation theory where one system responds to a combination of magnitude and delay of the reward. The more recent theories suggest that one system computes goal value and another system can modulate this value by exercising 'self-control' (Knoch and Fehr (2007), Hare et al. (2009), Luo et al. (2009), Figner et al. (2010), Hare et al. (2011)). In this case, exerting self-control means typically integrating high-order considerations –such as health or other long term consequences– in the decision.

The goal of this paper is to formalize the self-control idea described in this recent strand of neuroscience experiments and discuss its behavioral implications. For this, we build a model of decision-making that incorporates the two distinct processes emphasized in this literature, *value computation* and *value modulation*. The model closely follows the experimental setup by Hare, Camerer and Rangel (2009) (hereafter [HCR]). According to their theory, some regions of the ventromedial prefrontal cortex (vmPFC) perform 'value computation', and the consumption decision depends on this value. Ideally, this value computation integrates all the attributes of the good. The key issue is that information about the high-order attribute is not encoded in this region. Instead, it is encoded in some regions of the dorsolateral prefrontal cortex (dlPFC). Self-control is defined as a 'value modulation', that is, a transmission of information about the high-order attribute from dlPFC to vmPFC. This information transmission is costly but allows a more informed, hence better, decision. Using this dual-process model, we first determine the optimal modulation and choice strategy of the individual and then derive implications that depart from self-control theories of hyperbolic discounting, cue-triggered behavior or myopic vs.

[1] These include, but are not limited to, hyperbolic discounting theories (Strotz (1956), Prelec (1989), Laibson (1997), Carrillo and Mariotti (2000), etc.), decision theoretic models of temptation (Gul and Pesendorfer (2001), Dekel, Lipman and Rustichini (2009), etc.), cue-triggered theories of consumption (Laibson (2001), Bernheim and Rangel (2004)), and dual-self models of intertemporal choice (Thaler and Shefrin (1981), Fudenberg and Levine (2006), Brocas and Carrillo (2008), etc.).

farsighted systems.

The full characterization of the optimal information transmission and consumption constitutes the main theorem of the paper (Theorem 1). Technically, we show that information about the high-order attribute is not transmitted (that is, value is not modulated) when the realization is in a certain compact set H. When there is no modulation, value is computed assuming the expectation of the high-order attribute within that set. The conditions for no modulation, that is, the size and location of H, depend on the cost of attention, the relative importance of the high-order attribute, and perhaps more importantly the distribution from which that attribute is drawn.

This simple result has a number of new behavioral implications. First, the likelihood of value modulation decreases in the presence of distractors or other tasks orthogonal to the decision and increases with the help of reminders regarding health considerations or long term benefits (Corollary 1). This is in line with the recent experimental findings in social psychology (Vohs et al., 2008) and neuroeconomics (Hare et al., 2011) that link self-control to multi-tasking. Second, the environment affects the decision to modulate value signal and consume the good. For example, a highly negative health signal about a tempting good is more likely to be transmitted if the situation is such that the individual is not supposed to encounter such unhealthy goods than if he is expected to (Corollaries 2 and 3). This relates to the literature on cue triggered behavior, which argues that environmental cues affect choices either by changing preferences (Laibson, 2001) or producing errors (Bernheim and Rangel, 2004). A notable difference is that, in our model, the effect of cues in modulation and consumption is endogenous: the environment determines the expectation of the high-order attribute, which in turn influences the decision to transmit information and consume. Third, the theory predicts that costly value modulation always results in excessive consumption of very unhealthy goods but also in insufficient consumption of mildly unhealthy ones (Corollary 4). While over-consumption is a standard result in models of self-control, under-consumption is perhaps more surprising. It implies that over- and under-regulation are two sides of the same coin: the very same reasons that provoke a self-indulgent behavior can also be responsible for a self-restraint one.

Next, we use the main theorem to study some anomalies in choice. We start with eating disorders, and consider subjects with an incorrect perception of the distribution of health ratings. If an individual believes that all goods are unhealthy (as, for example, patients suffering from Anorexia Nervosa) or no good is unhealthy (as, for example, patients suffering from Bulimia Nervosa), his modulation decision will be compromised resulting in systematic under-consumption in the first case and systematic over-consumption in the second (Corollary 5). This contrasts with the case of correct perceptions in which over-

and under-consumption occur for the same individual depending on the health realization. We then use the model to think about decisions with long term consequences. More precisely and in line with Luo et al. (2009), we reinterpret the high-order consideration as the date where the health effect is suffered (instead of the level). It follows that an exponential discounting individual subject to costly value modulation will behave like a hyperbolic discounting individual with no cost of value modulation: goods with short-term negative consequences are over-consumed relative to time-consistency whereas goods with long-term negative consequences are under-consumed (Corollary 6). In other words, the dual process of signal valuation and signal modulation can provide a microfoundation for decreasing impatience, a behavioral anomaly long noted in psychology and economics (Strotz (1956); Ainslie (1975); Prelec (1989)). Finally, we analyze modulation and consumption of addictive substances. We model habit formation in the standard Becker and Murphy (1988) tradition: total and marginal utility of current consumption are decreasing and increasing in the level of past consumption, respectively. In this framework, we show that an addicted individual will be less likely to incur the cost of modulating the value signal due to his higher propensity to consume. This implies that ignorance of long term health considerations will be more prevalent for addictive substances, not because of self-delusion or other irrational motives but as the result of an optimal trade-off between the costs and benefits of signal modulation (Corollary 7).

The paper is organized as follows. In section 2, we present the basic neurophysiological dual process model of value modulation and value computation and derive the main result. In section 3, we discuss some implications of the theory: the effect of distractors and environmental cues on modulation and choice, and the possibility of under- and over-regulation. In section 4, we study in the context of the model some behavioral anomalies related to eating disorders, time-inconsistent choices and consumption of addictive substances. In section 5, we provide some final comments. Proofs of the main theorem and ensuing corollaries are relegated to the appendix.

2 The model

2.1 The basic decision problem

We propose a theoretical model of self-control based on the *multi-attribute* neurophysiological approach of [HCR]. According to their experimental evidence, value signals are encoded by one system (some regions of vmPFC), and this value dictates choices. A second system (some regions of dlPFC) can affect the final choices by modulating some attributes of the signal. More precisely, in the experiment an individual chooses whether to con-

sume a good with a basic attribute that has an immediate pleasurable effect (taste) and a high-order attribute that requires long term considerations (healthiness). If self-control is not exercised (i.e., value is not modulated), information about the high-order attribute is not transmitted and choice is dictated exclusively by taste. If self-control is exercised (i.e., value is modulated), then information about healthiness is transmitted and both attributes are taken into consideration when making a choice. In the paper, we put special emphasis in tightening our theory as close as possible to this neurobiological evidence.

To formalize the theory, we consider an individual choosing whether to consume ($z = 1$) or not consume ($z = 0$) a "tempting" good with a taste attribute $\theta \in [\underline{\theta}, \overline{\theta}]$ and a health attribute $h \in [\underline{h}, \overline{h}]$. The utility of the individual is given by:

$$\begin{cases} \theta - \alpha h & \text{if} \quad z = 1 \\ 0 & \text{if} \quad z = 0 \end{cases} \tag{1}$$

By definition of being tempting, the good is pleasurable ($\underline{\theta} \geq 0$) but unhealthy ($\underline{h} \geq 0$). The parameter α captures the importance of healthiness relative to taste in the decision. For example, an individual in a diet or with a physical condition will exhibit a high α. For the rest of the paper, we make the following parametric assumptions.

Assumption 1 *$\theta \sim U[0,1]$, $h \in [0,1]$ with continuous, differentiable and strictly log-concave p.d.f. $f(h)$, and $\alpha \in (0,1)$.*

There is of course some loss of generality in assuming a uniform distribution of taste. For tractability, we could keep a general distribution function for only one variable and opted for h because the health attribute of a good is more objectively measurable and therefore easier to manipulate experimentally than the taste attribute. Log-concavity is assumed for purely technical reasons. Finally, identical support for θ and h together with $\alpha \in (0,1)$ ensure an interior optimal solution, where consumption is always optimal if $\theta = 1$ and never optimal if $\theta = 0$.

2.2 Representation of information in the brain

We now model how the brain represents the information relevant to the decision. Following [HCR], we posit an interplay between two systems. The dlPFC encodes a signal regarding the high-order attribute, the healthiness h. It then decides whether to exert self-control, that is, to transmit that signal to the vmPFC or not. This transmission of information is called *value modulation*. The vmPFC encodes a signal regarding the basic attribute, the taste θ. It receives the information from dlPFC (or not) and determines the goal value of the good on the basis of the signals available. This aggregation of information is called

value computation. Under modulation, the goal value incorporates both the health and taste attributes (appropriately weighted) whereas under no modulation it includes only the taste attribute. The consumption decision is made as a function of this goal value. From now on, we will generically call $\mathcal{M}$ the system responsible for value modulation and $\mathcal{C}$ the system responsible for value computation, and defer to section 2.4 a more in-depth discussion of the brain regions involved.

As has been evidenced in the literature, signal modulation is not pervasive, which suggests that it must be costly to a certain extent. Indeed, modulation requires energy and attention to be spent in the coding and decoding of information and this energy is not freely available. To capture this feature, we introduce a cost of modulation c internalized by system $\mathcal{M}$. Signal modulation is also discriminative, that is, it depends on the realization of the health parameter. This means that system $\mathcal{M}$ is sophisticated enough to anticipate how system $\mathcal{C}$ computes the goal value, and uses this knowledge to decide whether to send the information.

The remaining (but crucial) issue is to determine the way in which system $\mathcal{C}$ interprets an absence of modulation. One possibility is a mechanistic approach where system $\mathcal{C}$ incorrectly takes at face value that no information implies no health concerns (formally, it assumes $h = 0$). Alternatively, one may argue that system $\mathcal{C}$ correctly 'infers' that no information transmission from system $\mathcal{M}$ indicates a certain level of the health parameter.[2] The first option, quite reasonable as a first approximation, is in fact unsatisfactory for theoretical, intuitive, neurophysiological and empirical reasons. From a theory standpoint, $h = 0$ is only one of many plausible "non-rational" ways to (not) incorporate information, and there is no solid ground for adopting this focal point rather than any other. From an intuitive viewpoint, it seems that inferences about the high-order attribute under no modulation should be different for a savory, unhealthy good (e.g., potato chips) than for an unsavory, healthy good (e.g., cod liver oil). In other words, the general properties of the good are always incorporated. From a neurophysiology standpoint and consistent with the previous argument, the evidence that will be reviewed in section 2.4 suggests that some information is updated and some inferences are made by system $\mathcal{C}$ even under no modulation. Last, from an empirical standpoint and as we will see in sections 3 and 4, predictions of a theory based on correct expectations match empirical observations better than those of a theory based on an incorrect face value interpretation of no evidence. For the rest of the paper, we will focus on the bayesian model with correct expectations under no modulation.

[2]Needless to say, both approaches are *as if* abstraction of the decision processes involved. It is in no way implied that systems $\mathcal{M}$ and $\mathcal{C}$ literally perform such unrealistic and sophisticated calculations.

Finally, note that contrary to the recent dual process literature (e.g., Fudenberg and Levine (2006), Brocas and Carrillo (2008) and Alonso et al. (2013)), in this model *there is no conflict between systems.* Both $\mathcal{M}$ and $\mathcal{C}$ care about representing the utility of the individual correctly. Some choices are suboptimal only because systems have different access to information and communication is costly. These features of the model are shared with the recent work by Cunningham (2013) on automatic vs. reflective judgment.

2.3 Modulation and choice

Under the "rational expectations" approach, system $\mathcal{C}$ computes goal value anticipating that no information transmission by system $\mathcal{M}$ has some informational content about the health realization. Formally, we use superscript "+" to denote modulation and "−" to denote no modulation. Under modulation, system $\mathcal{C}$ learns the health and taste realizations of the good and incorporates this information in the decision. The utility of consuming ($z = 1$) and not consuming ($z = 0$) are given by (1):

$$u_z^+(\theta, h) = \begin{cases} \theta - \alpha h & \text{if} \quad z = 1 \\ 0 & \text{if} \quad z = 0 \end{cases}$$

Under no modulation, inferences are more subtle. Denote by H the set of health realizations such that system $\mathcal{M}$ does not modulate the goal value, that is, does not transmit health information to system $\mathcal{C}$ (naturally, this set needs to be determined endogenously). Let $f(h \mid h \in H)$ be the revised probability distribution function when no information is transmitted: it is understood that $h \in H$ but the realization remains unknown. Also, let $E[h \mid h \in H]$ be the conditional expectation. Under no modulation, system $\mathcal{C}$ computes the goal value given the taste realization and the health inferred by the absence of information transmission. The expected utility of consuming ($z = 1$) and not consuming ($z = 0$) are now given by:

$$u_z^-(\theta) = \begin{cases} \theta - \alpha E[h \mid h \in H] & \text{if} \quad z = 1 \\ 0 & \text{if} \quad z = 0 \end{cases}$$

The optimal recommendation by system $\mathcal{C}$ under modulation (+) and no modulation (−) is based in each case in a comparison between the goal value of consuming and not consuming:

$$\begin{cases} z^+ = 1 \quad \text{iff} \quad u_1^+(\theta, h) > u_0^+ \quad \Leftrightarrow \quad \theta > \theta^+ \equiv \alpha h \\ z^- = 1 \quad \text{iff} \quad u_1^-(\theta) > u_0^- \quad \Leftrightarrow \quad \theta > \theta^- \equiv \alpha E[h \mid h \in H] \end{cases}$$

When system $\mathcal{M}$ decides between modulation and no modulation, it knows the health

rating of the good but not its taste.[3] However, it anticipates correctly how system $\mathcal{C}$ will represent the value of consumption and recommend a decision. Therefore, the values from its viewpoint of modulating and not modulating the signal given a health rating h are:

$$V^+(h) = \int_{\theta^+}^{1} (\theta - \alpha h) d\theta - c \qquad \text{and} \qquad V^-(h) = \int_{\theta^-}^{1} (\theta - \alpha h) d\theta$$

respectively, where c (> 0) is the attentional effort or cost of signal modulation. The condition for the optimality of modulation is $V^+(h) > V^-(h)$, which can be rewritten as:

$$\Delta(h, E[h \,|\, h \in H]) \;\equiv\; \int_{\alpha E[h|h\in H]}^{\alpha h} (\alpha h - \theta) d\theta \;>\; c$$

$$\Leftrightarrow \quad \Big| h - E[h \,|\, h \in H] \Big| > \eta(c, \alpha) \;\equiv\; \frac{\sqrt{2c}}{\alpha} \tag{2}$$

Fixing H, one can notice that the benefit of modulation, Δ, is convex in h and with a minimum at $E[h \,|\, h \in H]$:

$$\frac{\partial \Delta}{\partial h} = \alpha^2 \left(h - E[h \,|\, h \in H] \right) \gtreqless 0 \quad \text{iff} \quad h \gtreqless E[h \,|\, h \in H]$$

and

$$\frac{\partial^2 \Delta}{\partial h^2} = \alpha^2 \;>\; 0$$

This means that H, the set of values such that system $\mathcal{M}$ chooses no modulation, is necessarily compact. The main theorem of the paper consists in a complete characterization of the optimal modulation strategy as function of c, α and $F(\cdot)$.

Theorem 1 *The optimal modulation strategy is unique.*

- *When* $\eta \geq 1$, *then* $H = [0, 1]$.
- *When* $\eta \in [1/2, 1)$, *then* $H = [0, h^{**}]$ *if* $E[h] < 1 - \eta$,
  $H = [0, 1]$ *if* $E[h] \in [1 - \eta, \eta]$,
  $H = [h^*, 1]$ *if* $E[h] > \eta$.
- *When* $\eta \in (0, 1/2)$, *then* $H = [0, h^{**}]$ *if* $E[h|h < 2\eta] < \eta$,
  $H = [h^o, h^o + 2\eta]$ *if* $E[h|h < 2\eta] > \eta$ *and* $E[h|h > 1 - 2\eta] < 1 - \eta$,
  $H = [h^*, 1]$ *if* $E[h|h > 1 - 2\eta] > 1 - \eta$.

where (h^{**}, h^o, h^*) *are the solutions of:* $h^{**} - E[h|h \leq h^{**}] = \eta$, $E[h|h^o \leq h \leq h^o + 2\eta] = h^o + \eta$, *and* $E[h|h \geq h^*] - h^* = \eta$.

[3] This assumption matches the neuroscience evidence that vmPFC is involved in the computation of all types of values whereas dlPFC is involved in computation of abstract high-order attributes but not in value computation in general (see e.g. Plassmann et at. 2007, Baumgartner et al. 2011, Hutcherson et al. 2012 and the meta analysis by Clithero and Rangel 2013). Our model however does not hinge on this hypothesis and would also extend with minor modifications to the case where system $\mathcal{M}$ decides between modulation and no modulation under complete information of θ and h.

The first noticeable characteristic of the optimal modulation mechanism is that it is unique for any given c, α and $F(\cdot)$. This is a priori not obvious since it involves a fixed point argument: the benefit of modulation, Δ, depends on the set of health realizations for which there is no modulation, H, which itself depends on the benefit of modulation. The second feature of the mechanism is that modulation can be optimal for health realizations above a certain threshold h^{**} (transmit the health rating only when it is sufficiently bad) but also for the opposite case of health realizations below a certain cap h^{*} (transmit the health rating when it is not too bad) or even both for low and high realizations of health. The third general property is that the extent of modulation is inversely related to η, the cost of self-control and the relative importance of taste.[4] However, the main value of our theory is that it provides a unified framework with new and testable implications for issues related to self-control problems. These implications are explored in sections 3 and 4. But before, for the reader interested in the neurophysiology of the decision making process, we discuss the neural correlates of the modulation mechanism outlined in this section.

2.4 Neural correlates of value modulation and value computation

Our result suggests the existence of two different processes. One process represents a rough outline of the high-order features of the environment in which consumption takes place. It is characterized by a default costless representation. For example, in a bar, it captures an unhealthy environment, and it tends to trigger abstinence. It takes the form of a *rule-based* process. Another process represents the exact high-order characteristic of the item to consume. It is characterized by a costly computation leading to an efficient choice and can be viewed as a *cognitive* process. For a given modulation cost, the switch between the rule-based process and the cognitive process occurs when the rule is not appropriate and likely to generate a decision that departs substantially from the optimal one. We are therefore looking for: (i) system(s) that implement the rule-based process, (ii) system(s) that implement the cognitive process, and (iii) system(s) that operate the switch between the two.

Previous studies have shown that vmPFC, dlPFC and orbito-frontal cortex (OFC) are central actors of the goal value system (Hare et al. (2008), Plassmann et al. (2007)). Basic information about taste is computed in the mesolimbic regions involved in reward processing and sent to the vmPFC/OFC regions where the goal value is computed given the information received. As explained in section 2.2, dlPFC sometimes communicates costly information regarding the high-order health attribute to vmPFC, while the latter

[4]Formally, the size of the no-modulation set H is h^{**}, 2η or $1-h^{*}$. As shown in the proof of Theorem 1, $dh^{**}/d\eta > 0$ and $dh^{*}/d\eta < 0$.

encodes subjective taste and integrates the high-order attributes only when dlPFC sends that information (Hare et al. (2009)). Together, vmPFC and dlPFC playing the roles of systems $\mathcal{C}$ and $\mathcal{M}$ are part of the cognitive process.

In parallel, several studies document both the ability to incorporate environmental features into value and to switch from rule-based effortless to cognitive effortful processes, the two other key elements to implement our mechanism. Recent experiments have shown that cues and frames affect the activity pattern in value-related regions of the OFC (Plassman et al. (2008), De Araujo et al. (2005)). Also, the literature on habitual control shows that the habit system (mostly the basal ganglia) assigns through repeated exposure values to actions performed in specific contexts, and is capable of selecting actions yielding the highest values (Niv and Montague (2008), Daw and O'Doherty (2013)). These results provide a framework to explain how associations between external factors and actions are learned over time and how they are represented in each consumption episode. In particular, we should observe different activity patterns in OFC/vmPFC regions when subjects are asked to consume in different environments as a function of their prior experiences.

Other studies have identified key regions involved in rule setting and behavior adjustment. In particular, the medial frontal cortex (MFC), particularly the anterior cingulate cortex (ACC), is involved in motivation of behavior (Rushworth et al. (2007)) and is known to participate in action monitoring to prevent undesirable actions (Bush et al. (2000), Paus (2001), Ridderinkhof et al. (2004)). Posterior lateral PFC regions have been found to select decisions as a function of contextual information while medial lateral PFC regions motivate decisions based on the temporal episode in which decision-making takes place (Koechlin et al. (2003)). Also, MFC has been shown to be involved in representing rule-context associations to implement executive control in lPFC (Kouneiher et al. (2009)). Converging evidence supports the view that MFC and dlPFC support jointly cognitive control, with MFC responsible for monitoring performance and dlPFC responsible for adjusting behavior. Taren et al. (2011) show that those regions share a common hierarchical organization and are functionally connected at different levels. This is consistent with both the known anatomical connections between the two systems (Petrides (2005)) and the documented joint contributions to cognitive control (Botvinisk et al. (2001); Botvinik (2008); Kerns et al. (2004); Koechlin and Summerfield (2007); Koechlin et al. (1999, 2000)).

Taken together, the evidence suggests that a rule-based process is in place to represent contextual information and intrinsic motivation. The process features PFC/MFC regions constructing and representing high-order information that is passed to vmPFC/OFC regions. Furthermore systems involved in the rule-based process are functionally connected to those involved in the cognitive process used to adjust behavior according to task de-

mands. More precisely, dlPFC selects which information is sent: a contextual (default) information independent of the current episode through the costless rule-based process, or a precise information about the food to be consumed through the costly cognitive process. This general description is in line with recent studies showing that the decision to engage in cognitive control rather than rule-based control is driven by expectations about prospects and the eventual selection involves dlPFC (Dixon and Christoff (2012), Bahlmann et al (2012)). It is also consistent with our current understanding of how dietary choices are regulated via either the habitual control system that uses contexts/actions associations built from past experience or the more flexible goal-directed control system capable of representing future consequences more accurately (Rangel (2013)).

3 Behavioral implications

3.1 Pointers and distractors

One key element of our theory is that modulation requires costly attentional effort. In the absence of a cost and given the congruence of objectives between $\mathcal{M}$ and $\mathcal{C}$, system $\mathcal{M}$ would always modulate the signal and system $\mathcal{C}$ would always incorporate that information in computing goal value. It is obviously difficult for an experimenter to estimate the cost of modulation. On the other hand, it is possible to study experimentally the effect of *variations* of this cost. The exercise can be performed across individuals. For example, a higher IQ or a greater capacity to focus attention should translate into a lower cost of modulation. Perhaps more easily testable, comparisons can also be performed within individuals and across trials. In particular, the experimenter can increase the cost of self-control by showing some distractors or asking the subject to multi-task, and she can decrease the cost by presenting information related to the benefits of healthy food. The basic model has clear comparative statics predictions.

Corollary 1 - Manipulation of decisions. *Choices can be affected with the use of distractors and multi-tasking:* $\frac{d\eta}{dc} > 0$, $\frac{d\eta}{d\alpha} < 0$, $\frac{d^2\eta}{dc^2} < 0$, $\frac{d^2\eta}{d\alpha^2} > 0$ *and* $\frac{d^2\eta}{dc\,d\alpha} < 0$.

Recall from Theorem 1 that value modulation is inversely related to η. Therefore, $d\eta/dc > 0$ means that modulation is more prevalent the lower the cost of exercising it. This comparative statics is straightforward and consistent with the recent neuroexperimental evidence. Hare et al. (2011) shows that health cues makes it easier, that is, less effortful for the individual to integrate the health variable in the decision. The result holds both behaviorally (tasty but unhealthy snacks are avoided more often) and neurophysiologically (there is more value modulation as witnessed by stronger activation of dlPFC).

The prediction is also in line with recent research in social psychology on self-control in multitasking contexts. When engaged in both self-control and an unrelated cognitive task, there is competition between tasks for the scarce resources (such as glucose) which results in lower exertion of self-control (that is, less modulation) and also lower performance in the other task (Vohs et al. (2008); Masicampo and Baumeister (2008)).

Marginal effects can also be tested in a controlled setting by suitably selecting pointers and distractors. According to Corollary 1, there are increasing returns in lowering the cost of modulation ($d^2\eta/dc^2 < 0$). Also, as individuals become more and more concerned with the health component of the good, value modulation increases but at a decreasing rate ($d^2\eta/dcd\alpha < 0$). Finally, subjects on a diet or with a delicate physical condition engage more in value modulation ($d\eta/d\alpha < 0$) but, again, at a decreasing rate ($d^2\eta/d\alpha^2 > 0$). Overall, the result shows that tasks which are orthogonal to the consumption decision (pointers, distractors, irrelevant tasks) can alter the benefit of modulating the goal value and, through this channel, affect the final choices.

3.2 Environmental cues

Traditional models of self-control ignored the fact that the environment can affect the choices of individuals, an issue long recognized in the psychopharmacological literature, notably for addictive substances (Zinberg (1984); Falk and Feingold (1987); Caprioli et al. (2007)). Realizing this shortcoming, the economics literature has proposed more comprehensive models where the environment provides *cues* that either change the preferences of the individual (Laibson, 2001) or trigger mistakes (Bernheim and Rangel, 2004). In both cases, however, the effect of cues on the behavior of subjects is an exogenous feature of the model. Our model proposes an endogenous relationship between the environment and the likelihood of consumption.

Corollary 2 - Accounting for the environment. *The likelihood of consuming the tempting good depends not only on the health realization h but also on the distribution $F(h)$ from which it is drawn.*

One way of thinking about environmental cues is in terms of the information they provide about the distribution of the high-order attribute from which the good is drawn. For example, if I am offered a drink in a bar or a meal in a fast food parlor, chances are that the alcoholic and fat content of those goods are greater than if I am offered them at a friend's dinner party. Under this interpretation, the environment in our model has an *endogenous* effect on the behavior of subjects. Indeed, for the optimal transmission of the high-order signal, not only the health realization (alcohol and fat content) but also

whether the subject typically faces healthy or unhealthy goods in this environment (bar vs. friend's house) matter.[5] In our example, an alcoholic drink at a bar may trigger a different response than *the same* drink at a friend's party. This results from the fact that the communication strategy of system $\mathcal{M}$ depends on its expectation regarding health, which in turn affects the interpretation by system $\mathcal{C}$ when no evidence is transmitted.[6] Therefore, the realized health rating of a good may be communicated to system $\mathcal{C}$ in one environment and not in another, leading to different consumption decisions. Interestingly, these decisions are also accompanied by different neural responses, which means that the theory has new testable implications at two levels, behavioral and neural. The effect of the environment in the choice of subjects has further implications.

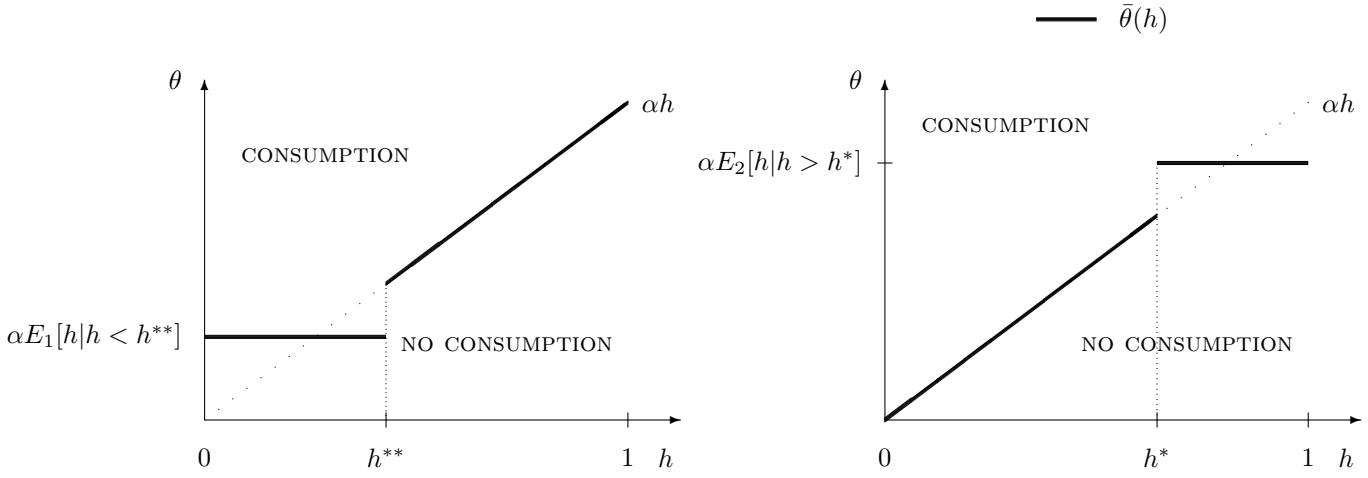
Corollary 3 - Modulation of the improbable. *Value modulation occurs when the realized health rating is unexpected: high if the expectation is low, low if the expectation is high, and high and low if the expectation is intermediate.*

According to this result, not only the amount but even the type of modulation depends on the distribution of the health rating. Formally, as $f(h)$ shifts towards higher values (in a Monotone Likelihood Ratio Property sense), the optimal strategy of system $\mathcal{M}$ shifts from modulation only at the top, to never modulation or modulation at the top and bottom, and finally to modulation only at the bottom. The intuition is simple. By understanding when system $\mathcal{M}$ modulates the signal, system $\mathcal{C}$ makes inferences regarding the health rating. If $\mathcal{M}$ communicates only bad (good) health ratings, then in the absence of modulation $\mathcal{C}$ assumes the rating is good (bad) and consumption is likely (not likely) to take place. Therefore, to prevent the individual from consuming highly unhealthy items, there are two possible strategies. System $\mathcal{M}$ communicates only bad ratings, in which case the individual is likely to consume unless the rating is disclosed. Or system $\mathcal{M}$ communicates only good ratings, in which case the individual is likely to not consume unless the rating is disclosed. When the item is drawn from an unhealthy distribution, the second strategy is optimal, as it requires to "pay" the modulation cost less often. By contrast, when the item is drawn from a healthy distribution, the first strategy dominates for the same cost saving reasons. Overall, since information transmission is costly it should be utilized when news are striking and unexpected. When news are expected, information is not transmitted, costs are spared and no-news is correctly interpreted as ordinary news.

[5] Formally, suppose that $H = [0, h^{**}]$. Fixing c and α, h^{**} solves $h^{**} - E[h|h \leq h^{**}] = \eta$ so it depends on $F(\cdot)$. The same is true for h^o and h^*.

[6] Notice the similarities between this result and the recent work on "rational inattention" (see e.g. Caplin and Martin (2011) and Caplin and Dean (2013)) where subjects decide whether to spend costly resources into incorporating signals as a function of the expected informational benefits.

The result is illustrated in Figure 1. The left graph corresponds to an environment where goods are typically healthy and the right graph to an environment where goods are typically unhealthy.[7] With no modulation cost, optimal consumption occurs whenever $\theta > \alpha h$. With a positive modulation cost c, consumption occurs when, for a given health h, taste is above the cutoff $\bar{\theta}(h)$, as represented by the bold line. In the healthy environment, there is modulation for bad ratings ($h > h^{**}$) followed by consumption when $\theta > \alpha h$ and no modulation for good ratings ($h < h^{**}$) followed by consumption when $\theta > \alpha E_1[h \,|\, h < h^{**}]$. In the unhealthy environment, there is modulation for good ratings ($h < h^{*}$) followed by consumption when $\theta > \alpha h$ and no modulation for bad ratings ($h > h^{*}$) followed by consumption when $\theta > \alpha E_2[h \,|\, h > h^{*}]$).



Mildly unhealthy goods ($F_1(h)$). Strongly unhealthy goods ($F_2(h)$).

Figure 1. Consumption as a function of taste and healthiness

In Corollary 1, we discussed a test of the effects of c and α on behavior. It is possible to induce variations on $F(h)$ to test this new aspect of the theory. For instance, one can start an experiment by presenting all the goods that are going to be offered to the subject, or the set of goods from which choices are randomly drawn. The experimenter can manipulate the "environment" by choosing whether the majority of goods are mildly unhealthy ($F_1(h)$) or highly unhealthy ($F_2(h)$). According to Corollaries 2 and 3, fixing the health of the current item, the amount and type of modulation– reflected both in the

[7] Formally, in the first environment $f_1(h)$ is such that $E_1[h \,|\, h < 2\eta] < \eta$ and in the second environment $f_2(h)$ is such that $E_2[h \,|\, h > 1 - 2\eta] > 1 - \eta$. There is a third environment (not shown in the figure) with $f_3(h)$ such that $E_3[h \,|\, h < 2\eta] > \eta$ and $E_3[h \,|\, h > 1 - 2\eta] < 1 - \eta$. In this environment, modulation occurs for $h \in [0, h^o] \cup [h^o + 2\eta, 1]$.

choices and the neural activity– should depend on the distribution.

In light of the literature previously reviewed (Kouneiher et al. (2009); Taren et al. (2011)), we conjecture that PFC/MFC regions make an initial assessment of the expected high-order information and gradually construct the representation of the contextual information unless dlPFC (system $\mathcal{M}$) judges necessary to switch to the cognitive process. When the initial assessment suggests the health rating is likely to be bad, dlPFC triggers the cognitive process only if the actual draw goes against that prediction, and lets the rule-based process compute and send contextual information to vmPFC (system $\mathcal{C}$) otherwise.

Finally, there is an interesting connection between these results and a strand of the literature that links self-control to intertemporal choices. As pointed out by Ainslie and Monterosso (2003) and formalized by Bodner and Prelec (2003) and Mijovic-Prelec and Prelec (2010), succumbing to temptation today may signal the individual his likelihood of succumbing in the future. If each decision has a small cost but the accumulated stream is large, then by bundling behavior (that is, mentally comparing never succumbing to always succumbing) the individual can achieve some self-control. Here we propose a different mechanism that links *indirectly* current choices to future opportunities even when decisions are not inter-temporally connected: the set of upcoming options, defined as the likelihood of encountering high vs. low unhealthy goods in the future, determines the distribution from which goods are drawn, which itself affects the current decision to modulate value signals and consume.

3.3 Imperfect self-regulation

Our next goal is to describe in more detail consumption choices. We take the perspective of an experimenter who observes decisions over repeated trials and would like to retrieve the consumption patterns of the individual. Recall that $\bar{\theta}(h)$ represents the taste cutoff above which the individual consumes (see Figure 1). The fact that $\bar{\theta}(h)$ is a step function is, in part, an artifact of our dichotomous model with binary consumption and binary choice of information transmission. To make the model more realistic, we now assume that the cost of modulation varies across trials, and this may not be directly observed by the experimenter. For example, one can argue that some days it requires less attention to modulate the signal than others simply because it is easier to focus your mind or there are fewer distractors present. Technically, a variable cost has the property of smoothing out $\bar{\theta}(h)$.[8] To capture this variability, we assume that c is stochastic and drawn from a

[8]The function would be also smoothed out if we included stochastic changes in preferences, variable consumption quantities z or variable information transmission technologies (e.g., signals that are more or

continuous distribution. This means that $\eta \equiv \sqrt{2c}/\alpha$ is also stochastic. Denote by $G(\eta)$ its cumulative distribution function and assume a sufficiently wide support $[0, \eta^*]$ (with $\eta^* \geq 1$) to ensure that modulation and no-modulation may both occur in equilibrium.[9] The realization of c is known to both systems before the modulation and consumption decisions so that, in each trial, the optimal strategy presented in Theorem 1 applies. From the perspective of the experimenter who does not observe c, the relevant variable is the *average* taste cutoff above which the individual consumes. This is computed by integrating over all possible values of c. We obtain the following result.

Corollary 4 - Self-indulgent and self-restraint behavior. *On average, the individual under-consumes low tempting goods and over-consumes high-tempting goods.*

An individual with no cost of modulation consumes the good optimally, that is, when $\theta > \alpha h$. With a positive and stochastic cost of modulation, the individual exhibits on average under-consumption of low-tempting goods and over-consumption of high-tempting goods. Formally, let $\bar{\theta}_c(h)$ be the expected taste cutoff above which the individual consumes (i.e., the analogue of $\bar{\theta}(h)$ when c is stochastic). The expected cutoff is increasing in h. Furthermore, there exist two values h_1 and h_2 with $0 < h_1 \leq h_2 < 1$ such that $\bar{\theta}_c(h) > \alpha h$ for all $h < h_1$ and $\bar{\theta}_c(h) < \alpha h$ for all $h > h_2$.[10]

The reason is simple and holds for any distribution $G(\cdot)$. When system $\mathcal{M}$ modulates the goal value, the individual consumes optimally. When $\mathcal{M}$ does not modulate the goal value, system $\mathcal{C}$ infers that health is in a certain set, and assumes the expectation within that set. Such expectation is necessarily above the true realization when h is sufficiently low and below the true realization when h is sufficiently high. Also, other things being equal, modulation is less frequent the higher the cost which results in a strictly increasing expected taste cutoff function $\bar{\theta}_c(h)$. Following our previous example, Corollary 4 implies that an individual in a bar will succumb too frequently to very desirable beverages with high alcoholic content whereas the same individual in a friend's party will refrain too often from consuming a light cocktail with moderate alcoholic content. Figure 2 illustrates the under- and over-consumption result when $\eta \sim U[0,1]$ and $f'(h) \leq 0$.

less precise). The results in this section would be similar under these alternative formulations.

[9] Recall that, by definition, modulation is always optimal when $\eta = 0$ and by Theorem 1 modulation is never optimal when $\eta \geq 1$. Also, since all that matters for our theory is the stochasticity of η, identical results are obtained when c is fixed but α varies.

[10] Under some conditions, $h_1 = h_2$ so that the cutoff separating under-consumption from over-consumption is unique. These include, among others, $\eta \sim U[1/2, 1]$ or $\eta \sim U[0, 1]$ and $f(h)$ either monotonic or symmetric around its inflection point (see appendix B for details).

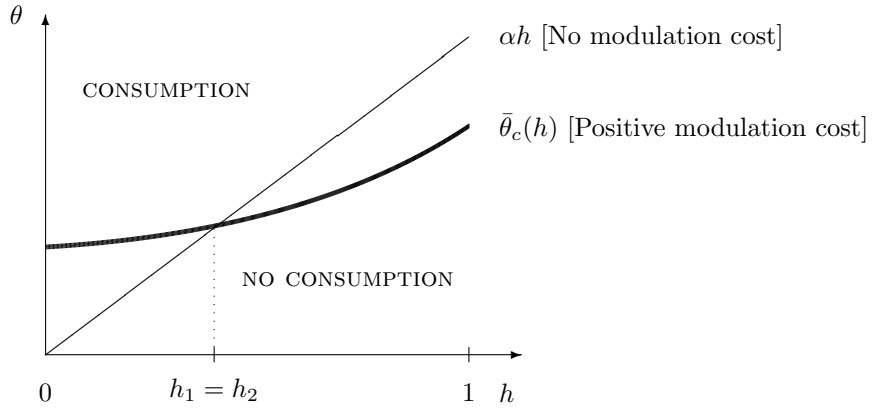


Figure 2. Expected consumption as a function of the health realization

Notice that traditional models of imperfect self-regulation emphasize the natural idea that subjects with self-control problems have a tendency to succumb into temptations and over-consume. Our theory has a novel implication. It suggests that the cost to modulate the goal value can result in either over-consumption or under-consumption depending on the taste and health realizations. In other words, these opposite deviations are in fact two sides of the same coin, so that the very same mechanism that induces self-indulgence can also result in self-restraint behavior.[11] The result also implies that consumption is more sensitive to the taste component of goods than it optimally should.

Remember that overconsumption arises when food items are tasty but very unhealthy because system $\mathcal{C}$'s assessment of the expected rating is better than the true rating. During this consumption event, the subject faces a tempting alternative and ends up succumbing to it. This case corresponds to the failure of willpower. Recent studies have shown that exertion of willpower is associated with the activation of dlPFC (Figner et al. (2010); Hare et al. (2009); Crockett et al. (2013)). In our case, dlPFC is not sending information to the goal value system when the rule-based process operates, which is in line with this evidence. Under-consumption arises when food items are healthy but not very tasty. In that case, $\mathcal{C}$'s assessment of the expected rating is not as good as the true rating, and the subject fails to consume an item that is good for his health. Consistently with the case of willpower failure, we conjecture that dlPFC is not activated when this occurs.

[11] The result has some resemblance to Benabou and Tirole (2004)'s model of compulsive behavior. The mechanisms, however, are very different since their theory relies on imperfect memory and an intertemporal conflict of preferences.

4 Behavioral anomalies

4.1 Modulatory dysfunctions and eating disorders

Our model makes predictions regarding modulation and behavior given beliefs about the health effects of the consumption good. In this section, we study self-regulation when individuals have incorrect *perceptions* of the health distribution. Formally, we assume that h is drawn from distribution $f(h)$ in $[0,1]$. However, the individual believes that it is drawn from distribution $p(h)$ with full support in $[0,1]$. We consider two polar cases for this (incorrect) perceived distribution: extremely healthy $p(h) = \underline{p}(h)$ such that $E_{\underline{p}(h)}[h] \to 0$ and extremely unhealthy $p(h) = \overline{p}(h)$ such that $E_{\overline{p}(h)}[h] \to 1$. Importantly, however, the true realization of h is correctly assessed by system $\mathcal{M}$ and, if transmitted, also by system $\mathcal{C}$.[12] As in section 3.3, we assume that attention costs vary across trials so that $\eta \sim G(\eta)$. We then study how the choice of modulation and the expected consumption given the true realization of health (i.e., the cutoff $\bar{\theta}_c(h)$) depend on the perception distribution $p(h)$. We have the following result.

Corollary 5 - Dysfunctions. *When $p(h) = \underline{p}(h)$, there is over-consumption for all $h \in (0,1]$. When $p(h) = \overline{p}(h)$, there is under-consumption for all $h \in [0,1)$.*

According to this result, an individual with excessively positive prior beliefs about the health rating of goods will over-consume them, even when the health realization is not as good as expected whereas an individual with excessively negative prior beliefs will under-consume them even when the health realization is not as bad as expected. Formally and as graphically depicted in Figure 3, when $p(h) = \underline{p}(h)$ then $\bar{\theta}_c(h; \underline{p}(h)) < \alpha h$ for all $h \in (0,1]$ and when $p(h) = \overline{p}(h)$ then $\bar{\theta}_c(h; \overline{p}(h)) > \alpha h$ for all $h \in [0,1)$.

It is crucial to realize that over- and under-consumption are not trivially driven by the fact that goods are always (incorrectly) seen as fully healthy and fully unhealthy, respectively. In fact and as we can see form Figure 3, an individual who believes that the environment is extremely healthy exhibits first-best consumption only when the realization is $h = 0$. Similarly, an individual who believes that the environment is extremely unhealthy exhibits first-best consumption only when the realization is $h = 1$. The interesting case arises when the true realization of h lies in $(0,1)$. This is a low perceived probability event under either $\underline{p}(h)$ or $\overline{p}(h)$ but occurs generically given the true distribution $f(h)$. When it happens, the individual with perception $\underline{p}(h)$ consumes more often than optimal, though not always, whereas the individual with perception $\overline{p}(h)$ consumes less often than

[12] One natural question is whether and, if so, how fast individuals realize that their perceived distribution about health is incorrect. We do not address this issue here.

optimal, though still sometimes. The reason is simple. Consider an interior realization of h. As a function of the attention cost c, system $\mathcal{M}$ will either modulate the signal or not. Modulation triggers optimal consumption by system $\mathcal{C}$, that is, whenever $\theta > \alpha h$. No modulation triggers highly optimistic beliefs and always consumption under $\underline{p}(h)$ or highly pessimistic beliefs and never consumption under $\overline{p}(h)$. For an individual with a given belief, the final consumption is an appropriately weighted combination of the choices under modulation and no modulation. Overall, the result shows that prior beliefs are key in shaping modulation and consumption. Incorrect perceptions of the distribution can have an impact in choice even when the realizations are correctly interpreted. This also means that making individuals aware of h, the health effects of the good they are about to consume (or not), is not enough to prevent suboptimal choices: as long as their perceived distribution is incorrect, decisions may be distorted.

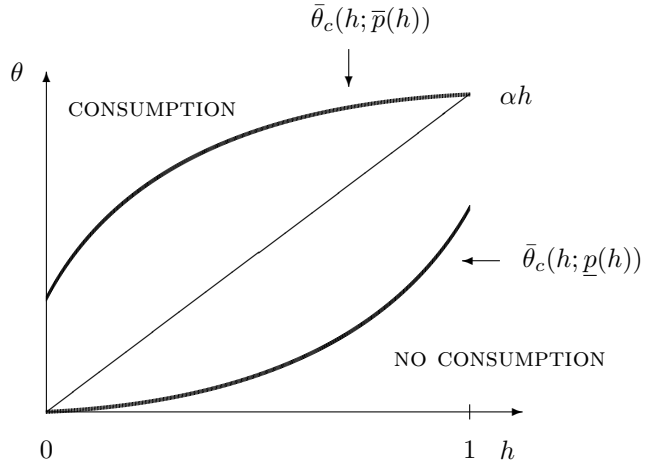


Figure 3. Expected consumption with incorrect perceptions $\overline{p}(h)$ and $\underline{p}(h)$

The findings are consistent with the existing evidence on Anorexia Nervosa (AN) and Bulimia Nervosa (BN). These two conditions are characterized by extreme behavior patterns in which subjects systematically under- and over-consume, respectively. In our model, these behaviors emerge for subjects whose prior belief is that all goods are excessively unhealthy ($\overline{p}(h)$) and unusually healthy ($\underline{p}(h)$), respectively. More specifically, inefficiencies result from a dysfunction of the rule-based mechanism because contextual information is represented through a biased assessment of the environment. As developed below, this prediction is consistent with recent findings.

There is evidence (Brooks et al. (2012)) that eating disorders emerge from an imbalance between mesolimbic regions (ventral striatum which has an activity correlated with vmPFC) and prefrontal regions involved in cognitive evaluations (dlPFC, MPFC, OFC and ACC).[13] Recent neuroimaging studies on patient with AN and BN support the hypothesis that behavior reflects different activity patterns in dlPFC. More precisely, an excessive activation pattern in dlPFC is associated with a strict refusal to eat in AN patients while an insufficient activation pattern in that same region predicts binge eating in BN patients. The reversed patterns are observed in mesolimbic regions: reduced activation in the case of AN patients and hyperactivation in the case of BN patients (van Kuyck et al. (2009), Kaye et al. (2009), Kaye et al. (2011), Brooks et al. (2011)). Consistent results are obtained for the case of obesity. Mesolimbic regions are found to be hypersensitive in response to food stimuli while regions of the prefrontal cortex are found to have activation deficits.[14]

Taken together, those results suggest that eating disorders emerge from a dysfunction of the processes in charge of representing the attributes and the contextual information. AN patients compute low taste attributes and act according to a rule-based mechanism that represents an excessively unhealthy environment, thereby triggering under-consumption. This is reflected by a low activity in the mesolimbic regions (low θ) and a high activity in regions involved in rule-based processing (high $E[h]$). BN patients compute high taste attributes and act according to a rule-based mechanism that represents an unusually healthy environment, thereby triggering over-consumption. This is reflected by a high activity in the mesolimbic regions (high θ) and a low activity in regions involved in rule-based processing (low $E[h]$). By focusing on the effect of health expectations on choices, our model captures the dysfunction of the rule-based process.[15] The fact that transcranial magnetic stimulation can effectively correct behavior suggests that dlPFC supervises the construction of contextual information, another key feature of our model.

4.2 Dynamic inconsistencies

The difficulty for individuals to incorporate high-order motivations may come simply from the problem of correctly evaluating the temporal distance until the realization of the distant component of the good. Although this is not the interpretation in [HCR]'s multi-attribute model of self-control, it corresponds to the more traditional neuroscience ap-

[13] Interestingly, and as will be discussed later, similar structures are implicated in the case of drug and behavioral addictions (Reuter et al. (2005)).

[14] Studies using transcranial magnetic stimulation also showed that symptoms were reduced in AN patients (feeling fat or anxious) and BN patients (suffering binge eating disorders) in conjunction with stimulation of dlPFC (Van den Eynde et al. (2010); Van den Eynde et al. (2013)).

[15] A richer model of expectations regarding the taste parameter is necessary to also capture the dysfunction occurring in the mesolimbic regions.

proaches to dynamic choices and self-control (see e.g. Luo et al. (2009)). Our basic model can be straightforwardly extended to encompass this alternative interpretation. Indeed, suppose that consumption at date 0 has a fixed and known negative health effect equal to one which occurs at an unknown date t ($\in \{1, ..., \tau\}$) with τ finite but possibly large. Signal modulation by system $\mathcal{M}$ consists in the costly transmission of t, the delay between the time where the pleasurable taste is enjoyed and the time where the negative health effects are suffered. Under the standard assumption of exponential discounting of payoffs, the analogue of the utility described in (1) is:

$$\begin{cases} \theta - \delta^t & \text{if} \quad z = 1 \\ 0 & \text{if} \quad z = 0 \end{cases} \tag{3}$$

where the basic and high-order components of the model are now taste and delay, θ and t, rather than taste and healthiness. Following the same methodology as in section 2.3, it is straightforward to show that modulation occurs when the realization of t is in a certain set T. At equilibrium, modulation dominates no modulation when:

$$\left| \delta^t - E[\delta^t \mid t \in T] \right| > \eta(c) \equiv \sqrt{2c} \tag{4}$$

The equilibrium modulation interval has similar properties to those in section 2.3, as evidenced by equations (2) and (4). In particular, there is information transmission when the realization of t is improbable. When the harm is likely to be suffered in a distant date, system $\mathcal{M}$ signals only near dates, resulting in no modulation and possibly inefficient consumption decisions when the health realization occurs in the distant future. When the harm is likely to be suffered in a near date, $\mathcal{M}$ signals only distant dates and inefficiencies occur when the health realizations occurs in the near future.

We follow the same methodology as in section 3.3 and assume that c is stochastic. We can determine the expected taste cutoff above which the individual consumes as a function of the delay between the pleasurable consumption and its health consequences, both with zero and positive cost of modulation. We obtain the following conclusion.

Corollary 6 - Optimal modulation and hyperbolic discounting. *Costly signal modulation generates decreasing impatience.*

From (3), it is immediate that an individual with no cost of modulation learns the realization of t and consumes whenever $\theta > \delta^t$. By construction, this corresponds to the traditional exponential discounting model. Denote by $\bar{\theta}_c(t)$ the analogue of $\bar{\theta}_c(h)$ when the high-order attribute is the date at which the health component is suffered. Using a straightforward change in variable, we can graphically depict in Figure 4 the analogue of Figure 2 in the (θ, t) space.

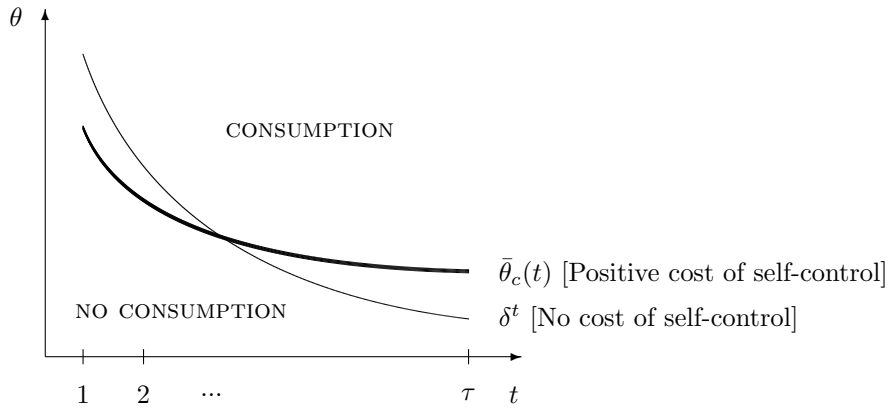


Figure 4. Expected consumption as a function of the *delay* realization

Relative to the case where modulation is costless, an individual will on average exhibit increased consumption when the negative health effect is in the near future (t close to 1) and decreased consumption when the negative health effect is in the distant future (t close to τ). This means that the implicit discount function when the cost of modulation is positive exhibits properties consistent with hyperbolic discounting: consumption is more frequent than in the exponential case when the externality is close and less frequent when it is far away. Notice that this is different from arguing that a cost of modulation implies more impatience or a steeper discounting. Instead, we show that it endogenously generates *decreasing* impatience, that is, a marginal rate of substitution between two consecutive periods that falls monotonically. This behavioral anomaly has long been noted in psychology and economics (Ainslie (1975); Prelec (1989); Laibson (1997)) and can be rationalized with costly modulation.

Even though the specific paradigm outlined in this section has not been studied in the literature, recent neuroimaging studies suggest that similar regions and processes as those detailed in the previous sections are involved. In particular, time discounting studies show that ventral striatum, mPFC and OFC, together labeled as limbic reward areas, are activated in response to immediate rewards while regions in the lPFC are involved in intertemporal trade-offs (McClure et al. (2004); McClure et al. (2007)). In apparent contradiction, according to Kable and Glimcher (2007) those limbic reward areas are also encoding delayed values. These findings can be reconciled. Indeed they are both consistent with the hypothesis that system $\mathcal{C}$ (here, the limbic reward areas) encodes information about the immediate taste attribute but is also capable of representing general properties of the high-order attribute, that is, the delayed health effects. In parallel, system $\mathcal{M}$

(here, the lPFC) has access to the high-order information and supervises decisions. We conjecture that a cognitive process involving these two regions is in place, with $\mathcal{M}$ sending precise information to $\mathcal{C}$ about delayed health ratings when the expected delays computed via a rule-based process are likely to result in significant inefficiencies.

4.3 Addiction

We now extend the basic model to account for the possibility of habit formation or inter-period consumption externalities. Suppose that the individual chooses consumption during two periods, 1 and 2. In period 1, the individual's utility is given by (1) just as in the basic model. In period 2, the utility depends on both present and past consumption decisions. More precisely, denote by $u_{zz'}$ the utility in period 2 of consumption choice z' ($\in \{0,1\}$) given a consumption choice z ($\in \{0,1\}$) in period 1. In keeping with the most traditional definition of addiction and habit formation (Becker and Murphy (1988)), we assume that:

$$\begin{array}{ll} u_{01} = \theta - \alpha h & \\ u_{00} = 0 & \end{array} \quad \text{and} \quad \begin{array}{ll} u_{11} = p\,\theta - \alpha h & \\ u_{10} = -q\,\theta & \end{array} \tag{5}$$

where $p < 1$ and $p + q > 1$. In words, if the individual abstains in period 1 ($z = 0$), his period 2 utility remains unchanged. If the individual consumes in period 1 ($z = 1$), then the total utility of consumption in period 2 is reduced ($u_{11} < u_{01}$) but the marginal utility is increased ($u_{11} - u_{10} > u_{01} - u_{00}$). Notice also that the individual is subject to "craving", that is, a disutility of abstaining in period 2 after having consumed in period 1 ($u_{10} < u_{00}$), and craving is increasing in the realization of taste θ.

We take a myopic approach and assume that, in each period, the individual bases his consumption decision exclusively on the present cost and benefit.[16] The decision in period 1 and the decision in period 2 given abstinence in period 1 are the same as in Theorem 1. Also, and by construction, if information about health is transmitted from system $\mathcal{M}$ to system $\mathcal{C}$, consumption in period 2 is more likely to occur after consumption in period 1 than after abstinence (formally, $\theta > \alpha h/(p+q)$ vs. $\theta > \alpha h$). A more interesting question is to determine whether the decision to modulate the goal value is affected by past consumption. To answer that question, we determine the value of modulation and no modulation after period 1 consumption:

$$V_1^+(h) = \int_0^{\frac{\alpha h}{p+q}} -q\,\theta d\theta + \int_{\frac{\alpha h}{p+q}}^1 (p\,\theta - \alpha h) d\theta - c$$

[16] This allows us to better focus on the inter-period consumption externality without polluting it with the (important but for our model peripheral) forward-looking considerations of current behavior on future utility. Naturally, with forward looking agents, other effects would add to the one studied here.

and

$$V_1^-(h) = \int_0^{\frac{\alpha E[h|h\in H]}{p+q}} -q\,\theta d\theta + \int_{\frac{\alpha E[h|h\in H]}{p+q}}^1 (p\,\theta - \alpha h)d\theta$$

Again, self-control dominates no self-control when $V_1^+(h) > V_1^-(h)$, which translates into:

$$\begin{aligned}\Delta_1(h, E[h|h\in H]) &\equiv \int_{\frac{\alpha E[h|h\in H]}{p+q}}^{\frac{\alpha h}{p+q}} \Big(\alpha h - (p+q)\theta\Big)d\theta \;>\; c \\ &\Leftrightarrow \quad \Big|h - E[h|h\in H]\Big| > \eta_1(c,\alpha,p,q) \equiv \frac{\sqrt{2c(p+q)}}{\alpha} \qquad (6)\end{aligned}$$

and we obtain the following result.

Corollary 7 - Under-modulation of addictive goods. *Signal modulation is less prevalent for addictive than for non-addictive substances.*

The result is obtained by direct inspection of equations (2) and (6). The very fact that the marginal effect of taste in current consumption increases with past consumption ($p + q > 1$), also implies that the incentives to modulate the signal decrease with past consumption. Indeed, increasing the marginal value of taste makes consumption more desirable, which implies that spending the cost of knowing the exact health realization is less valuable (since it has a lower impact on decision). Overall, the set of health realizations for which there is no modulation expands. Moreover, as the individual is exposed to an addictive good, his decision-making may switch from a regime in which modulation is optimal at the top or the bottom, to a regime in which modulation is never optimal. The direct consequence of both effects is that blissful ignorance of the long term health repercussions of consumption is *endogenously more frequent* for addicted than for non-addicted individuals. The result is not based on an irrational self-delusion motive but, instead, on the optimal decision to avoid costly signal modulation.

Interestingly, the current evidence on the neurobiology of addiction points to the importance of both vmPFC and dlPFC, our main candidates for systems $\mathcal{C}$ and $\mathcal{M}$, respectively. The mechanisms underlying addiction are intricate. However, recent studies show that addiction disrupts PFC functions that are fundamental for regulation and decision-making (Goldstein and Volkow (2011)). In particular, activity in vmPFC decreases with the severity of addiction. This pattern of activity is consistent with our modeling of addiction: a given item becomes less desirable when it has been consumed often in the past. Hypoactivity of dlPFC is also noted and interpreted as a deficiency of inhibitory control in drug-addicted individuals. Our model suggests that hypoactivity may well be the natural consequence of the interplay between the two systems, rather than the result of an

impairment due to drug exposure. Understanding the causes of a lack of inhibitory control is important for intervention.

The result also suggests two ways of combating addiction. First, focusing the subjects' attention (i.e., decreasing c) enhances the incentives to restore communication between systems $\mathcal{M}$ and $\mathcal{C}$. Second and perhaps more importantly, offering an environment in which addictive goods are scarce (i.e., changing the true distribution $f(h)$ or even the perceived distribution $p(h)$) makes addicted individuals less prone to consume goods that escalate their addiction, even conditional on the true realization of h.

5 Conclusion

In this paper we build a novel theory of self-control based on the neurophysiological experiment of [HCR]. Our model provides new testable implications regarding self-control that depart from traditional analyses in economics. We highlight the importance of the environment in triggering value modulation and therefore endogenously affecting consumption decisions. We also argue that self-indulgence and self-restraint are two sides of the same coin: they are both the result of the same costly value modulation phenomenon. Finally, our theory sheds light on some behavioral anomalies. Eating disorders can be the result of an incorrect perception of the general health properties of goods, decreasing impatience can be rationalized by a positive cost of modulation, and ignorance of health considerations naturally occur more frequently for individuals suffering from addiction.

The paper illustrates the value of combining the constrained optimization tools of microeconomic theory with the insights obtained from neuroscience experiments. Indeed, existing theories of temptation and self-control are insightful but incomplete. Behavioral economics uses introspection, casual observation and empirical observation of choices to refine those models. However, widely different behavioral theories can be constructed to explain the same departure and it is difficult to know a priori which theory is more adequate. In our methodology, the discipline of the model comes from the neuroscience experimental evidence. The model has a rich set of testable predictions, both behavioral and neurophysiological, which can form the basis of new experiments. In that respect, the paper provides an argument regarding the value of neuroscience for economics that has not been emphasized much in the recent methodological debates (Gul and Pesendorfer (2008); Camerer (2008)).

Finally, we view this approach as complementary (not substitute) to computational neuroscience. Indeed, one could build a mechanical model where system $\mathcal{M}$ automatically sends health "warnings" to system $\mathcal{C}$ for certain realizations of h. This seems a more plau-

sible implementation of what is actually going on in the brain than the "as if" optimization mechanism developed in the paper. The trouble with such approach is that it provides little guidance as to when is the health information supposed to be transmitted. As a first approximation, one could argue that warnings should be sent for health realizations above a certain threshold. The added value of our optimization approach is to show that optimal information transmission generically requires a threshold and a cap and that these values depend on the cost of modulation, the relative importance of health, and the distribution from which the health parameter is drawn. Once this is understood, one can think of a mechanical implementation along the lines described above.

Appendix A

Proof of Theorem 1

Step 1. Δ^{Ns} is convex with a minimum at $E[h \mid h \in H]$. This implies that H is necessarily a compact set. As a result, there are only four possible types of equilibria depending on whether no modulation is exerted when h is at the bottom (called **B** and formalized as $H = [0, h^{**}]$), at the middle (**M**, with $H = [h^o, h^{oo}]$), at the top (**T**, with $H = [h^*, 1]$) or always (**A**, with $H = [0, 1]$).[17] Naturally $h^*, h^{**}, h^o, h^{oo} \in (0, 1)^4$. We now characterize those sets and provide conditions for existence.

Bottom (**B**): $H = [0, h^{**}]$ where h^{**} is the value such that $\Delta^{Ns}(h^{**}, E[h \mid h \leq h^{**}]) = c$. Using (2), we get:

$$h^{**} - E[h \mid h \leq h^{**}] = \eta \tag{7}$$

where $\eta(c, \alpha) = \sqrt{2c}/\alpha$. It is well-known that if $f(h)$ is log-concave then $x - E[h \mid h \leq x]$ is increasing in x (see e.g. Bagnoli and Bergstrom (2005)), so h^{**} is unique and $dh^{**}/d\eta > 0$. The conditions for this equilibrium to exist are $\Delta^{Ns}(0, E[h \mid h \leq h^{**}]) < c$ and $\Delta^{Ns}(1, E[h \mid h \leq h^{**}]) > c$, which can be rewritten as:

$$E[h \mid h \leq h^{**}] < \eta \tag{B1}$$

and

$$E[h \mid h \leq h^{**}] < 1 - \eta \tag{B2}$$

Middle (**M**): $H = [h^o, h^{oo}]$ where h^o and h^{oo} are the values such that $\Delta^{Ns}(h^o, E[h \mid h^o \leq h \leq h^{oo}]) = c$ and $\Delta^{Ns}(h^{oo}, E[h \mid h^o \leq h \leq h^{oo}]) = c$ respectively. Using (2), we get:

$$E[h \mid h^o \leq h \leq h^{oo}] - h^o = \eta \quad \text{and} \quad h^{oo} - E[h \mid h^o \leq h \leq h^{oo}] = \eta$$

Combining them, we get $h^{oo} = h^o + 2\eta$ and therefore:

$$E[h \mid h^o \leq h \leq h^o + 2\eta] = h^o + \eta \tag{8}$$

We will prove later that, in this equilibrium, h^o must be unique. The conditions for this type of equilibrium to exist are $\Delta^{Ns}(0, E[h \mid h^o \leq h \leq h^o + 2\eta]) > c$ and $\Delta^{Ns}(1, E[h \mid h^o \leq h \leq h^o + 2\eta]) > c$, which can be rewritten as:

$$E[h \mid h^o \leq h \leq h^o + 2\eta] > \eta \tag{M1}$$

[17] Never no modulation (that is, always modulation or $H = \emptyset$) cannot be an equilibrium if $c > 0$. Indeed, for any belief distribution that follows a deviation to no modulation, there exists a value h characterized by the expectation of such belief such that $\Delta^{Ns} = 0$ ($< c$), making the deviation to no modulation profitable.

and

$$E[h \mid h^o \leq h \leq h^o + 2\eta] < 1 - \eta \tag{M2}$$

Top (**T**): $H = [h^*, 1]$ where h^* is the value such that $\Delta^{Ns}(h^*, E[h \mid h \geq h^*]) = c$. Once again using (2), we have:

$$E[h \mid h \geq h^*] - h^* = \eta \tag{9}$$

If $f(h)$ is log-concave then $E[h \mid h \geq x] - x$ is decreasing in x (Bagnoli and Bergstrom, 2005), so h^* is unique and $dh^*/d\eta < 0$. The conditions for this type of equilibrium to exist are $\Delta^{Ns}(0, E[h \mid h \geq h^*]) > c$ and $\Delta^{Ns}(1, E[h \mid h \geq h^*]) < c$, which can be rewritten as:

$$E[h \mid h \geq h^*] > \eta \tag{T1}$$

and

$$E[h \mid h \geq h^*] > 1 - \eta \tag{T2}$$

Always (**A**): $H = [0, 1]$. The conditions for this equilibrium to exist are $\Delta^{Ns}(0, E[h]) < c$ and $\Delta^{Ns}(1, E[h]) < c$, which can be rewritten as:

$$E[h] < \eta \tag{A1}$$

and

$$E[h] > 1 - \eta \tag{A2}$$

Step 2. We now prove that an equilibrium exists and is unique for all parameter values of η and for all strictly log concave distributions $F(\cdot)$. From (h^{**}, h^o, h^*) as defined in (7)-(8)-(9) and the equilibrium conditions $(B1)$-$(B2)$-$(M1)$-$(M2)$-$(T1)$-$(T2)$-$(A1)$-$(A2)$, we have three cases.

Case 1: $\eta \geq 1$. Trivially, the unique equilibrium is **A** for all $F(h)$.

Case 2: $\eta \in [1/2, 1)$. In this case, $1 - \eta \leq \eta$ so **M** can never be an equilibrium.

• If $E[h] < 1 - \eta$ (that is, $1 - E[h] > \eta$), then a **B**-equilibrium exists and is unique. Indeed, $x - E[h \mid h \leq x]$ increasing in x implies there exists a unique $h^{**} < 1$ such that $h^{**} - E[h \mid h \leq h^{**}] = \eta$. Furthermore, $E[h] < 1 - \eta$ implies that a **A**-equilibrium cannot exist. Finally, $E[h] < 1 - \eta$ implies $E[h] < \eta$. $E[h \mid h \geq x] - x$ decreasing in x implies $E[h \mid h \geq x] - x < \eta$ for all x so a **T**-equilibrium does not exist either.

• If $E[h] > \eta$, then a **T**-equilibrium exists and is unique. Indeed, $E[h \mid h \geq x] - x$ decreasing in x implies there exists a unique $h^* > 0$ such that $E[h \mid h \geq h^*] - h^* = \eta$. Furthermore, $E[h] > \eta$ implies that a **A**-equilibrium cannot exist. Finally, $E[h] > \eta$ implies $1 - E[h] < \eta$. $x - E[h \mid h \leq x]$ increasing in x implies $x - E[h \mid h \leq x] < \eta$ for all x so a **B**-equilibrium does not exist either.

• If $E[h] \in [1-\eta, \eta]$, then a **A**-equilibrium exists. $E[h \,|\, h \geq x] - x$ decreasing in x implies $E[h \,|\, h \geq x] - x < \eta$ for all x so a **T**-equilibrium does not exist. $x - E[h \,|\, h \leq x]$ increasing in x implies $x - E[h \,|\, h \leq x] < \eta$ for all x so a **B**-equilibrium does not exist either.

To summarize, when $\eta \in [1/2, 1)$ the equilibrium is unique given $F(\cdot)$ and η:

$$\begin{array}{ll} \mathbf{B} & \text{if} \quad E[h] < 1-\eta \\ \mathbf{A} & \text{if} \quad E[h] \in [1-\eta, \eta] \\ \mathbf{T} & \text{if} \quad E[h] > \eta \end{array}$$

Case 3: $\eta \in (0, 1/2)$. In this case, $\eta < 1 - \eta$ so **A** can never be an equilibrium.

In a **B**-equilibrium, $h^{**} - E[h \,|\, h < h^{**}] = \eta$ and $E[h \,|\, h < h^{**}] < \eta$. Therefore, $h^{**} < 2\eta$. Since $x - E[h \,|\, h \leq x]$ is increasing in x, then $2\eta - E[h \,|\, h < 2\eta] > \eta$ or:

$$E[h \,|\, h < 2\eta] < \eta \tag{1'}$$

In a **M**-equilibrium, $E[h \,|\, h^o \leq h \leq h^{oo}] - h^o = \eta$ and $h^{oo} - E[h \,|\, h^o \leq h \leq h^{oo}] = \eta$. Therefore, $h^{oo} = h^o + 2\eta$, and therefore:

$$E[h \,|\, h^o \leq h \leq h^o + 2\eta] = h^o + \eta \tag{2'}$$

In a **T**-equilibrium, $E[h \,|\, h > h^*] - h^* = \eta$ and $E[h \,|\, h > h^*] > 1 - \eta$. Therefore, $h^* > 1 - 2\eta$. Since $E[h \,|\, h < x] - x$ is decreasing in x, $E[h \,|\, h > 1 - 2\eta] - (1 - 2\eta) > \eta$ or:

$$E[h \,|\, h > 1 - 2\eta] > 1 - \eta \tag{3'}$$

Note that $f(h)$ strictly log-concave and differentiable implies single-peakedness of $f(h)$ resulting in three possible cases:[18] (i) $f'(h) < 0$ for all h (i.e. $F''(h) < 0$), (ii) $f'(h) > 0$ for all h (i.e. $F''(h) > 0$), or (iii) $f'(h) \gtrless 0$ for all $h \lessgtr \tilde{h}$ (i.e. $F''(h) \gtrless 0$ for all $h \lessgtr \tilde{h}$).

Consider intervals $[a, b] \subset [0, 1]$ and $h \in [a, b]$, we have the following relationship:[19]

$$F''(h) \lessgtr 0 \quad \Rightarrow \quad (b-a)\frac{F(b)+F(a)}{2} \lessgtr \int_a^b F(h)dh \quad \Leftrightarrow \quad E[h \,|\, a < h < b] \lessgtr \frac{a+b}{2} \tag{4'}$$

Combining (1')-(2')-(3') with (4'), there are three subcases:

(3i) $F''(h) < 0$ for all h. Only (1') can be satisfied so only **B** can be an equilibrium. Furthermore, $h^{**} < 1$ if and only if $1 - E[h] > \eta$ which is always true since $F''(h) < 0$ implies $E[h] < 1/2$ $(< 1 - \eta)$.

[18] Given $f(h)$ is positive, strict log concavity of $f(h)$ is equivalent to strict concavity of $log(f(h))$, i.e. $f''(h)f(h) < (f'(h))^2$. Suppose there exists x such that $f'(x) = 0$, then it has to be the case that $f''(x) < 0$. At most one value satisfies this property.

[19] The first implication relies on obvious properties of increasing and concave as well as increasing and convex functions. The second equivalence is obtain by integrating $\int_a^b F(h)dh$ by parts.

(3ii) $F''(h) > 0$ for all h. Only $(3')$ can be satisfied so only **T** can be an equilibrium. Furthermore, $h^* > 0$ if and only if $E[h] > \eta$ which is always true since $F''(h) > 0$ implies $E[h] > 1/2$ $(> \eta)$.

(3iii) $F''(h) \gtrless 0$ for all $h \lessgtr \tilde{h}$. Fix η and $F(\cdot)$ and suppose that $\eta\, F(2\eta) < \int_0^{2\eta} F(h)dh$ (condition $(1')$ holds). This means there is $h^{**} < 1$ such that a **B**-equilibrium exists. Furthermore, it is by construction such that $F''(2\eta) < 0$ (or $2\eta > \tilde{h}$). This equilibrium is unique if $(2')$ and $(3')$ never hold, that is, if:

$$\eta[F(x+2\eta) + F(x)] < \int_x^{x+2\eta} F(h)dh \qquad \forall x \in (0, 1-2\eta] \tag{5'}$$

Fix x and notice that $F''(x+2\eta) < 0$ which implies $F'(x+2\eta) < \frac{F(x+2\eta)-F(x)}{2\eta}$ for all x. If $F''(x) < 0$, then $F''(h) < 0$ for all $h \in [x, x+2\eta]$ and therefore $(5')$ automatically holds as in case (3i). If $F''(x) > 0$, then $F'(x) < \frac{F(x+2\eta)-F(x)}{2\eta}$. This means that $\eta[F'(x)+F'(x+2\eta)] < F(x+2\eta) - F(x)$. So, as x increases the left hand side of $(5')$ increases at a lower rate than the right hand side of $(5')$. Hence, if the inequality holds at $x = 0$ it also holds at all $x > 0$ and no **M** or **T** equilibrium exists.

Suppose now that $\eta[1 + F(1-2\eta)] > \int_{1-2\eta}^1 F(h)dh$ (condition $(3')$ holds). This means there is $h^* > 0$ such that a **T**-equilibrium exists. Furthermore, it is by construction such that $F''(1-2\eta) > 0$ (or $1-2\eta < \tilde{h}$). This equilibrium is unique if $(1')$ and $(2')$ never hold, that is, if:

$$\eta[F(1-x) + F(1-2\eta-x)] > \int_{1-2\eta-x}^{1-x} F(h)dh \qquad \forall x \in [0, 1-2\eta] \tag{6'}$$

Fix x and notice that $F''(1-2\eta-x) > 0$ which implies $F'(1-2\eta-x) < \frac{F(1-x)-F(1-2\eta-x)}{2\eta}$ for all x. If $F''(1-x) > 0$, then $F''(h) > 0$ for all $h \in [1-2\eta-x, 1-x]$ and therefore $(6')$ automatically holds as in case (3ii). If $F''(1-x) < 0$, then $F'(1-x) < \frac{F(1-x)-F(1-2\eta-x)}{2\eta}$. This means that $\eta[F'(1-x)+F'(1-2\eta-x)] < F(1-x)-F(1-2\eta-x)$. So, as x increases the left hand side of $(6')$ decreases at a lower rate than the right hand side of $(6')$. Hence if the inequality holds at $x = 0$ it also holds at all $x > 0$ and no **M** or **B** equilibrium exists.

Suppose last that $\eta[F(h^o+2\eta) + F(h^o)] = \int_{h^o}^{h^o+2\eta} F(h)dh$ (condition $(2')$ holds). Consider the function $M(x) = \int_x^{x+2\eta} F(h)dh - \eta[F(x+2\eta) + F(x)]$. It represents the area between the curve $F(x)$ between x and $x+2\eta$ and the line joining points $(x, F(x))$ to point $(x+2\eta, F(x+2\eta))$. By construction, the line is above the curve whenever $x < \tilde{h} - 2\eta$ and therefore we have $M(x) < 0$. By contrast, the line is below the curve when $x > \tilde{h}$ and therefore, we have $M(x) > 0$. For all $x \in [\tilde{h} - 2\eta, \tilde{h}]$, there exists a unique point $k(x)$ where the line and the curve cross. The line is above the curve on $[x, k(x)]$ and it is

above the curve on $[k(x), x+2\eta]$. Let $\underline{M}(x)$ be the area between the curve and the line on $[x, k(x)]$ and $\overline{M}(x)$ the area between the curve and the line on $[k(x), x+2\eta]$, we have $M(x) = \underline{M}(x) + \overline{M}(x)$, $\underline{M}(x) < 0$ and $\overline{M}(x) > 0$. By construction, as x increases, $k(x)$ decreases and therefore both $\underline{M}(x)$ and $\overline{M}(x)$ increase. Assume there exists a point x^* such that $M(x^*) = 0$. This point must lie in $[\tilde{h}-2\eta, \tilde{h}]$ and must be such that $\underline{M}(x^*) = -\overline{M}(x^*)$. Given the previous points, $M(x) < 0$ for all $x < x^*$, and $M(x) > 0$ and for all $x > x^*$.This proves that h^o is unique when it exists. It also proves that $M(0) < 0$ (condition (1') does not hold) and $M(1-2\eta) > 0$ (condition (3') does not hold).Therefore, if a **M** equilibrium exists, no **B** or **T** equilibrium can exist. Last, note that by construction the two cutoffs h^o and $h^{oo} = h^o + 2\eta$ need to move away from each other when η increases. Therefore, we necessarily have $h^{o'}(\eta) < 0$ and $h^{oo'}(\eta) > 0$.

To summarize, when $\eta \in (0, 1/2)$ the equilibrium is unique given $F(\cdot)$ and η:

$$\begin{array}{ll} \mathbf{B} & \text{if} \quad E[h \,|\, h < 2\eta] < \eta \\ \mathbf{M} & \text{if} \quad E[h \,|\, h < 2\eta] > \eta \quad \text{and} \quad E[h \,|\, h > 1-2\eta] < 1-\eta \\ \mathbf{T} & \text{if} \quad E[h \,|\, h > 1-2\eta] > 1-\eta \end{array}$$

Proof of Corollary 1

We analyze the quantity of modulation as a function of η.

Step 1. We first study the sequence of equilibria for a given distribution $F(\cdot)$ as a function of η.

Case 1: $\eta > 1$. The equilibrium is always **A**. There is no modulation.

Case 2: $\eta \in (1/2, 1)$. This is immediate from Step 2 - case 2 in the proof of the main theorem. If the distribution is such that $E[h] < 1/2$, the equilibrium is **B** when $\eta \in (1/2, 1-E(h))$ and **A** when $\eta \in (1-E(h), 1)$. If the distribution is such that $E[h] \geqslant 1/2$, the equilibrium is **T** when $\eta \in (1/2, E(h))$ and **A** when $\eta \in (E(h), 1)$.

Case 3: $\eta \in (0, 1/2)$. We have 3 subcases

(2i) $F''(h) < 0$ for all h. The only equilibrium for all $\eta \in (0, 1/2)$ is **B**. This is immediate from the previous proof.

(2ii) $F''(h) > 0$ for all h. The only equilibrium for all $\eta \in (0, 1/2)$ is **T**. This is also immediate.

(2iii) $F''(h) \gtrless 0$ for all $h \lessgtr \tilde{h}$. Let

$$A(\eta) = \eta F(2\eta) - \int_0^{2\eta} F(h)dh$$

$$B(\eta) = \eta[1 + F(1-2\eta)] - \int_{1-2\eta}^{1} F(h)dh$$

We remind the reader that $A(\eta) < 0$ in a **B**-equilibrium and $B(\eta) > 0$ in a **T**-equilibrium. We have $A(1/2) = B(1/2) = E[h] - 1/2$ and $A(0) = B(0) = 0$. Note that $A''(\eta) = 4\eta f'(2\eta)$, hence $A(\eta)$ is convex for all $\eta \leqslant \tilde{h}/2$ and concave for all $\eta > \tilde{h}/2$. We also have $A'(\eta) = 2\eta f(2\eta) - F(2\eta)$ and given by construction $hf(h) > F(h)$ for all $h < h$, we deduce that $lim_{\eta \to 0} A'(\eta) > 0$. Combining these findings, when $E(h) > 1/2$, $A(\eta) > 0$ for all $\eta \in (0, 1/2)$. When $E(h) < 1/2$, there exists a unique $\underline{\eta}$ such that $A(\eta) > 0$ for all $\eta < \underline{\eta}$ and $A(\eta) < 0$ for all $\eta \in (\underline{\eta}, 1/2)$.

Note also that $B''(\eta) = 4\eta f'(1 - 2\eta)$, hence $B(\eta)$ is concave for all $\eta \leqslant (1 - \tilde{h})/2$ and convex for all $\eta > (1 - \tilde{h})/2$. We also have $B'(\eta) = -2\eta f(1 - 2\eta) + 1 - F(1 - 2\eta)$ and given by construction $(1-h)f(h) > 1-F(h)$ for all $h > \tilde{h}$, we deduce that $lim_{\eta \to 0} B'(\eta) < 0$. Combining these findings, when $E(h) < 1/2$, $B(\eta) < 0$ for all $\eta \in (0, 1/2)$. When $E(h) > 1/2$, there exists a unique $\overline{\eta}$ such that $B(\eta) < 0$ for all $\eta < \overline{\eta}$ and $A(\eta) > 0$ for all $\eta \in (\overline{\eta}, 1/2)$.

Overall, when $E[h] > 1/2$, the equilibrium is **M** when $\eta < \overline{\eta}$ and **T** when $\eta \in (\overline{\eta}, 1/2)$. When $E[h] < 1/2$, the equilibrium is **M** when $\eta < \underline{\eta}$ and **B** when $\eta \in (\underline{\eta}, 1/2)$.

To summarize, as η increases, the sequence of equilibria given $F(\cdot)$ is:

$$\begin{array}{ll} \textbf{(M) B A} & \text{if} \quad E[h] < 1/2 \\ \textbf{(M) T A} & \text{if} \quad E[h] > 1/2 \end{array}$$

where parentheses indicate that the equilibrium may not occur for some distributions. Recall that **(M)** is a possible equilibrium only when there exists $\tilde{h} \in (0, 1)$ such that $F''(h) \gtrless 0$ for all $h \lesseqgtr \tilde{h}$.

Note that $A(\underline{\eta}) = 0$ implies that $h^o(\underline{\eta}) = 0$ and $h^{oo}(\eta^*) = 2\underline{\eta}$. Combining (1') and $A(\underline{\eta}) = 0$ implies $h^{**}(\underline{\eta}) = 2\underline{\eta}$ as well. Therefore, as η increases, the transition from **M** to **B** is smooth. Similarly, $B(\overline{\eta}) = 0$ implies that $h^o(\overline{\eta}) = 1 - 2\overline{\eta}$ and $h^{oo}(\overline{\eta}) = 1$. Combining (3') and $B(\overline{\eta}) = 0$ implies $h^*(\overline{\eta}) = 1 - 2\overline{\eta}$. Therefore, as η increases, the transition from **M** to **T** is also smooth.

Step 2. Note that h^{**} increases in η and therefore, there is less modulation in **B** equilibria as η increases. Similarly, h^* decreases in η and therefore, there is less modulation in **T** equilibria as η increases. When **(M)** exists, no modulation occurs only between h^o and $h^{oo} = h^o + 2\eta$. As η increases, the interval in which no modulation occurs increases. Given the transition from equilibria **(M)** and **B** or **(M)** and **T** are smooth, the quantity of modulation decreases continuously as η increases.

Proof of Corollary 2

We now identify the properties of the equilibrium for a given η as a function of the distribution. Consider a family of functions parametrized by z which are ordered by the

standard monotone likelihood ratio property (MLRP). Formally:

$$\left(\frac{f_z(h)}{f_{z+1}(h)}\right)' < 0 \qquad \forall\, z$$

It is well-know that MLRP implies $E_{z+1}[h \,|\, a < z < b] > E_z[h \,|\, a < z < b]$. Suppose now that we are on a **B**-equilibrium. Denote h_z^{**} the no-self control cutoff given h is distributed according to $f_z(h)$. We have: $h_z^{**} - E_z[h \,|\, h \leq h_z^{**}] = \eta$ and $h_{z+1}^{**} - E_{z+1}[h \,|\, h \leq h_{z+1}^{**}] = \eta$. Suppose $h_{z+1}^{**} = h_z^{**} - \delta^{Ns}$ with $\delta^{Ns} > 0$. Then:

$$h_z^{**} - E_z[h \,|\, h \leq h_z^{**}] = h_z^{**} - \delta^{Ns} - E_{z+1}[h \,|\, h \leq h_z^{**} - \delta^{Ns}]$$

Log-concavity of $f(h)$ implies that $E_{z+1}[h \,|\, h \leq h_z^{**} - \delta^{Ns}] > E_{z+1}[h \,|\, h \leq h_z^{**}] - \delta^{Ns}$. Combining this inequality with the previous equation, we get:

$$E_{z+1}[h \,|\, h \leq h_z^{**}] < E_z[h \,|\, h \leq h_z^{**}]$$

which is a contradiction. Therefore, $h_{z+1}^{**} > h_z^{**}$. The same argument demonstrates that in a **M**-equilibrium $h_{z+1}^{o} > h_z^{o}$ and $h_{z+1}^{oo} > h_z^{**}$ and in a **T**-equilibrium $h_{z+1}^{*} > h_z^{*}$. Thus, as z increases (that is, as the distribution shifts towards higher values of h in a MLRP sense), we move from **B** to **A** to **T**-equilibrium if $\eta \in [1/2, 1)$ and from **B** to **M** to **T**-equilibrium if $\eta \in [0, 1/2)$.

Proof of Corollary 3: immediate and therefore omitted.

Proof of Corollary 4

Assume $E(h) < 1/2$. There are 2 possible cases.

Case 1. $f'(h) < 0$ for all h or $f'(h) \gtrless 0$ for all $h \lessgtr \tilde{h}$ such that a **M** equilibrium does not exist. From (7), for all $\eta < 1 - E[h]$, the equilibrium is **B** and there exists $h^{**}(\eta)$ such that $h^{**}(\eta) - E[h \,|\, h < h^{**}(\eta)] = \eta$. For all $\eta \geqslant 1 - E[h]$, the equilibrium is **A** and $H = [0, 1]$. For all h, there exists $i(h)$ such that $h^{**}(i(h)) = h$ and such that the equilibrium is **B** and modulation occurs at h when $\eta < i(h)$, the equilibrium is **B** but no modulation occurs at h when $\eta \in (i(h), 1 - E(h))$ and the equilibrium is **A** and no modulation occurs at h when $\eta > 1 - E[h]$. Given $h^{**}(\eta)$ is increasing, $i(h')$ is increasing. For each h', there are three consumption scenarii: (i) when $\eta < i(h')$, consumption takes place if $\theta > \alpha h'$; (ii) when $\eta \in (i(h'), 1 - E(h))$, consumption takes place when $\theta > \alpha E[h \,|\, h' < h^{**}(\eta)]$; (iii) when $\eta > 1 - E[h]$, consumption takes place when $\theta > \alpha E[h]$. Overall, for all h', consumption takes place when $\theta > \bar{\theta}_c(h)$ where

$$\bar{\theta}_c(h') = \int_{\eta=0}^{i(h')} \alpha h' \, dG(\eta) + \int_{i(h')}^{1-E[h]} \alpha E[h \,|\, h < h^{**}(\eta)] \, dG(\eta) + \int_{1-E[h]}^{\eta^*} \alpha E[h] dG(\eta)$$

Note that $\bar{\theta}_c(0) > 0$ and $\bar{\theta}_c(1) < \alpha$. We also have

$$\frac{d\bar{\theta}_c}{dh'} = \alpha\, G(i(h'))\left[1 + \frac{g(i(h'))}{G(i(h'))} i(h) i'(h')\right] > 0 \tag{10}$$

Case 2. $f'(h) \gtrless 0$ for all $h \lesseqgtr \tilde{h}$ such that a **M** equilibrium exists. For all $\eta < \underline{\eta}$, the equilibrium is **M** and there exists $h^o(\eta)$ such that $E[h|h^o(\eta) \leqslant h \leqslant h^o(\eta) + 2\eta] = h^o + \eta$. For all $\eta \in [\underline{\eta}, 1 - E[h]]$, the equilibrium is **B** and there exists $h^{**}(\eta)$ such that $h^{**}(\eta) - E[h|h < h^{**}(\eta)] = \eta$. For all $\eta > 1 - E[h]$, then $h^{**} = 1$. Note that the unique health rating that solves $E[h|h^o(\eta) \leqslant h \leqslant h^o(\eta) + 2\eta] = h^o + \eta$ when $\eta \to 0$ is $\tilde{h}$. There are three cases: (i) $h' < \tilde{h}$ and modulation occurs only for low values of η in the **M** equilibrium; (ii) $h' \in [\tilde{h}, 2\underline{\eta}]$ no modulation occurs when η is high enough in **M**, and modulation never occurs in the **B** equilibrium. (iii) $h' > 2\underline{\eta}$ and modulation occurs always in the **M** and only if η is not too high in the **B** equilibrium.

Case 2i: $h' < \tilde{h}$, there exists $i^o(h') < \underline{\eta}$ such that $h^o(i^o(h')) = h'$ and modulation occurs only when $\eta < i^o(h')$. Consumption takes place on average when $\theta > \bar{\theta}_c(h')$ where

$$\bar{\theta}_c(h') = \int_{\eta=0}^{i^o(h')} \alpha h'\, dG(\eta) + \int_{i^o(h')}^{\underline{\eta}} \alpha E[h \,|\, h^o(\eta) < h < h^{oo}(\eta)]\, dG(\eta)$$

$$+ \int_{\underline{\eta}}^{1-E[h]} \alpha E[h \,|\, h < h^{**}(\eta)]\, dG(\eta) + \int_{1-E[h]}^{\eta^*} \alpha E[h] dG(\eta)$$

We have

$$\frac{d\bar{\theta}_c}{dh'} = \alpha\, G(i^o(h'))\left[1 - \frac{g(i^o(h'))}{G(i^o(h'))} i^{o'}(h') i^{o'}(h')\right] \tag{11}$$

Note that $h^o(\eta)$ decreases in η and therefore $i^o(h')$ decreases which proves that $\frac{d\bar{\theta}_c}{dh'} > 0$. Note also that $\bar{\theta}_c(0) > 0$.

Case 2ii: $h \in [\tilde{h}, 2\underline{\eta}]$, there exists $i^{oo}(h) < \underline{\eta}$ such that $h^{oo}(i^{oo}(h)) = h$. Consumption takes place on average when $\theta > \bar{\theta}_c(h)$ where

$$\bar{\theta}_c(h') = \int_{\eta=0}^{i^{oo}(h')} \alpha h'\, dG(\eta) + \int_{i^{oo}(h')}^{\underline{\eta}} \alpha E[h \,|\, h^o(\eta) < h < h^{oo}(\eta)]\, dG(\eta)$$

$$+ \int_{\underline{\eta}}^{1-E[h]} \alpha E[h \,|\, h < h^{**}(\eta)]\, dG(\eta) + \int_{1-E[h]}^{\eta^*} \alpha E[h] dG(\eta)$$

We have

$$\frac{d\bar{\theta}_c}{dh'} = \alpha\, G(i^{oo})(h')\left[1 + \frac{g(i^{oo}(h'))}{G(i^{oo}(h'))} i^{oo'}(h') i^{oo}(h'\right] \tag{12}$$

Note that $h^{oo}(\eta)$ increases in η and therefore $i^{oo}(h')$ increases which proves that $\frac{d\bar{\theta}^e}{dh'} > 0$.

Case 2iii: $h' > 2\underline{\eta}$, there exists $i(h')$ such that $h^{**}(i(h')) = h'$. Consumption takes place on average when $\theta > \bar{\theta}_c(h)$ where

$$\bar{\theta}_c(h') = \int_{\eta=0}^{i(h')} \alpha h' \, dG(\eta) + \int_{i(h')}^{1-E[h]} \alpha E[h \,|\, h < h^{**}(\eta)] \, dG(\eta) + \int_{1-E[h]}^{\eta^*} \alpha E[h] dG(\eta)$$

which we know to be increasing in h'. Note also that $\bar{\theta}_c(1) < \alpha$.

When $E[h] < 1/2$, the function $\bar{\theta}_c(h')$ is increasing in h' and such that $\bar{\theta}_c(0) > 0$ and $\bar{\theta}_c(1) < \alpha$. Therefore, there exist $\underline{h} \in (0,1)$ and $\bar{h} \in (0,1)$ such that $\bar{\theta}_c(h') > \alpha h$ for all $h' < \underline{h}$ and $\bar{\theta}_c(h') < \alpha h'$ for all $h' > \bar{h}$.

Assume now that $E(h) > 1/2$. There are 2 possible cases.

Case 3. $f'(h) > 0$ for all h or $f'(h) \gtrless 0$ for all $h \lessgtr \tilde{h}$ such that a **M** equilibrium does not exist. From (9), for all $\eta < E[h]$, the equilibrium is **T** and there exists $h^*(\eta)$ such that $E[h \,|\, h > h^*(\eta)] - h^*(\eta) = \eta$. For all $\eta \geqslant E[h]$, the equilibrium is **A** and $H = [0,1]$. For all h, there exists $j(h)$ such that $h^*(j(h)) = h$ and such that the equilibrium is **T** and modulation occurs at h when $\eta < j(h)$, the equilibrium is **T** but no modulation occurs at h when $\eta \in (j(h), E(h))$ and the equilibrium is **A** and no modulation occurs at h when $\eta > E[h]$. Given $h^*(\eta)$ is decreasing, $j(h')$ is decreasing. For each h', there are three consumption scenarii: (i) when $\eta < j(h')$, consumption takes place if $\theta > \alpha h'$; (ii) when $\eta \in (j(h), E(h))$, consumption takes place when $\theta > \alpha E[h \,|\, h > h^*(\eta)]$; (iii) when $\eta > E[h]$, consumption takes place when $\theta > \alpha E[h]$. Overall, for all h', consumption takes place when $\theta > \bar{\theta}_c(h)$ where

$$\bar{\theta}_c(h') = \int_{\eta=0}^{j(h')} \alpha h' \, dG(\eta) + \int_{j(h')}^{E[h]} \alpha E[h \,|\, h > h^*(\eta)] \, dG(\eta) + \int_{E[h]}^{\eta^*} \alpha E[h] dG(\eta)$$

Note that $\bar{\theta}_c(0) > 0$ and $\bar{\theta}_c(1) < \alpha$. We have

$$\frac{d\bar{\theta}_c}{dh'} = \alpha \, G(j(h')) \left[1 - \frac{g(j(h'))}{G(j(h'))} j(h') j'(h') \right] > 0 \tag{13}$$

Case 4. $f'(h) \gtrless 0$ for all $h \lessgtr \tilde{h}$ such that a **M** equilibrium exists. For all $\eta < \overline{\eta}$, the equilibrium is **M** and there exists $h^o(\eta)$ such that $E[h | h^o(\eta) \leqslant h \leqslant h^o(\eta) + 2\eta] = h^o + \eta$. For all $\eta \in [\overline{\eta}, E[h]]$, the equilibrium is **T** and there exists $h^*(\eta)$ such that $E[h | h < h^*(\eta)] - h^*(\eta) = \eta$. For all $\eta > E[h]$, then $h^* = 0$. Again, the unique health rating that solves $E[h | h^o(\eta) \leqslant h \leqslant h^o(\eta) + 2\eta] = h^o + \eta$ when $\eta \to 0$ is $\tilde{h}$.

There are three cases: (i) $h' > \tilde{h}$ and modulation occurs only for low values of η in the **M** equilibrium; (ii) $h' \in [1 - 2\overline{\eta}, \tilde{h}]$ no modulation occurs when η is high enough in

M, and modulation never occurs in the **T** equilibrium. (iii) $h' < 1 - 2\overline{\eta}$ and modulation occurs always in the **M** and only if η is not too high in the **T** equilibrium.

Case 4i: $h' > \tilde{h}$, there exists $j^{oo}(h') < \overline{\eta}$ such that $h^{oo}(j^{oo}(h')) = h'$ and modulation occurs only when $\eta < j^{oo}(h')$. Consumption takes place on average when $\theta > \bar{\theta}_c(h')$ where

$$\bar{\theta}_c(h') = \int_{\eta=0}^{j^{oo'}(h')} \alpha h' \, dG(\eta) + \int_{j^{oo}(h')}^{\overline{\eta}} \alpha E[h \,|\, h^o(\eta) < h < h^{oo}(\eta)] \, dG(\eta)$$

$$+ \int_{\overline{\eta}}^{E[h]} \alpha E[h \,|\, h > h^*(\eta)] \, dG(\eta) + \int_{E[h]}^{\eta^*} \alpha E[h] dG(\eta)$$

We have

$$\frac{d\bar{\theta}_c}{dh'} = \alpha \, G(j^o(h')) \left[1 + \frac{g(j^{oo}(h'))}{G(j^{oo}(h'))} j^{oo'}(h') j^{oo'}(h') \right] \tag{14}$$

Note that $h^{oo}(\eta)$ increases in η and therefore $j^{oo}(h')$ increases which proves that $\frac{d\bar{\theta}_c}{dh'} > 0$. Note also that $\bar{\theta}_c(1) < \alpha$.

Case 4ii: $h \in [1 - 2\overline{\eta}, \tilde{h}]$, there exists $j^{oo}(h) < \overline{\eta}$ such that $h^o(j^o(h)) = h$. Consumption takes place on average when $\theta > \bar{\theta}_c(h)$ where

$$\bar{\theta}_c(h') = \int_{\eta=0}^{j^o(h')} \alpha h' \, dG(\eta) + \int_{j^o(h')}^{\overline{\eta}} \alpha E[h \,|\, h^o(\eta) < h < h^{oo}(\eta)] \, dG(\eta)$$

$$+ \int_{\overline{\eta}}^{E[h]} \alpha E[h \,|\, h < h^{**}(\eta)] \, dG(\eta) + \int_{E[h]}^{\eta^*} \alpha E[h] dG(\eta)$$

We have

$$\frac{d\bar{\theta}_c}{dh'} = \alpha \, G(j^o)(h') \left[1 - \frac{g(j^o(h'))}{G(j^o(h'))} j^{o'}(h') j^o(h' \right] \tag{15}$$

Note that $h^o(\eta)$ decreases in η and therefore $j^o(h')$ decreases which proves that $\frac{d\bar{\theta}_c}{dh'} > 0$.

Case 4iii: $h' < 1 - 2\underline{\eta}$, there exists $j(h')$ such that $h^*(j(h')) = h'$. Consumption takes place on average when $\theta > \bar{\theta}_c(h)$ where

$$\bar{\theta}_c(h') = \int_{\eta=0}^{j(h')} \alpha h' \, dG(\eta) + \int_{j(h')}^{1-E[h]} \alpha E[h \,|\, h > h^*(\eta)] \, dG(\eta) + \int_{E[h]}^{\eta^*} \alpha E[h] dG(\eta)$$

which we know to be increasing in h'. Note also that $\bar{\theta}^e(0) > 0$.

When $E[h] > 1/2$, the function $\bar{\theta}_c(h')$ is increasing in h' and such that $\bar{\theta}_c(0) > 0$ and $\bar{\theta}_c(1) < \alpha$. Therefore, there exist $\underline{h} \in (0,1)$ and $\bar{h} \in (0,1)$ such that $\bar{\theta}_c(h') > \alpha h$ for all $h' < \underline{h}$ and $\bar{\theta}_c(h') < \alpha h'$ for all $h' > \bar{h}$.

Proof of Corollary 5

Under beliefs $\underline{p}(h)$, we have $E[h] \to 0$. For each η, modulation occurs at the top and $h^{**}(\eta) \to \eta$. Overall, consumption takes place on average when $\theta > \bar{\theta}_c(h'; \underline{p}(h'))$ where

$$\bar{\theta}_c(h'; \underline{p}(h')) = \int_0^{h'} \alpha h' \, dG(\eta) \leqslant \alpha h'$$

which is increasing in h' and such that $\bar{\theta}_c(0) = 0$ and $\bar{\theta}_c(1) < \alpha$.

Under beliefs $\overline{p}(h)$, we have $E[h] \to 1$. For each η, modulation occurs at the bottom and $h^{*}(\eta) \to 1 - \eta$. Overall, consumption takes place on average when $\theta > \bar{\theta}_c(h', \overline{p}(h'))$ where

$$\bar{\theta}_c(h', \overline{p}(h')) = \int_0^{1-h'} \alpha h' \, dG(\eta) + \int_{1-h'}^{\eta^*} \alpha \, dG(\eta) \geqslant \alpha h'$$

which is increasing in h' and such that $\bar{\theta}_c(0) > 0$ and $\bar{\theta}_c(1) = \alpha$.

Proof of Corollary 6: This corresponds to a straightforward change of variable. Note that δ^t is monotonic (decreasing) in t. Let $m = \delta^t$ and replace h by m in the previous proofs. More precisely, there is no modulation when $m \in M$ where

$$\Delta(m, E[m \,|\, m \in M]) = \int_{E[m \,|\, m \in M]}^{m} (m - \theta) d\theta > c$$

which is convex in m. Therefore M is compact, and given monotonicity of δ^t, T is also compact. The rest of the proofs are straightforward adaptations of the previous ones and therefore omitted.

Proof of Corollary 7: immediate and therefore omitted.

Appendix B

Sufficient conditions in Corollary 4

Case 1. When $\eta^* = 1$ and $G(\eta) = \eta$, we get:

$$\frac{d\bar{\theta}_c}{dh'} = \alpha\, i(h')\left[1 + i'(h')\right] \quad \text{and} \quad \frac{d^2\bar{\theta}_c}{dh'^2} = \alpha\, i'(h')\left[1 + i'(h')\right] + \alpha\, i(h')i''(h') \tag{16}$$

From the definition of $i(h')$, we have $h' - E[h|h < h'] = i(h')$, which implies:

$$i'(h') = 1 - \frac{\partial}{\partial h'}E[h|h < h'] \quad \text{and} \quad i''(h') = -\frac{\partial^2}{\partial h'^2}E[h|h < h']$$

Also:

$$E[h|h < h'] = h' - \frac{\int_0^{h'} F(h)dh}{F(h')} \quad \Rightarrow \quad \frac{\partial}{\partial h'}E[h|h < h'] = \frac{\int_0^{h'} F(h)dh}{F(h')}\frac{f(h')}{F(h')} = i(h')\mathcal{J}(h')$$

where $\mathcal{J}(h') = f(h')/F(h')$ which is decreasing in h'. Therefore

$$i''(h') = -\Big(1 - i(h')\mathcal{J}(h')\Big)\mathcal{J}(h') - i(h')\mathcal{J}'(h')$$

Using these equalities we obtain

$$\frac{d^2\bar{\theta}_c}{dh'^2} = 2\alpha\,[i'(h')]^2 - \alpha\,[i(h')]^2\mathcal{J}'(h') > 0.$$

Combining $\bar{\theta}_c(0) > 0$, $\bar{\theta}_c(1) < \alpha$, $\frac{d\bar{\theta}_c}{dh'} > 0$ and $\frac{d^2\bar{\theta}_c}{dh^2} > 0$, we get that there exists a unique value $\bar{h} \in (0,1)$ such that $\bar{\theta}_c(h') > \alpha h$ for all $h' < \bar{h}$ and $\bar{\theta}_c(h') < \alpha h'$ for all $h' > \bar{h}$.

Case 2.

Case 2i When $\eta^* = 1$ and $G(\eta) = \eta$, we have

$$\frac{d^2\bar{\theta}_c}{dh'^2} = \alpha\, i^{o'}(h')\left[1 - i^{o'}(h')\right] - \alpha\, i^o(h')i^{o''}(h') \tag{17}$$

Consider functions that satisfy the following condition P*: $f(x) = f(1-x)$ for $x < \tilde{h}$ when $\tilde{h} < 1/2$ and $f(x) = f(1-x)$ for $x > \tilde{h}$ when $\tilde{h} > 1/2$.

When P* is satisfied, we also have $h^o(\eta) = \tilde{h} - \eta$ and therefore $i^o(h') = \tilde{h} - h'$ and therefore $\frac{d^2\bar{\theta}_c}{dh'^2} < 0$.

Case 2ii: When $\eta^* = 1$ and $G(\eta) = \eta$, we have

$$\frac{d\bar{\theta}_c}{dh'} = \alpha\, i^{oo}(h')\left[1 + i^{oo'}(h')\right] \quad \text{and} \quad \frac{d^2\bar{\theta}_c}{dh'^2} = \alpha\, i^{oo'}(h')\left[1 + i^{oo'}(h')\right] + \alpha\, i(h')i^{oo''}(h') \tag{18}$$

When P^* is satisfied, we also have $h^{oo}(\eta) = \tilde{h} + \eta$ and therefore $i^{oo}(h') = h' - \tilde{h}$ and therefore $\frac{d^2\bar{\theta}_c}{dh'^2} > 0$.

Case 2iii. This case is similar to case 1. We have $\frac{d^2\bar{\theta}_c}{dh'^2} > 0$.

Case 3. Assuming $\eta^* = 1$ and $G(\eta) = \eta$, we get:

$$\frac{d\bar{\theta}_c}{dh'} = \alpha\, j(h')\left[1 - j'(h')\right] \quad \text{and} \quad \frac{d^2\bar{\theta}_c}{dh'^2} = \alpha\, j'(h')\left[1 - j'(h')\right] - \alpha\, j(h')j''(h') \tag{19}$$

From the definition of $j(h')$, we have $E[h|h > h'] - h' = j(h')$, which implies:

$$j'(h') = \frac{\partial}{\partial h'}E[h|h > h'] - 1 \quad \text{and} \quad j''(h') = \frac{\partial^2}{\partial h'^2}E[h|h > h']$$

Also:

$$E[h|h > h'] = \frac{1 - h'F(h') - \int_{h'}^{1} F(h)dh}{1 - F(h')} \;\Rightarrow\; \frac{\partial}{\partial h'}E[h|h > h'] = \frac{1 - h' - \int_{h'}^{1} F(h)dh}{1 - F(h')}\frac{f(h')}{1 - F(h')}$$

that is

$$\frac{\partial}{\partial h'}E[h|h > h'] = j(h')\mathcal{K}(h')$$

where $\mathcal{K}(h') = f(h')/(1 - F(h'))$ which is increasing in h'. Therefore

$$j''(h') = \Big(j(h')\mathcal{K}(h') - 1\Big)\mathcal{K}(h') + j(h')\mathcal{K}'(h')$$

Using these equalities we obtain

$$\frac{d^2\bar{\theta}_c}{dh'^2} = -2\alpha\left[j'(h')\right]^2 - \alpha\left[j(h')\right]^2\mathcal{K}'(h') < 0.$$

Case 4.

Case 4i: When $\eta^* = 1$ and $G(\eta) = \eta$, we have

$$\frac{d^2\bar{\theta}_c}{dh'^2} = \alpha\, j^{oo'}(h')\left[1 + j^{oo'}(h')\right] + \alpha\, j(h')j^{oo''}(h') \tag{20}$$

When P^* is satisfied, we also have $h^{oo}(\eta) = \tilde{h} + \eta$ and therefore $j^{oo}(h') = h' - \tilde{h}$ and therefore $\frac{d^2\bar{\theta}_c}{dh'^2} > 0$.

Case 4ii When $\eta^* = 1$ and $G(\eta) = \eta$, we have

$$\frac{d^2\bar{\theta}_c}{dh'^2} = \alpha\, j^{o'}(h')\left[1 - j^{o'}(h')\right] - \alpha\, j^{o}(h')j^{o''}(h') \tag{21}$$

When P* is satisfied, we also have $h^o(\eta) = \tilde{h} - \eta$ and therefore $j^o(h') = \tilde{h} - h'$ and therefore $\frac{d^2\bar{\theta}_c}{dh'^2} < 0$.

Case 4iii. This case is similar to case 3. We have $\frac{d^2\bar{\theta}_c}{dh'^2} < 0$.

Overall, when η is drawn from a uniform distribution on $[0,1]$, $\underline{h} = \overline{h}$ for distributions satisfying cases 1 and 3. When the extra property P^* holds, $\underline{h} = \overline{h}$ also for distributions satisfying cases 2 and 4.

References

1. Ainslie, G. (1975), "Specious Reward: a Behavioral Theory of Impulsiveness and Impulse Control", *Psychological Bulletin*, 82, 463-509.

2. Ainslie, G. and J. Monterosso (2003), "Building Blocks of Self-Control: Increased Tolerance for Delay with Bundled Rewards", *Journal of the Experimental Analysis of Behavior* 79(1), 37-48.

3. Alonso, R., Brocas, I. and J.D. Carrillo (2013), "Resource Allocation in the Brain", forthcoming in *Review of Economic Studies*.

4. Bagnoli, M. and T. Bergstrom (2005), "Log-concave Probability and its Applications", *Economic Theory* 26(2), 445-469.

5. Bahlmann J., F.M. Korb, C. Gratton and A.D. Friederici (2012), "Levels of Integration in Cognitive Control and Sequence Processing in the Prefrontal Cortex", *PLoS One*, 7(8), 1-10.

6. Baumgartner, T., Knoch, D., Hotz, P., Eisenegger C., Fehr E. (2011), "Dorsolateral and ventromedial prefrontal cortex orchestrate normative choice", *Nature Neuroscience* 14, 1468-1474.

7. Becker, G.S. and M.K. Murphy (1988), "A Theory of Rational Addiction", *Journal of Political Economy* 96, 675-700.

8. Bénabou, R. and J. Tirole (2004), "Will-power and Personal Rules", *Journal of Political Economy* 112, 848-886

9. Bernheim, B.D. and A. Rangel (2004), "Addiction and Cue-Triggered Decision Processes", *American Economic Review* 94(5), 1558-90.

10. Bodner, R. and D. Prelec (2003), "Self-Signaling and Diagnostic Utility in Everyday Decision Making", in I. Brocas and J.D. Carrillo (Eds.), *The Psychology of Economic Decisions (Vol. 1: Rationality and Well-being).* Oxford: Oxford University Press.

11. Botvinick M.M., T.S. Braver, D.M. Barch, C.S. Carter, J.D. Cohen (2001), "Conflict monitoring and cognitive control", *Psychological Review*, 108, 624-652.

12. Botvinick M.M.(2008), "Hierarchical models of behavior and prefrontal function", *Trends in Cognitive Sciences*, 12, 201-208.

13. Brocas, I. and J.D. Carrillo (2008), "The Brain as a Hierarchical Organization", *American Economic Review* 98, 1312-1346.

14. Brooks S.J., O.G. O'Daly, Uher R., Friederich H.C., Giampietro V., Brammer M., Williams S.C.R., Schith H.B., Treasure J., Campbell I.C. (2011), "Differential neural responses to food images in women with bulimia versus anorexia nervosa", *PLoS One*, 6(7), 1-8.

15. Brooks S.J. , M. Rask-Andersen, C. Benedict, H.B. Schith (2012), "A Debate on Current Eating Disorder Diagnoses in Light of Neurobiological Findings. Is It Time for a Spectrum Model?", *BMC Psychiatry*, 12(76), 1-18.

16. Bush G., P. Luu, M.I. Posner (2000), "Cognitive and emotional influences in anterior cingulate cortex", *Trends in Cognitive Sciences*, 4, 215-222.

17. Camerer, C. (2008), "The case for mindful economics", in A. Caplin and A. Schotter Eds., *The foundations of positive and normative economics*, 43-61.

18. Caplin A. and M. Dean (2013), "Rational Inattention, Entropy, and Choice: The Posterior-Based Approach", *mimeo, NYU.*

19. Caplin A. and D. Martin (2011), "A Testable Theory of Imperfect Perception", *mimeo, NYU.*

20. Caprioli D., Celentano M., Paolone G. and A. Badiani (2007), "Modeling the role of environment in addiction", *Progress in Neuro-Psychopharmacology and Biological Psychiatry* 31, 16391653.

21. Carrillo, J.D. and T. Mariotti (2000), "Strategic Ignorance as a Self-Disciplining Device", *Review of Economic Studies* 67(3), 529-544.

22. Clithero, J. and A. Rangel (2013) "Informatic parcellation of the network involved in the computation of subjective value", forthcoming in *Social Cognitive and Affective Neuroscience.*

23. Crockett M., B. Braams, L. Clark, P. Tobler, T. Robbins, T. Kalenscher (2013), "Restricting Temptations: Neural Mechanisms of Precommitment", *Neuron*, 79(2), 391-401.

24. Cunningham, T. (2013), " Biases and Implicit Knowledge", *mimeo*, IIES Stockholm.

25. Daw, N. and O'Doherty, J.P. (2013), "Multiple systems for value learning" in Neuroeconomics: Decision Making and the Brain 2nd ed. (eds. Glimcher, P. and Fehr, E.), Academic Press, New York.

26. De Araujo I.E., Rolls E.T., Velazco M.I., Margot C., Cayeux I. (2005), "Cognitive modulation of olfactory processing", *Neuron*, 46, 671-679.

27. Dekel, E., B. Lipman and A. Rustichini (2009), "Temptation-Driven Preferences", *Review of Economic Studies*, 76(3), 937-971.

28. Dixon M.L. and K. Christoff (2012), "The Decision to Engage Cognitive Control Is Driven by Expected Reward-Value: Neural and Behavioral Evidence", *PLoS One*, 7(12), 1-12.

29. Falk J.L. and D.A. Feingold (1987), "Environmental and cultural factors in the behavioral actions of drugs". In: Meltzer H.Y., editor. Psychopharmacology: the third generation of progress. New York: Raven Press, 1503-1510.

30. Figner, B., Knoch, D., Johnson, E.J., Krosch, A.R., Lisanby, S.H., Fehr, E., and E.U. Weber (2010), "Lateral Prefrontal Cortex and Self-Control in Intertemporal Choice", *Nature Neuroscience* 13(5), 538-539.

31. Fudenberg, D. and D.K. Levine (2006), "A Dual Self Model of Impulse Control", *American Economic Review* 96(5), 1449-1476.

32. Goldstein R.Z. and N. Volkow (2011), "Dysfunction of the prefrontal cortex in addiction: neuroimaging findings and clinical implications", *Nature reviews* 12, 652-669.

33. Gul F. and W. Pesendorfer (2001), "Temptation and Self-Control", *Econometrica* 69, 1403-1435.

34. Gul, F. and W. Pesendorfer (2008), "The case for mindless economics", in A. Caplin and A. Schotter Eds., *The foundations of positive and normative economics*, 3-39.

35. Hare, T., Camerer, C. and A. Rangel (2009), "Self-Control in Decision-Making Involves Modulation of the vmPFC Valuation System", *Science* 324, 646-648.

36. Hare, T., Malmaud, J. and A. Rangel (2011), "Focusing Attention on the Health Aspects of Food Changes Value Signals in the vmPFC and Improves Dietary Choice", *Journal of Neuroscience* 31, 11077-11087.

37. Hare, T. , O'Doherty, J. , Camerer, C. , Schultz, W., and A. Rangel (2008) "Dissociating the Role of the Orbitofrontal Cortex and the Striatum in the Computation of Goal Values and Prediction Errors", *The Journal of Neuroscience* 28(22), 5623-5630.

38. Hutcherson, C., H. Plassmann, J.J. Gross and A. Rangel (2012), "Cognitive Regulation during Decision Making Shifts Behavioral Control between Ventromedial and Dorsolateral Prefrontal Value Systems", *The Journal of Neuroscience* 32(39), 13543-13554.

39. Kable, J.W. and P.W. Glimcher (2007), "The Neural Correlates of Subjective Value During Intertemporal Choice", *Nature Neuroscience* 10(12), 1625-1633.

40. Kaye W.H., Fudge J.L., Paulus M. (2009), "New insights into symptoms and neurocircuit function of anorexia nervosa", *Nature Reviews Neuroscience*, 10(8), 573-584.

41. Kaye W.H., Wagner A., Fudge J.L. and M. Paulus (2011), "Neurocircuitry of eating disorders", *Current Topics in Behavioural Neuroscience*, 6, 37-57.

42. Kerns J.G., Cohen J.D., MacDonald A.W., Cho R.Y., Stenger V.A. and C.S. Carter (2004), "Anterior cingulate conflict monitoring and adjustments in control", *Science* 303, 1023-1026.

43. Knoch, D. and E. Fehr (2007), "Resisting the Power of Temptations. The Right Prefrontal Cortex and Self-Control", *Annals of the New York Academy of Sciences* 1104, 123-134.

44. Koechlin E., Basso G., Pietrini P., Panzer S., Grafman J. (1999), "The role of the anterior prefrontal cortex in human cognition", *Nature*, 399, 148-151.

45. Koechlin E., Corrado G., Pietrini P., Grafman J. (2000), "Dissociating the role of the medial and lateral anterior prefrontal cortex in human planning", *Proceedings of the National Academy of Sciences*, 97, 7651-7656.

46. Koechlin E., Ody C., Kouneiher F. (2003), "The architecture of cognitive control in the human prefrontal cortex", *Science*, 302, 1181-1185.

47. Koechlin E., Summerfield C. (2007), "An information theoretical approach to prefrontal executive function", *Trends in Cognitive Sciences*, 11, 229-235.

48. Kouneiher F., S. Charron and E. Koechlin (2009), "Motivation and cognitive control in the human prefrontal cortex", *Nature Neuroscience*, 12(7). 939-945

49. Laibson, D. (1997), "Golden Eggs and Hyperbolic Discounting", *Quarterly Journal of Economics* 112(2), 443-477.

50. Laibson, D. (2001), "A Cue-Theory of Consumption", *Quarterly Journal of Economics* 116(1), 81-119.

51. Luo, S, Ainslie, G. Giragosian, L., and J. Monterosso (2009), "Behavioral and Neural Evidence of Incentive Bias for Immediate Rewards Relative to Preference-Matched Delayed Rewards", *Journal of Neuroscience* 29(47), 14820-14827.

52. Masicampo, E.J., and Baumeister, R.F. (2008), "Toward a physiology of dual-process reasoning and judgment: Lemonade, willpower, and expensive rule-based analysis", *Psychological Science* 19, 255-260.

53. McClure, S., Ericson, K.M., Laibson, D., Loewenstein, G. and J.D. Cohen (2007), "Time Discounting for Primary Rewards", *The Journal of Neuroscience* 27(21), 5796-5804.

54. McClure, S., Laibson, D., Loewenstein, G. and J.D. Cohen (2004), "Separate Neural Systems Value Immediate and Delayed Monetary Rewards", *Science* 306, 503-507.

55. Mijovic-Prelec, D. and D. Prelec (2010), "Self-deception as self-signalling: a model and experimental evidence", *Philosophical Transactions of the Royal Society - B* 365(1538), 227-240.

56. Niv, Y. and Montague, P.R. (2008), "Theoretical and empirical studies of learning" in Neuroeconomics: Decision-Making and the Brain (eds. Glimcher, P.W., Fehr, E., Camerer, C. and Poldrack, R.), Elsevier.

57. Paus T. (2001), "Primate anterior cingulate cortex: where motor control, drive and cognition interface", *Nature Review Neuroscience*, 2, 417-424.

58. Petrides M. (2005), "Lateral prefrontal cortex: architectonic and functional organization", *Philosophical Transactions of the Royal Society Biological Sciences*, 360, 781-795.

59. Plassmann, H., O'Doherty, J., and A. Rangel (2007), "Orbitofrontal Cortex Encodes Willingness to Pay in Everyday Economic Transactions", *The Journal of Neuroscience* 27(37), 9984-9988.

60. Plassmann H., J. O'Doherty, B. Shiv and A. Rangel (2008), "Marketing actions can modulate neural representations of experienced pleasantness", *Proceedings of the National Academy of Sciences*, 105(3), 1050-1054.

61. Prelec, D. (1989), "Decreasing Impatience: Definition and Consequences", *working paper*, Harvard Business School.

62. Rangel A. (2013), "Regulation of dietary choice by the decision-making circuitry", *Nature Neuroscience*, 16(12), 1717-1724.

63. Reuter J., T. Raedler, M. Rose, I. Hand, J. Glascher and C. Buchel (2005), "Pathological gambling is linked to reduced activation of the mesolimbic reward system", *Nature Neuroscience* 8(2), 147-148.

64. Ridderinkhof K.R. , M. Ullsperger, E.A. Crone and S. Nieuwenhuis (2004), "The Role of the Medial Frontal Cortex in Cognitive Control", *Science*, 306 (5695), 443-447.

65. Rushworth, M.F., Buckley M.J., Behrens T.E., Walton M.E., Bannerman D.M. (2007) "Functional organization of the medial frontal cortex", *Current Opinion in Neurobiology* 17(2), 220-227.

66. Strotz, R.H. (1956), "Myopia and Inconsistency in Dynamic Utility Maximisation", *Review of Economic Studies* 23(3), 166-180.

67. Taren A.A., V.Venkatraman and S. A. Huettel (2011), "A Parallel Functional Topography between Medial and Lateral Prefrontal Cortex: Evidence and Implications for Cognitive Control", *The Journal of Neuroscience*, 31(13), 5026-5031.

68. Thaler, R. and H. Shefrin (1981) "An Economic Theory of Self-Control", *Journal of Political Economy*, 89(2): 392-406.

69. Van den Eynde F., Claudino A.M., Mogg A., Horrell L., Stahl D., Ribeiro W., Uher R., Campbell I., Schmidt U. (2010), "Repetitive transcranial magnetic stimulation reduces cue-induced food craving in bulimic disorders", *Biological Psychiatry*, 67(8), 793-795.

70. Van den Eynde F., S. Guillaume, H. Broadbent, Campbell I., Schmidt U. (2013), "Repetitive transcranial magnetic stimulation in anorexia nervosa: A pilot study", *European Psychiatry* 28, 98-101.

71. van Kuyck K., Gerard N., Van Laere K., Casteels C., Pieters G., Gabriels L., Nuttin B. (2009), "Towards a neurocircuitry in anorexia nervosa: evidence from functional neuroimaging studies", *Journal of Psychiatric Research*, 43(14), 1133-1145.

72. Vohs, K.D., Baumeister, R.F., Schmeichel, B.J., Twenge, J.M., Nelson, N.M., and Tice, D.M. (2008), "Making choices impairs subsequent self-control: A limited resource account of decision making, self-regulation, and active initiative", *Journal of Personality and Social Psychology* 94, 883-898.

73. Zinberg N.E. (1984) "Drug, set, and setting: the basis for controlled intoxicant use", New Haven: Yale University Press.