A Theory of Discrete Choice with Information Costs

Anton Cheremukhin Anna Popova Antonella Tutino*

February 2015

Abstract

We present a theory of discrete choice with information costs that supports deliberate stochastic choice. We use a unique experimental dataset to distinguish between errors arising from limitations on a decision maker's cognitive abilities and conscious disregard of information. Experimental evidence strongly favors the latter explanation. The data also allows us to directly estimate the shape and size of information costs for individual participants. Furthermore, in line with a dynamic extension of our theory, we find that accumulated knowledge of the environment improves response consistency.

JEL: D81, D03, C91, C44.

Keywords: Bounded Rationality, Information Theory, Rational Inattention, Discrete Choice, Behavioral Experiments.

*Anton Cheremukhin (corresponding author): Federal Reserve Bank of Dallas, 2200 N Pearl St, Dallas TX 75201, chertosha@gmail.com, 214-922-6785. Anna Popova: University of Illinois at Urbana-Champaign, 603 East Daniel St, Champaign, IL 61820, apopova2@illinois.edu. Antonella Tutino: Federal Reserve Bank of Dallas, 2200 N Pearl St, Dallas TX 75201, tutino.antonella@gmail.com, 214-922-6804. We are grateful to Michel Regenwetter for access to the data, encouragement and helpful comments. We also thank Tony Marley and Duncan Luce for questions and suggestions which helped improve the draft of the paper. All remaining errors are our own. Popova's work and data collection were supported by National Science Foundation grant SES # 08-20009 (PI: M. Regenwetter, University of Illinois at Urbana-Champaign), entitled *A Quantitative Behavioral Framework for Individual and Social Choice,* awarded by the Decision, Risk and Management Science Program. IRB approval (Protocol: 08387, RPI: M. Regenwetter) has been obtained for the experiment on human subjects. Any opinions, findings or recommendations expressed in this publication are those of the authors and do not necessarily reflect the views of their colleagues, the National Science Foundation, the University of Illinois, the Federal Reserve Bank of Dallas or the Federal Reserve System.

1 Introduction

For decades behavioral choice literature has confronted the challenge of modeling bounded rationality. The stochasticity of choices when a decision maker faces the same stimuli repeatedly is a particularly troubling aspect.[1] Most responses to this challenge have focused on developing probabilistic choice models that can fit the observed error distributions.[2] These approaches are unsatisfactory because they do not explain the source of the errors. Understanding whether the errors arise from a physical limitation that prevents the decision maker from making the right choice or from a lack of interest has important public policy implications. For instance, if disinterest is the reason, the decision maker could be incentivized to improve their selection; in the case of a physical bound, little can be done. This paper seeks to distinguish between errors that come from decision makers' inability to identify the better choice and conscious mistakes.

We describe a theory that rationalizes conscious errors as an outcome of a tradeoff between expending the effort to identify the superior option and realizing the potential benefit from picking that option. The theory is centered around the assumption that there is a cognitive cost associated with processing information about the options. More information requires more effort but leads to higher confidence that the choice is correct. The presence of an information cost is crucial for the stochastic nature of choice, and the shape of the cost reflects the degree to which choice is deliberate.

As a microfoundation of this cost, our framework builds on the rational inattention theory of Sims (2003). Rational inattention theory measures the amount of information processed using a precise statistical definition from information theory and postulates a cost that limits the capacity of agents to process information. This cost has been modeled either as a fixed marginal cost or a fixed capacity limit. Our framework extends the specification of the cost function to accommodate both. This makes it flexible enough to simultaneously account for physical bounds and allow for conscious errors.

The main contribution of this paper is its use of a behavioral experiment to discriminate between the two types of cognitive limitations and shed light on their relative importance. To estimate the

[1] For a discussion of these issues see among others Mosteller and Nogee (1951), Hey and Orme (1994), Hey (2001), Regenwetter et. al. (2010) and (2011).

[2] Major works on probabilistic choice models include, among others, Fechner (1860), Thurstone (1927), Luce (1959), Block and Marschak (1960), Yellott (1977) and Falmagne (1978).

shape and size of information processing costs and to put our theory on firm empirical ground, we use data from a unique behavioral experiment in which each participant is subjected to the same stimuli many times. This property allows us to directly observe the probabilities of cognitive error. Since picking the wrong option implies a utility loss, we can trace the relationship between the probability of an error and the size of the loss. The shape of this relationship informs us about the curvature of the cost function.

Our estimates of information costs suggest that the majority of participant errors arise from deliberate decisions to ignore some information. More specifically, the experimental data favors a functional form of the cost that implies no physical bounds on information processing for the majority of participants.

We are also the first to provide direct estimates of information costs, a development with far reaching implications. These estimates can serve as both a benchmark calibration in bounded rationality models in macroeconmics[3] and to facilitate appropriate selection of probabilistic choice model in the behavioral literature.

There are two additional implications of our findings. First, we find that the majority of participants in the experiment respond to incentives by processing more information and being more accurate when the stakes are higher. Thus, models in which agents can rationally adjust information processing capacity, as if facing a linear subjective cost of information, are empirically more sound than models with constraints on information in the form of fixed thresholds, commonplace in the macro literature. This finding is also useful to calibrate the shape of error when studying other experimental environments. Second, we find that pooling responses from participants in the experiment can introduce substantial bias into the estimates of information costs and perception of risk. Thus, our results provide words of caution and guidance on the interpretation of pooled responses.

To evaluate the possibility that errors may come from limited familiarity with the experimental setup, we extend our theoretical framework to account for the evolution of beliefs as individuals learn from previous answers in a repeated-choice environment. The dynamic version of our theory predicts that over time a decision maker uses their capacity to acquire information about the environment, which leads to an improvement in the consistency of choices. In agreement with this prediction, we find that the observed ability of participants to process information and the

[3]See, e.g., Rubinstein (1998); Gabaix (2012); Mankiw and Reis (2002); Mackowiak and Wiederholt (2009).

consistency of their choices gradually increase during the experiment. This finding is a warning for the experimental literature that a large number of repetitions is necessary to distinguish both probabilistic choice theories and decision theories.[4]

We contribute to three strands of literature. First, our model of discrete choice under rational inattention is directly comparable with probabilistic choice models employed by the experimental behavioral choice literature.[5] Hence, our empirical results place restrictions on the selection of probabilistic choice models when analyzing experimental data obtained in discrete choice environments. Our findings emphasize the linear cost (logit) model as the preferable model of stochastic choice, in contrast to the fixed capacity (tremble) model, and encourage experimental setups with a large number of repetitions of the same stimuli.

Second, we contribute to the active field of dynamic behavioral choice. Our dynamic results corroborate the experimental finding of Agranov and Ortoleva (2013) that stochastic choice is deliberate. An extensive overview of theories suggesting that stochastic choice is the product of optimization of multiple goals can be found in Swait and Marley (2013). However, our emphasis is on testing whether stochastic choices are deliberate.

The relationship between rational inattention theory and the logit model of discrete choice has been independently discovered by Matějka and McKay (2015). They study a special case of our static model with non-stochastic choice options and linear costs. Additional implications of that model for state-dependent choice are explored by Caplin and Dean (2013a, 2013b). Our paper extends rational inattention theory of discrete choice to repeated settings in a stochastic environment with non-linear costs and tests its predictions using experimental data. Related models of imperfect attention, such as Masatlioglu et. al. (2012) and Manzini and Mariotti (2013), differ from our approach in their consideration set formalism and their emphasis on preference revelation, while we focus on endogenizing the properties of stochastic choice.

Finally, our findings are relevant for the macroeconomic literature studying implications of bounded rationality. Much in the spirit of Gabaix (2012), we propose a tractable cost function capturing the limits of an individual's ability to process information. In contrast to the assumptions of Gabaix (2012), as well as to Mankiw and Reis (2001) and Mackowiack and Wiederholt (2009),

[4]Most experiments repeat each set of choice options no more than 5 times, and pool choices among participants. Hey and Orme (1994), Harless and Camerer (1994), Holt and Laury (2002) and Birnbaum (2008) are prominent examples among this majority, while Regenwetter et. al. (2011) belongs to rare exceptions.

[5]For a detailed discussion of these models see Section 2.4.

we find that a fixed bound on the information-processing capacity is not supported by experimental data. Meanwhile, a fixed marginal cost of processing information associated with varying capacity is in agreement with the data. Our estimates could be a starting point to calibrate information costs in macroeconomic models.

The paper is organized as follows. Section 2 introduces the model and describes its connection to models of probabilistic choice. Section 3 describes the experiment and the methodology applied to experimental data. Section 4 presents the results and compares different decision theories. Section 5 discusses the extension of the model to dynamic environments and tests its predictions. Section 6 concludes. The full description of the dynamic model and all the proofs are relegated to the appendix. Appendix C elaborates on coding and knowledge.

2 Theoretical Framework

This section formally establishes the theoretical environment of the paper. First, we describe rational inattention theory and its relationship to probabilistic choice and introduce the information processing constraint. Second, we describe how rational inattention theory works in a discrete choice environment. Third, we describe the mapping between costs of processing information and existing models of probabilistic choice.

2.1 Rational Inattention as a Theory of Probabilistic Choice

Consider a choice between two options, A and B. Standard rational decision theories predict that whenever option A is preferred to option B, the decision maker will always choose the preferred option. In reality, people make errors. Experimental evidence suggests that the frequency of choosing the preferred option is at odds with deterministic predictions of decision theories. To illustrate this contradiction, Figure 1 displays on the vertical axis the probability of choosing option A, the horizontal axis displays the value differential between the two options according to a decision theory. The dashed line shows the predictions of decision theory and the solid line represents a typical observed mean response averaged across a sample population in an experimental setting, approximated by Luce's probabilistic choice model. Experimental evidence shown in Figure 1 suggests that the frequency of making an error depends on the difference in values of the two options. The bigger this difference, the more consistent are people in their choices. To account for

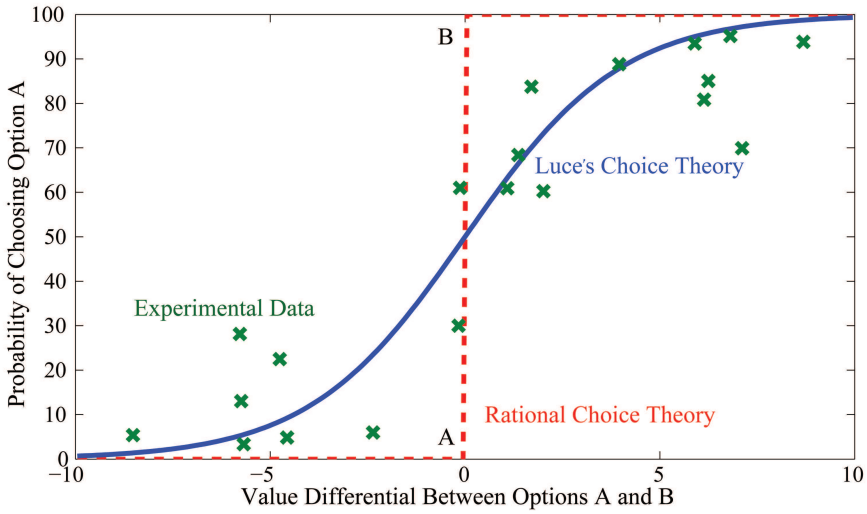

Figure 1: Data and Predictions of a Decision Theory

this mismatch between data and theory, the behavioral choice literature routinely postulates some functional form for the probability distribution of errors and augments decision theory with this statistical theory of probabilistic choice.

In this paper, we construct a model where the observed frequency of choosing option A can be fully taken into account as an outcome of rational choice. Our model builds on the fact that making a decision on which option to choose involves processing information about the options. We think of someone processing information in order to understand which option is better for them. If determining the better option requires some effort, and if the gain from finding the correct answer is not very large, the person may choose to leave some residual uncertainty about the correct answer. In that case, the person may choose the worse option with some probability.

To gather some intuition, let us consider a person working in a windowless room who must decide whether an umbrella is needed outside. Without knowing what the weather was over the past few hours, the person thinks there is a 50 percent chance of rain. To find out the precise answer the person needs to leave the room and check. However, the person is going to reduce uncertainty about the weather only if he cares. For instance, if the person plans to spend the next few hours in the room, then the current weather is immaterial. If the person just needs to get to their car in a nearby paring lot, they may opt to check the weather online. If the person plans to walk home,

they may seek out a colleague who just came in.

Whether the decision maker faces a choice of taking an umbrella or not, or among different gambles, the person will process information to reduce uncertainty to the extent to which they care about the outcome. That means that residual uncertainty may remain even after information has been processed, making the decision maker's choice prone to error. Model-wise, assuming that a person chooses the amount of information to process is equivalent to assuming that the person chooses the probability of picking each option.

The challenge in formalizing these ideas is in measuring the amount of information processed and the associated effort. We build on Shannon's (1948) information theory that defines the amount of information processed as the reduction in the uncertainty when going from a prior distribution to a posterior distribution. For instance, in our example the person deciding on whether to take an umbrella started off uninformed regarding the weather. By choosing whether to do nothing, check online or ask a colleague, the person sharpens their knowledge of the weather, which implies a posterior distribution over the possibility of rain and the likelihood of taking an umbrella. The amount of information processed, according to Shannon's information theory, is a statistical measure of distance between the prior and the posterior.

We build on rational inattention theory by associating a cost to the amount of information processed that is traded off against the decision maker's utility gain from choosing the preferred option. More specifically, we adopt the framework of the rational inattention literature of Sims (2003), (2006).[6]

Rational inattention theory's key difference from standard rational decision theories is that it allows the decision maker to rationally choose how much information to process and maps this choice onto their choice frequency. The decision maker is able to select the pieces of information deemed most relevant and ignore the rest. So long as the decision maker takes into account potential errors stemming from a disregard of information, inattentive behavior is a natural outcome of the optimizing framework postulated by rational choice theory. We now turn to the formal description of the theoretical framework of the paper.

[6] An early description of the ideas behind rational inattention theory can be found in Sims (1998). An accessible exposition of rational inattention theory can be found in Wiederholt (2010).

2.2 The Static Model

Our model builds on the model of Matějka and McKay (2015) by considering a discrete choice problem under rational inattention. While their environment studies choice among deterministic options, our model extends their setting to accommodate options of any nature. For example, as we discuss in more detail, our theory accommodates choices among gambles, the outcome of which is uncertain at the moment the choice is being made by the decision maker.

Consider a decision maker (DM) faced with a choice among K options indexed by $k \in \{1, ..., K\}$. We endow the DM with a prior distribution over the set of options denoted by$\{g(k)\}$:

$$\{g(k)\}_{k=1}^{K}, \qquad \sum_{k=1}^{K} g(k) = 1, \qquad g(k) \geq 0, \quad k \in \{1, ..., K\}.$$

We denote the posterior probability distribution over the set of options chosen by the DM by $\{s(k)\}$, where each $s(k)$ denotes her probability of choosing option k:

$$\{s(k)\}_{k=1}^{K}, \qquad \sum_{k=1}^{K} s(k) = 1, \qquad s(k) \geq 0, \quad k \in \{1, ..., K\}.$$

We use the insight from rational inattention theory that the amount of information processed by the DM is measured by the distance between the prior $g(k)$ and the posterior $s(k)$:

$$\kappa = \sum_{k=1}^{K} s(k) \log_2 \frac{s(k)}{g(k)}. \tag{1}$$

In the literature, this measure is known as Shannon's relative entropy of two distributions. As noted in Cover and Thomas (1991),[7] this formulation constitutes a special case of Shannon's Mutual Information when there is only one random variable that affects the DM's utility. The interpretation of this quantity is that the more information the DM processes about the options with respect to her original prior $\{g(k)\}$, the higher the relative entropy.

For instance, consider an experimental setting where a participant is asked to choose between options A and B. In this case, $g(k)$ represents the participant's prior beliefs about the odds that A is better than B before seeing the options. If ex ante the participant has no idea which option is better, the prior, $g(k)$, will allocate a 50 percent probability to each option. Once the options are laid out in front of the participant, they must to look carefully at each option to make an informed selection. The extent to which the participant chooses to pay attention to the options sharpens

[7] See Cover and Thomas (1991), Chapter 2.

confidence in one identified as superior. The attention that the participant pays is represented by strategy $s(k)$ and the reduction in uncertainty that that strategy achieves is given by information capacity κ. Suppose that both options are of similar value. Then staring at the options to tell them apart does not justify the effort. In this case, $s(k)$ would be close to $g(k)$, resulting in little information processed and κ close to zero. As a result, the participant will choose an option randomly. By contrast, if the options differ greatly, the participant may want to determine more precisely which one is better. In that case, $s(k)$ would place a higher probability on the better option compared with the uninformed prior. In the limit, if $s(k)$ takes on a unit probability for the better option, the amount of information processed equals 1 bit and all uncertainty is resolved.

Following rational inattention theory of Sims (2003), we model the DM's trade-off between the gain from informed choice and the cognitive effort involved in processing information by constraining the amount of information that can be processed. We assume that the decision maker has a cost associated with processing information, represented by a cost function $C(\kappa)$. The cost is an increasing convex function of the information processing capacity, κ, whose functional form is described in (1). We refer to elastic capacity as the notion that κ may vary depending on the options. The interpretation of elastic capacity is that people may choose to vary the amount of attention they pay to the options depending on the options themselves. For instance, a choice that involves a large sum of money may call for a bigger cognitive effort than a choice where a modest amount of money is involved.

We assume that the cost function enters additively into the objective function. The objective of the DM is to maximize the expected value of the options, $V(k)$, $k \in \{1, ..., K\}$, net of the subjective cost of processing information. The decision-maker's problem amounts to:

$$\max_{s(k)} \sum_{k=1}^{K} V(k)\, s(k) - C(\kappa) \tag{2}$$

s.t.

$$\kappa = \sum_{k=1}^{K} s(k) \log_2 \frac{s(k)}{g(k)} \tag{3}$$

$$s(k) \geq 0, \quad \sum_{k=1}^{K} s(k) = 1 \tag{4}$$

where $s(k)$ in the objective function (2) represents the DM's probability of choosing option k. Equation (3) computes the amount of information processed by the DM and the constraint (4)

limits the choice of the decision maker to the space of proper distributions. The following theorem characterizes the optimal solution to the decision-maker's problem in our static environment:

Theorem 1 *If the cost of information, $C(\kappa)$, is a differentiable increasing convex function of the amount of information, then the optimal choice probabilities are given by:*

$$s(k) = \frac{g(k)\exp\left(\frac{V(k)}{\theta/\ln 2}\right)}{\sum_{\tilde{k}=1}^{K} g\left(\tilde{k}\right)\exp\left(\frac{V\left(\tilde{k}\right)}{\theta/\ln 2}\right)}. \tag{5}$$

where $\theta = \frac{\partial C(\kappa)}{\partial \kappa(s(k);g(k))}$ is the derivative of the cost function with respect to the amount of information, evaluated at the chosen amount of information.

Proof. See Appendix B. ■

The key implication of the rational inattention model is that the DM chooses to behave probabilistically if information processing is costly. This remains true so long as the prior is not degenerate. As we discuss in Section 5, if we extend the model to a repeated setting, we can rationalize the prior as an outcome of the decision-maker's learning process. In that case, even in the stationary distribution, after the learning process is finished, the DM's behavior remains probabilistic.

Equation (5) represents the central testable prediction of our model. It states that the DM's choice is more precise when the difference in value between the options is more sizeable. Equation (5) also determines how the relationship between the values of the options, $V(k)$, and the choice precision, $s(k)$, depends on the cost function $C(\kappa)$. We explore this relationship in Section 2.3 and exploit it to estimate the cost function that is consistent with the experimental data in Section 4.

Another key implication of the rational inattention model is that discrete choice among any finite number of options of any nature, so long as information costs are convex, implies a choice probability distribution of the multinomial logit form as described by (5). To illustrate that the options themselves can be stochastic consider the case of options being gambles. Assume that each option $k \in \{1, ..., K\}$ has J possible outcomes with X_{kj} representing payoffs and p_{kj} respective probabilities:

$$k: \qquad \{X_{kj}, p_{kj}\}_{j=1}^{J}, \qquad \sum_{j=1}^{J} p_{kj} = 1, \qquad p_{kj} \geq 0, \qquad j \in \{1, ..., J\}.$$

In this case, the amount of information processed by the DM is captured by Shannon's Mutual Information of the joint distribution of the random variables $\widetilde{K}$ and $\tilde{J}$ that represent the choice

of gamble and the outcome of the gamble. The mutual information between $\tilde{K}$ and $\tilde{J}$ denoted as $\mathcal{I}\left(\tilde{K};\tilde{J}\right)$ is given by:

$$\mathcal{I}\left(\tilde{K};\tilde{J}\right)=\sum_{k=1}^{K}\sum_{j=1}^{J}f\left(k,j\right)\log_2\left(\frac{f\left(k,j\right)}{g\left(k\right)p_{kj}}\right),$$

where $f\left(k,j\right)=s\left(k\right)p_{kj}$ is the joint distribution of random variables $\tilde{K}$ and $\tilde{J}$. Note that the DM has no means of influencing the possible outcomes in $\tilde{J}$ or affecting the probability p_{kj}. It follows that in this case the uncertainty that is beyond the DM's control washes away and mutual information simplifies to relative entropy (1):

$$\mathcal{I}\left(\tilde{K};\tilde{J}\right)=\sum_{k=1}^{K}\sum_{j=1}^{J}s\left(k\right)p_{kj}\log_2\left(\frac{s\left(k\right)p_{kj}}{g\left(k\right)p_{kj}}\right)=\sum_{k=1}^{K}s_k\log_2\left(\frac{s\left(k\right)}{g\left(k\right)}\right)\overset{\text{def}}{=}I\left(s\left(k\right)\right),$$

where we indicate by $I\left(s\left(k\right)\right)$ the dependence of information only on the choice distribution $s\left(k\right)$ for a given prior $g\left(k\right)$.

Let us review special cases of the result in Theorem 1. For the case where the cost function is linear, the result in Theorem 1 is a generalization of the choice model of Luce (1959). To see this, assume no prior bias regarding the gambles,[8] suppose that there are $K=2$ options labeled A and B. Then the formula reduces to:

$$s\left(A\right)=\frac{1}{1+\exp\left(\frac{V(B)-V(A)}{\theta/\ln 2}\right)}. \tag{6}$$

However, our result is more general, since it can account for the DM's prior bias towards one option over the other stemming from the way options are presented and from experience. Our model provides an additional source of generality. As we discuss in the next subsection, by varying the cost function, $C\left(\kappa\right)$, we can replicate as special cases the error distributions generated by most probabilistic choice models used in the literature.

More importantly, our theory provides a rationalization to probabilistic choice models. Note that the structural forms (5) and (6) are derived from first principles. Intuitively, rational inattention theory predicts that the DM should flip a biased coin when making a selection. The bias of the coin is endogenous. It depends on the trade-off between the marginal benefit of being more attentive and the marginal cost of processing more information, captured by the expression $\theta=\frac{\partial C(\kappa)}{\partial\kappa}$. This transforms parameter θ, interpreted as the curvature of the error distribution in most existing models, into a preference parameter that characterizes the DM's costs of processing information.

[8] That is, assume that $\{g_k\}$ is uniform and equals $\{\frac{1}{K}\}$.

The goal of the empirical part of the paper is to estimate the parameters of the cost function, $C(\kappa)$. Estimates of the information cost function shed light on how likely people are to err based on the environment they face and inform us about the way people change information processing capacity in response to changes in the values of choice options.

The shape of the error distribution depends on the DM's ability to adjust the amount of information being processed. For instance, if the DM faces a capacity threshold, then the probability of making an erroneous choice would be constant, independent of the options. By comparison, if the DM can choose how much information to process by putting varying degrees of effort, then they would respond to incentives. In this case, if the DM perceives that the difference between the options is so small that it is not worth paying close attention, the DM will lack a strong preference between the two options and will choose randomly. The more the DM cares about one option over another, the more frequently they will choose the preferred option.

2.3 Probabilistic Choice Models and Information Costs

Most empirical studies of choice under risk attribute observed deviations from behavior implied by a decision theory to random errors made. Probabilistic choice models take various functional forms linking the choice probability, $s(z)$, to the value differential, z, between an option and its alternative. The value differential comes directly from decision theory.

Three major shapes of the probabilistic choice function are commonly used. First, Fechner (1860)'s model of random errors used in Hey and Orme (1994) makes use of a Gaussian cumulative density function (probit). Second, Luce (1959)'s choice model used by Holt and Laury (2002) implies a logistic curve. Third, the "tremble" model of Harless and Camerer (1994) sets the probability of a misstep to a constant, τ. There are a number of generalizations building on these three models described in Table 1.

Note that all of these models are symmetric with respect to positive and negative values of the value differential. The left panel of Figure 2.1 demonstrates that all of these shapes can be well captured by a combination of two factors. The first is the slope of the function as it passes through the point of indifference. The second factor is the asymptotic probability of a misstep, when one choice option strongly dominates another.

Both of these factors have an intuitive interpretation in our rational inattention (RI) model. Recall that our RI model with a constant marginal cost of information, $\bar{\theta}$, reproduces the logit

specification of Luce (1959). Thus, the RI model interprets the slope factor as the marginal cost of information, when the cost function is linear.

Now consider the other, more common, assumption made in the inattention literature where agents face a fixed capacity constraint, $\bar{\kappa}$. In this case the cost of information is zero for all values below $\bar{\kappa}$, but becomes vertical exactly at $\bar{\kappa}$. In this case, the RI model predicts choice probabilities identical to the tremble model. Thus, the RI model interprets a constant misstep probability as evidence of a capacity constraint on information processing.

To capture both of these factors as well as all their combinations, we adopt a flexible specification for the information cost function. We assume the following functional form:

$$C'(x) = \bar{\theta}\pi / \operatorname{arccot}\left(\frac{x - \bar{\kappa}}{\rho}\right), \tag{7}$$

where the cost of information, $\bar{\theta}$, is non-negative, the capacity constraint, $\bar{\kappa}$, takes values in the unit interval, and the curvature parameter, ρ, takes a value much higher than 1.[9]

The right panel of Figure 2 illustrates the properties of this cost function and compares it to cost functions implied by other probabilistic choice models. Note that all of the existing models can be well approximated by a combination of a constant marginal cost, $\bar{\theta}$, turning into a capacity constraint, $\bar{\kappa}$. Table 1 reports the corresponding values of these two parameters for other choice models in the literature. Note that the additional factors introduced by the Contextual Utility model of Wilcox (2011) as well as the Decision Field Theory of Busemeyer and Townsend (1993) can be interpreted as distortions of the decision theory, while the implied probabilistic choice model remains logit.[10]

2.4 Heterogeneity and Aggregation Bias

The literature on behavioral choice and the macroeconomic literature often study combined choices of all participants in an experiment or market and treat them as if coming from a single "representative" decision maker. The macroeconomic literature commonly refers to this fictional decision maker as the "representative agent." This concept is different from the average participant of the

[9] In our estimation we set $\rho = 600$. This value is high enough to capture the transition, while maximum likelihood estimation tends to set this value even higher.

[10] Although these distortions are well-specified for our experimental setup, they are hard to map directly into probability weighting functions. Our hope is that our generalized beta weighting function, described in Section 3.2, is flexible enough to meaningfully capture these distortions for our specific gamble set.

Table 1: Functional Representation of Probabilistic Choice Models

Model	Formula	Cost Function	
Fechner/Probit	$s(z) = F\left(\frac{z}{\sigma}\right)$	$\bar{\theta} \approx .41\sigma$	$\bar{\kappa} = 1$
Luce/Logit	$s(z) = \Lambda(\lambda z)$	$\bar{\theta} = 1/\lambda$	$\bar{\kappa} = 1$
Tremble	$s(z) = \left(\frac{1}{2} + \frac{2\tau - 1}{2} sgn(z)\right)$	$\bar{\theta} = 0$	$\bar{\kappa} = I(\tau)$
Truncated Fechner	$s(z) = F\left([z]_{-z_0}^{z_0}/\sigma\right)$	$\bar{\theta} \approx .41\sigma$	$\bar{\kappa} = I\left(F\left(-\frac{z_0}{\sigma}\right)\right)$
Hetero. Fechner	$s(z) = F\left(z/e^{\lambda\lvert z\rvert}\right)$	$\bar{\theta} \approx .41\sigma$	$\bar{\kappa} < 1$
Contextual Utility	$s(z) = \Lambda\left(\lambda z/\left(u\left(\overline{X}\right) - u\left(\underline{X}\right)\right)\right)$	$\bar{\theta} = 1/\lambda$	$\bar{\kappa} = 1$
Decision Field Theory	$s(z) = \Lambda\left(\lambda z/\sqrt{Var(z)}\right)$	$\bar{\theta} = 1/\lambda$	$\bar{\kappa} = 1$
Rational Inattention	$s(z) = \Lambda\left(\frac{z}{C'(I(s(z)))}\right)$	$C'(x) = \bar{\theta}\pi/\operatorname{arccot}\left(\frac{x-\bar{\kappa}}{\rho}\right)$	

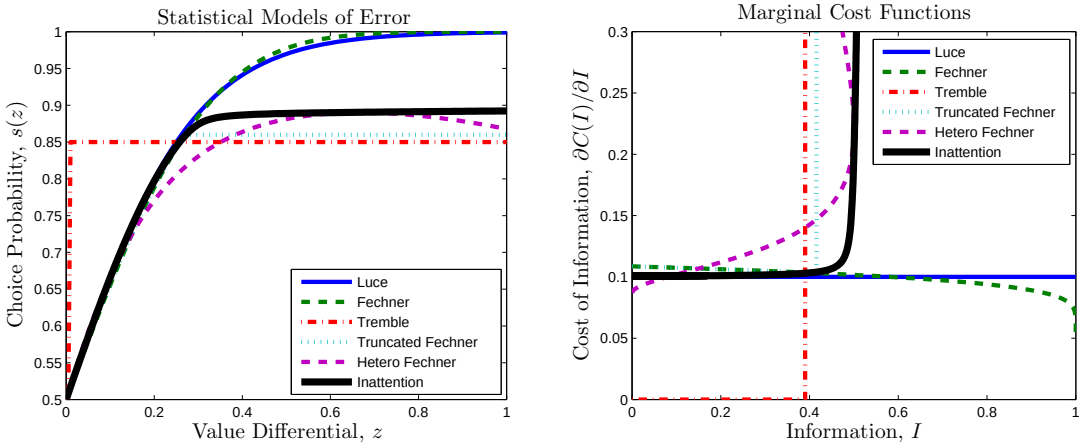

Figure 2: Error Models and Corresponding Cost Functions

experiment, i.e. the participant with average values of all parameters characterizing their behavior. We call the difference between the properties of choices of the average experiment participant and the representative agent an "aggregation bias."

The experimental data allow us to investigate the direction of aggregation bias in our sample. To this end, suppose that participants differ only in their cost of processing information, θ_i. Then, the following theorem predicts the direction of aggregation bias. The experiment allows us to measure the size of this bias.

Theorem 2 *When agents differ in their cost of information θ_i, the inverse cost of information of the representative agent is always biased downwards compared with the average across inverse costs of information of individual agents:*

$$\frac{1}{\theta_{RA}} < \frac{1}{N}\sum_{i=1}^{N}\frac{1}{\theta_i}.$$

Proof. See Appendix B. ■

The key implication of Theorem 2 is that we should expect aggregate behavior to appear as if produced by a more inattentive representative agent relative to the average individual. Thus, we should expect to encounter more inattentive behavior in the aggregate than in individual data.

3 Methodology

3.1 Experimental Setup

In order to assess the relevance of the cost of information and compare the models described in the previous section, we use experimental data collected in Michel Regenwetter's laboratory at the University of Illinois at Urbana-Champaign in Summer 2009. Each participant was asked the same question repeatedly a large number of times, the main property of interest in the experimental setup. The collected data contain observed frequencies with which each subject chose from each pair of gambles as well as the sequence of questions and answers. These data allow us to test rational inattention theory at the individual level.

Experimental data contains answers of $\mathcal{N}$ individuals who were repeatedly asked to compare $\mathcal{M}$ pairs of gambles. Each individual faced each gamble $\mathcal{L}$ times. The $\mathcal{M} \times \mathcal{L}$ overall gambles per individual were shuffled to ensure that memory effects did not impact the experiment. It

was conducted in a laboratory space at the University of Illinois at Urbana-Champaign. Forty individuals participated in the study, roughly evenly split by gender, all approximately of college age. The experiment was conducted over two sessions, separated by at least four days for each participant. Each session was not time constrained, taking roughly one hour to complete. At the beginning of each session, a clear description was given of what to expect from the experiment and several practice gambles were played. The participants were also warned that each pair of gambles could be selected at the end of the session to be played for real, making clear that their choices could affect their final payoff.

Each question contained two gambles, A and B, with parameters $\{X_1, p_X, X_2,\ 1 - p_X\}$ and $\{Y_1, p_Y, Y_2, 1 - p_Y\}$.[11] Gambles were randomly uniformly drawn from the whole domain of potential gambles following the procedure proposed by Rieskamp (2008).

One advantage of this procedure is that by construction, it does not favor any particular theory. Thus, the gamble space generates no a priori bias to the estimates of the parameters.[12] Second, this gamble selection procedure guarantees that costs associated with coding information about gambles to be processed are roughly the same for all gambles. This puts all choices on equal ground and eliminates bias associated with the possibility of inefficient coding.

Gamble outcomes were selected from a uniform distribution over [0,30] in 0.01 increments. Probabilities were selected from a uniform distribution over [0,1] in 0.01 increments. About 59% of these gambles were screened because either one gamble showed first-order stochastic dominance over the other or one gamble had at least double the expected value of the other. 20 pairs were randomly selected from the remaining gambles. Table 2 presents the gambles used in the experiment.

Participants were presented with a sequence of gamble pairs, one pair at a time. Probabilities were displayed in the form of pie charts. Participants could choose only one gamble from each pair. Gamble pairs were ordered by the computer quasi-randomly, i.e. drawn from a uniform distribution, with the condition that the same pair never be presented in succession. Over the course of a session, each gamble pair was presented 30 times, so participants made 600 choices in each of the two sessions. At the end of each session, a computer randomly selected one gamble

[11] In the dataset $\mathcal{N} = 40$, $\mathcal{M} = 20$, $\mathcal{L} = 60$.

[12] This setup is important. For instance, consider the classical Experiment I from Tversky (1969). The gambles for that experiment were selected in a subspace of all gambles which have almost the same expected payoff. Using such a set of gambles in our experimental setup would give us no ability to identify parameters characterizing costs of information.

Table 2: Gamble Payoffs and Probabilities

$\mathcal{N}_{-}^{o}$	$X_1, \$$	p_X	$X_2, \$$	$Y_1, \$$	p_Y	$Y_2, \$$
1	29.38	0.65	1.19	18.00	0.68	3.21
2	27.98	0.42	18.89	25.44	0.47	3.90
3	26.44	0.52	1.92	26.03	0.34	5.77
4	25.05	0.24	24.01	25.32	0.66	10.56
5	23.64	0.71	10.78	25.03	0.98	6.86
6	20.76	0.80	11.61	12.42	0.93	8.14
7	19.38	0.23	2.46	12.57	0.96	0.73
8	18.02	0.39	4.97	15.01	0.49	14.26
9	16.66	0.60	9.03	16.32	0.19	10.87
10	19.58	0.48	15.17	26.39	0.45	10.07
11	13.88	0.41	5.05	8.91	0.70	8.67
12	29.83	0.38	12.47	25.10	0.85	22.74
13	21.78	0.72	11.16	21.30	0.66	20.91
14	9.61	0.17	6.49	9.87	0.31	4.17
15	16.11	0.20	8.10	22.75	0.13	6.18
16	6.88	0.53	6.69	13.86	0.90	0.96
17	24.08	0.90	5.02	23.74	0.07	14.41
18	18.56	0.99	1.70	27.68	0.97	2.16
19	22.51	0.88	0.00	19.30	0.71	0.73
20	22.57	0.70	0.12	11.53	0.79	2.81

out of the 600 the participant had chosen and played for real. The outcome of the computer picks was paid to the participant together with a $5 flat payment. The average payment was $20.97 per session.

3.2 Decision Theories

Our theory of rational inattention complements decision theory, which determines valuations of choice options $V(k)$ depending on the payoffs X_{kj} and their objective probabilities, p_{kj}. In this paper we estimate individual information cost functions considering several decision theories for two-branch gambles. We follow the literature in assuming that individual valuations of sure outcomes are given by the utility function with constant relative risk aversion:

$$U(X) = \alpha \frac{X^{1-\gamma}}{1-\gamma}, \tag{8}$$

where α is a positive constant, $\gamma \in R$ represents risk-aversion of the decision-maker.[13]

Most existing decision theories, when applied to our setup, can be expressed as particular forms of the rank-dependent utility (RDU) model, developed by Quiggin (1982). RDU models commonly assume that the value of an option is determined as a weighted sum of utilities of payoffs:

$$V(k) = \sum_{j=1}^{J} w_j(p_{kj})\, U(X_{kj}), \tag{9}$$

but vary in their probability weighting function, $w_j(p)$. In the case of two-branch gambles, rank-dependence shows itself in the assumption that the weight $w(p)$ corresponds to the branch with a higher payoff, while the weight $1 - w(p)$ is attached to the lower payoff.

Prominent special cases include expected utility (EU) theory of von Neumann and Morgenstern (1944), where the weights are equal to objective probabilities, cumulative prospect theory (CPT) of Tversky and Kahneman (1992), the transfer of attention exchange (TAX) model of Birnbaum and Chavez (1997). Table 3 describes the various functional forms for the weighting function adopted in the literature. To allow for the possibility of each of these functional forms simultaneously, we extend Wilcox's (2010) beta weighting function by attaching a scale parameter to it. In our analysis,

[13] We also tried a more general specification of utility used by Holt and Laury (2002) which adds global absolute risk aversion to the utility function. We found that this specification does not improve the fit of the model.

Table 3: Weighting Function in RDU

Decision Theory	Weighting Function
EU, Von Neumann, Morgenstern (1944)	$w(p) = p$
Karmarkar (1979)	$\frac{w(p)}{1-w(p)} = \left(\frac{p}{1-p}\right)^{\phi} \left(\frac{\delta}{1-\delta}\right)^{1-\phi}$
Kumaraswamy (1980)	$w(p) = 1 - \left(1 - p^{\delta}\right)^{\phi}$
CPT, Tversky, Kahneman (1992)	$w(p) = \frac{p^{\phi}}{\left(p^{\phi}+(1-p)^{\phi}\right)^{\frac{1}{\phi}}}$
Goldstein, Einhorn (1987) Lattimore, Baker, Witte (1992)	$w(p) = \frac{\delta p^{\phi}}{\delta p^{\phi}+(1-p)^{\phi}}$
Wu, Gonzalez (1996)	$w(p) = \frac{p^{\phi}}{\left(p^{\phi}+(1-p)^{\phi}\right)^{\delta}}$
TAX, Birnbaum, Chavez (1997)	$w(p) = \frac{\delta p^{\phi}}{p^{\phi}+(1-p)^{\phi}}$
Prelec (1998)	$w(p) = \exp\left(-\delta\left(-\ln p\right)^{\phi}\right)$
Wilcox (2010)	$w(p) = B(p,\phi,\eta)/B(\phi,\eta)$
Generalized Beta	$w(p) = \left[\delta B(p,\phi,\eta)/B(\phi,\eta)\right]^{1}$

we adopt the following generalized beta weighting function:

$$w(p) = \min\left\{\delta \frac{\int_0^p (x)^{\phi-1}(1-x)^{\eta-1}\,dx}{\int_0^1 (x)^{\phi-1}(1-x)^{\eta-1}\,dx}, 1\right\}, \tag{10}$$

where behavioral parameters ϕ, η and δ take positive values. Our specification boils down to expected utility when $\phi = \eta = \delta = 1$. Also, when $\gamma = 0$ agents are risk-neutral.

Although we are unaware of closed-form expressions converting parameters of other decision theories into these parameters, it is possible to find a parameter combination for the generalized beta function that represents each of these decision theories with a high degree of accuracy. However, our functional form is more general: Under many parameter values, none of the existing decision theories can approximate choices implied by our specification.

3.3 Estimation of Parameters

We have four parameters, $\{\gamma, \phi, \eta, \delta\}$, which fully capture most existing decision theories, and two parameters of the generalized cost function (7), $\left\{\bar{\theta}, \bar{\kappa}\right\}$, which summarize the DM's limited ability to process information. Parameter α in equation (8) does not appear in the list because it scales both the utility function and the cost of information. It can be removed by converting the cost of

information θ from utils per bit to dollars per bit using a conversion factor: $\sum_{k=1}^{K} g(k) \frac{\partial V(k)}{\partial X_k}$ for each gamble. This is consistent with intuition, as the individual scale of utility affects the absolute cost of information in utils per bit, but does not affect the relative cost measured in dollars per bit.

The experimental data allows us to estimate jointly the values of all six parameters $\left\{\bar{\theta}, \bar{\kappa}, \gamma, \phi, \eta, \delta\right\}$ for each of $\mathcal{N}$ individuals, or any subset of parameters for any restricted version of the theory. The likelihood function of the data is the density of a binomial distribution where $s_{a,i}$ denotes the binomial choice probability of participant a on question i. The log likelihood of option A being chosen x times and option B being chosen y times given the deep parameters $\omega_a = \left\{\bar{\theta}, \bar{\kappa}, \gamma, \phi, \eta, \delta\right\}$ and parameters of the question $\zeta_i = \{X_1, X_2, p_X, Y_1, Y_2, p_Y\}$ is given by

$$\log L(x, y|\omega, \zeta) = \log \binom{y}{x+y} + x \log s_{a,i}(\omega, \zeta) + y \log (1 - s_{a,i}(\omega, \zeta)). \tag{11}$$

The choice probability, $s_{a,i}(\omega_a, \zeta_i)$, can be computed by solving numerically the equation:

$$s_{a,i}(\omega_a, \zeta_i) = \frac{1}{1 + \exp\left(-\frac{V(A)-V(B)}{C'(I(s_{a,i})|\omega_a)/\ln 2}\right)}, \tag{12}$$

where $C'(I|\omega_a)$ is the marginal cost function of participant a defined in (7), and $I(s_{a,i})$ denotes the amount of information implied by the choice probability $s_{a,i}$:

$$I(s_{a,i}) = s_{a,i} \log_2 s_{a,i} + (1 - s_{a,i}) \log_2 (1 - s_{a,i}) + 1. \tag{13}$$

Because our specification of the marginal cost function is convex, Theorem 5 implies that the solution of equation (12) is unique. We use this specification to estimate ω_a by maximizing the sum of log likelihoods of choices made by participant a defined as:

$$\Lambda_a = \sum_{i=1}^{\mathcal{M}} \log L(x_i, y_i|\omega_a, \zeta_i). \tag{14}$$

To compare models we sum up individual likelihoods, and then penalize the joint likelihood for over-parameterization using the Bayes Information Criterion (BIC) and the Akaike Information Criterion:

$$BIC = -2 \sum_{a=1}^{\mathcal{N}} \Lambda_a + n \ln O, \tag{15}$$

$$AIC = -2 \sum_{a=1}^{\mathcal{N}} \Lambda_a + 2n, \tag{16}$$

where n is the total number of estimated parameters, and O is the total number of observations.

To compare nested model specifications, where we allow participants to change a subset of parameters across two sessions, we use the likelihood ratio test that follows a chi-squared distribution with the number of restrictions, r, determining degrees of freedom:

$$LR = -2\left(\Lambda_a^R - \Lambda_a^U\right) \sim \chi^2(r). \tag{17}$$

4 Results

We start by describing static estimates of the parameters of cost functions. We use these estimates to study the properties of the cost functions to identify participants that face information processing capacity constraints. We describe the amount of heterogeneity in cost functions and measure the size of aggregation bias. We discuss in detail the estimates of parameters of decision theories, concluding that there are large variations in these estimates as well.

The first three columns of each panel in Table 4 report the estimates of parameters of the marginal cost function for all 40 participants in the experiment. For each participant we report the inverse of the estimated value of the marginal cost of information, $1/\hat{\bar{\theta}}$, measured in bits per cent, and the estimated value of the capacity constraint, $\hat{\bar{\kappa}}$, in bits.

As the first observation, we note that variations in the estimates of the capacity constraint are not particularly large across individuals; most estimates are indistinguishable from 1. The fourth column of each panel in Table 4 reports the likelihood ratio test statistic for the hypothesis that the capacity constraint is absent, i.e. $H_0 : \bar{\kappa} = 1$. Each test statistic has a chi-squared distribution with 1 degree of freedom. However, rejecting the null is not sufficient to conclude that a participant has a capacity constraint. We need to check two additional conditions.

First, we verify that the estimated value of the capacity is sufficiently below 1 to be meaningful. Note that if a participant accidentally made a single misstep while answering the remaining 59 repetitions of the question in line with decision theory, we would conclude that the participant processed $I(59/60) = 0.877$ bits per question. Hence, any $\bar{\kappa} > 0.877$ is indistinguishable from having no capacity constraint and answering all 60 questions consistently.

Second, we verify that the estimated value of $\bar{\kappa}$ is achieved at least theoretically in a few questions in our experiment. The low estimate of the capacity and the rejection of the null may be driven by restrictions we place on the functional form of the cost function, rather than evidence of the presence of a capacity constraint.

In Table 4, we mark with an † sign the participants with a capacity constraint, which we identified by checking three criteria: 1) that the likelihood ratio is above the critical value of 5.0; 2) that the estimated capacity is below 0.877; and 3) that the capacity constraint is achieved for at least 3 questions in our sample. We find that 12 of 40 participants of our experiment satisfy these conditions. Although these conditions might appear stringent at first glance, relaxing any one of the two additional requirements does not add more than a couple participants to the list.

The first conclusion that we draw from our estimates is that the number of participants with a capacity constraint does not exceed one third. This implies that the majority of participants respond to incentives by making more consistent choices when stakes are higher. Even those participants for whom we identify a capacity constraint have a positive cost associated with lower values of capacity. This implies that all participants respond to incentives for a large interval of value differentials. Only when stakes are especially large does the capacity constraint prevent some participants from further increasing the precision of their choices.

We observed that the estimates of marginal costs of information differ by more than an order of magnitude across participants of the experiment. Even after removing clear outliers, the set of estimates covers the whole range between 1.8 bit per cent and 25 bits per cent. This finding together with Theorem 2 suggests that we should expect some aggregation bias in estimates for the representative agent.

The cost estimates for the "representative agent" (RA), i.e. from treating the combined set of choices of all 40 participants as if coming from a single decision-maker, are reported at the bottom of Table 4. We find that the RA's marginal cost of information is 4.7 bits per cent, and has a capacity constraint of 0.73 bits. These estimates are in sharp contrast with the mean of individual costs of 15.3 bits per cent (median of 8.9) and the mean capacity constraint of 0.85 bits (median of 0.91). Restricting the comparison to subgroups of participants does not undermine this result.

Our second conclusion from the static estimates is that using combined choices of different participants introduces a substantial aggregation bias into an average individual's estimates of the cost function. As predicted by Theorem 2, the aggregation bias makes choices of the representative agent much less consistent relative to those of the average individual. The observation that the combined choices of individuals are inconsistent is often interpreted as showing that the average participant is very inconsistent and has substantial rationality limitations. We show that this observation may be a consequence of aggregation bias.

Table 4: Fixed capacity vs. fixed costs

$\mathcal{N}_-^o$	$1/\hat{\bar{\theta}}$	$\hat{\bar{\kappa}}$	LR	LR_θ	LR_κ	$\mathcal{N}_-^o$	$1/\hat{\bar{\theta}}$	$\hat{\bar{\kappa}}$	LR	LR_θ	LR_κ
1	16	.88	31	9.0§	9.6§	21	3.5	.47	8.6†	0.8	12§
2	1.8	.22	36†	0	1.3	22	5.0	.58	27†	5.3§	20§
3	5.9	.93	7.4	14♣	3.5	23	11	.88	9.7	38§	11§
4	7.6	1	0.4	0.3	0	24	5.4	.71	23†	34♣	25♣
5	25	1	0	18§	0	25	3.9	.32	40†	65♣	150♣
6	16	1	0.3	0.1	0	26	9.5	1	0.3	44§	0
7	12	1	0.2	0	0	27	6.5	.91	4.5	19§	5.0§
8	5.1	.95	1.1	7.7§	1.5	28	6.9	.83	30†	48§	24§
9	8.9	.93	18	0.3	2.5	29	14	1	0.1	9.6§	0
10	10	.44	65†	12§	0.6	30	6.8	.91	9.3	0.3	0.1
11	8.9	1	0.4	0	0	31	9.1	.74	22†	5.8§	1.6
12	51	.89	102	21	5.8	32	9.4	.75	22†	46§	11§
13	10	.75	36†	0.3	3.0	33	10	1	0.4	0	0
14	4.4	.43	62†	2.6	0	34	38	.95	47	68§	10§
15	3.4k	.49	169	0	62§	35	18	.91	11	0.6	0
16	11	.90	28	0.1	1.0	36	8.8	.69	29†	4.9	24§
17	5.8	.63	24	21§	0.8	37	10	1	0.2	22♣	0
18	6.9	.92	0.9	0	6.3§	38	7.0	1	0.5	0	0
19	566	.94	40	0	0	39	8.1	1	0.2	11§	0
20	8.8	.96	5.7	129§	5.5§	40	5.9	1	0.4	4.6§	0
RA	4.7	.73	443†	114§	0.1						

-N_-^o participants, $1/\hat{\bar{\theta}}$ -inverse of the estimated marginal cost of information in bits per cents, $\hat{\bar{\kappa}}$ estimated capacity in bits, LR - Likehood ratio test; LR_θ -LR test for the hypothesis that costs are equal across the two sections, LR_κ - LR test for the hypothesis that capacities are equal across the two sections. RA -representative agent, † - capacity constraint present (12 participants), § - lower information cost in second session (21 participant), ♣ - higher information cost in second session (4 participants). Likelihood ratio test statistics are distributed as $\chi^2(1)$. The critical values for this test are 3.8 at 5%, 5.0 at 2.5% and 6.6 at 1%. k means 10^3.

Table 5: Estimates of parameters of decision theories

$\mathcal{N}_{-}^{o}$	$\hat{\gamma}$	$\hat{\phi}$	$\hat{\delta}$	$\hat{\eta}$	$\mathcal{N}_{-}^{o}$	$\hat{\gamma}$	$\hat{\phi}$	$\hat{\delta}$	$\hat{\eta}$
1	0.62	5.57	4.78	0.85	21	-1.96	2.71	2.11	1.00
2	-2.27	3.81	2.88	1.83	22	0.42	0.68	1.49	0.89
3	-1.22	3.01	2.37	0.54	23	0.29	0.56	2.10	0.66
4	-0.63	0.44	0.41	0.70	24	-0.94	0.43	0.33	0.90
5	0.56	1.62	1.57	0.62	25	-1.68	4.02	3.72	0.92
6	0.22	2.37	1.97	0.59	26	1.80	1.38	1.09	0.76
7	-0.70	0.44	0.49	1.14	27	0.17	2.85	2.13	0.58
8	-0.41	1.03	0.78	0.31	28	0.29	5.05	4.35	0.52
9	-0.83	1.58	1.63	0.89	29	0.71	1.87	1.90	0.56
10	1.50	0.14	0.86	0.57	30	-1.54	0.42	0.00	3.8k
11	-0.18	1.08	1.46	0.84	31	1.06	0.21	0.59	0.69
12	-0.14	0.33	0.00	2.3k	32	-1.79	5.16	5.01	0.62
13	1.12	0.38	0.75	0.80	33	0.73	0.14	0.10	1.40
14	-7.5	174	176	1.00	34	0.53	0.32	0.00	1.7k
15	-1.08	0.59	0.96	0.52	35	1.94	1.74	1.61	0.79
16	-0.14	0.60	1.11	0.90	36	-3.23	3.84	3.34	0.60
17	0.77	0.58	1.06	0.79	37	0.78	1.61	1.73	0.54
18	-1.64	2.02	0.88	0.56	38	-1.23	2.98	2.86	0.32
19	-0.22	0.42	0.67	0.67	39	0.52	0.55	1.19	0.83
20	0.18	1.93	1.83	1.00	40	-0.54	9.9	15.9	0.31
RA	0.11	1.10	1.16	0.70					

Table 5 shows the estimates of parameters of decision theories for each participant over the whole sample. Variations in the risk-aversion parameter $\hat{\gamma}$ are quite substantial. Risk aversion covers a wide interval of values — from as high as +2— a relatively high level of risk-aversion compared to other experimental studies,[14] to as low as -2, which indicates a strong risk-loving attitude. However, there is only a mild difference between the average estimate of risk aversion, -0.39, and the RA's risk aversion value, 0.11.

There are similarly large variations in all the other parameter estimates. Estimates of the curvatures of likelihood functions indicate that only a tiny fraction of variation in these parameters across subjects can be attributed to measurement error. Most point estimates of parameters of the decision theory have relatively small standard errors.

One indication of variations in RDU parameters is the ability of different weighting functions from the literature to capture the observed weighting functions. We can roughly break down 40 participants into two groups. The first consists of participants for which the weighting function can be well approximated by a functional form from the literature. The functional forms that we find to fit best are TAX (15 participants), CPT(6 participants) and Prelec (2 participants). The second group includes 17 participants who cannot be approximated well (within 2 percent root mean-square error, RMSE) by any existing versions of RDU.

In addition to the generalized-beta specification of preferences, we redid the entire estimation exercise for an expected utility model. We find that all of our results hold in the EU specification as well: 1) the linear cost model fits most participants better than the fixed capacity model; 2) heterogeneity in cost and risk-aversion parameters is slightly bigger; 3) the dynamic behavior is very similar to that estimated in section 5 under the RDU specification.

Table 6 compares the fit of three models based on loglikelihood (LL), Bayesian Information Criterion (BIC) and Akaike Information Criterion (AIC): our generalized-beta specification of rank-dependent utility (RDU) and the expected utility model (EU). Overall, both the BIC and the AIC for RDU are much lower than that for EU (a lower value indicates better fit). We find that 36 out of 40 participants are better described by the RDU specification than by the EU specification, while for the other four participants the EU specification is more parsimonious.

Overall, we find strong evidence in favor of the rank-dependent model for most experiment

[14]The standard estimates of Holt and Laury (2002), who aggregate across subjects, are in the range of positive 0.3-0.5.

Table 6: Model fit across decision theories

Model	LL	n	BIC	AIC
1. EU	-11795	120	24884	23830
2. RDU	-7727	240	18041	15934

Legend: EU -Expected Utility theory, RDU -Rank-Dependent Utility theory, LL -log-likelihood, BIC -Bayes' Information Criterion, AIC - Akaike Information Criterion.

Note: 36 out of 40 participants are better described by RDU than EU whereas for 4 participants EU is more parsimonious.

participants. Meanwhile, the behavior of the RA is barely distinguishable from that predicted by EU model. The weighting function of the RA can be well approximated by a straight line that discounts each probability at a constant rate of 0.7. Recall that we estimated the RA's risk aversion parameter to be close to zero.

Heterogeneity among participants may be the main reason has been so hard to test and compare models of RDU in existing studies. The differences in parameters for participants of decision theories are so large that most of them would be attributed to noise if we pooled together the choices of all participants.

5 Dynamic Extension and Its Implications

The experimental setup with repeated questions introduces an additional dimension of information relevant for decision-making that participants can acquire. Before a participant faces the first question he has little knowledge of how many different pairs of options there are, the types of options, and their value differentials. As the experiment unfolds the participant accumulates knowledge about the experimental setup. Specifically, the participant may notice the number of unique questions and the likelihood of encountering them again. Accumulating this knowledge is beneficial because it can sharpen perception and reduce the likelihood of an error. This incremental processing of information must be taken into account in order to assure better assessment of conscious errors. This suggests that participants' choices should be less precise at the beginning of the experiment than at the end. As a result, we might confuse errors stemming from processing information about the value of the options with errors due to fuzzy knowledge of the environment. Discriminating between these two types of error requires a dynamic model.

In Appendix A we describe an environment where the DM faces questions repeatedly. The DM

starts with a prior perception that she rationally updates using information acquired answering the previous questions. We show that in this setting, the evolution of the decision maker's knowledge is closely related to the stochastic process from which the questions are drawn. Nevertheless, if questions are generated using a stationary transition process, then the prior converges to its stationary distribution. This is exactly the case for our experimental setting.

The dynamic version of our model predicts that the DM should behave probabilistically in the limit and that choice probabilities may evolve over time, eventually converging to a stationary distribution. Our model has closed-form predictions for this stationary distribution of choices, represented by the static model. Using the predictions of the static model, it is possible to use experimental data to uncover all of the DM's deep parameters.

However, it is much harder to find the mapping between these deep parameters and the dynamic changes in behavior before convergence. This is because the DM may start the experiment with different prior biases, and the DM's observed choices are insufficient to make inferences about this prior bias. Finally, it is hard to know the speed of convergence to the stationary distribution ex ante without knowing all the deep parameters and the prior bias. We can only hope that 60 repetitions of each question are enough for convergence to be achieved by the end of the experiment. For this reason, we do not attempt to estimate the dynamic model. Instead, we use the static model as an empirical tool. While precise in the limit, the static model serves as a good approximation of the DM's behavior before convergence has been achieved.

We use the first-order conditions of the dynamic model to simulate the model starting from different priors converging to the true stationary distribution. Using Monte-Carlo simulations, we show that this predicted behavior of beliefs maps uniquely into the dynamic behavior of rolling-window estimates of information costs. Specifically, if beliefs converge monotonically from some initial beliefs towards a uniform distribution, then the cost estimates decline over the course of the experiment converging to the true value of costs in the limit.

There are two main implications of the dynamic model.

The first implication is that acquaintance of the DM with the experimental setup lowers our estimates of the DM's costs of information. In particular, if two identical experimental sessions are confronted, our estimates of information costs should be lower in the second session than in the first session. This is because knowledge of the statistical properties of the experimental setup acquired in the first session affects the prior bias with which the DM enters the second session. This acquired

knowledge should make decisions sharper and more consistent in the second session.

The second implication is that, as the experiment unfolds, people make monotonically sharper and more consistent choices. This dynamic prediction involves the behavior of the DM within the same experimental session. The estimate of costs of information should monotonically decrease within an experimental session, while the consistency of choices should monotonically increase. This is because the dynamic process of updating the prior via Bayes' rule, predicted by our model, implies monotone convergence of the prior toward the uniform stationary distribution.

The first goal of the empirical analysis is to test the prediction of the dynamic model — if participants acquire information about the experimental environment then their estimated costs of information should fall between the two sessions. The second goal is to study the speed of convergence to the stationary distribution of attention.

Our inference proceeds in three steps. In the first step, we apply the static model to the whole dataset as if the DM starts in the stationary distribution. Estimates of the parameters of the DM's decision theory obtained this way should be close to the true parameters, because the decision theory determines the ordering of options, which remain unchanged over the course of the experiment.

Because more information capacity is diverted towards learning about the environment at early stages of the experiment, our estimates of the cost function give an upper bound on the costs of processing information rather than a precise estimate.

In a second step, we use the static model to estimate cost parameters separately for the two sessions, while keeping the parameters of decision theory constant across sessions for each participant. Then, we can test whether and how the parameters characterizing costs of processing information change between two sessions. This estimation procedure allows us to better estimate the cost function once convergence has been achieved, to test whether there is a difference in estimates between the two sessions and check whether this difference is consistent with predictions of our theory.

In a third step, we fix the parameters of the decision theory for each participant and run a rolling-window estimation of the cost function. This procedure allows us to get a good idea whether convergence has been achieved, the speed of convergence and its direction.

Note, that the design of the experiment implies a uniform transition function for answers. Thus, the probability of seeing any option on the left side of the computer screen equals the probability of seeing it on the right side. This experimental design eliminates any prior bias with respect to answers in the stationary distribution and validates the use of a uniform prior distribution when estimating

the static model. Thus, in the empirical part we can also abstract from concerns associated with possibly inefficient coding of information by participants and its effect on prior bias.

We use our dynamic model prediction that the DM's prior beliefs about the state variable and the transition process are updated during the experiment and converge to the true stationary distribution and true transition process. Combining this prediction with the first-order conditions in Theorem 1 allows us to simulate the dynamic model starting from different prior biases that then converge to the true stationary distribution. The data is not rich enough to infer the behavior of individual beliefs. To generate testable predictions, we apply a rolling-window estimation procedure both to artificial data generated by the model and to experimental data. Comparison of the two allows an indirect test of the model's predictions.

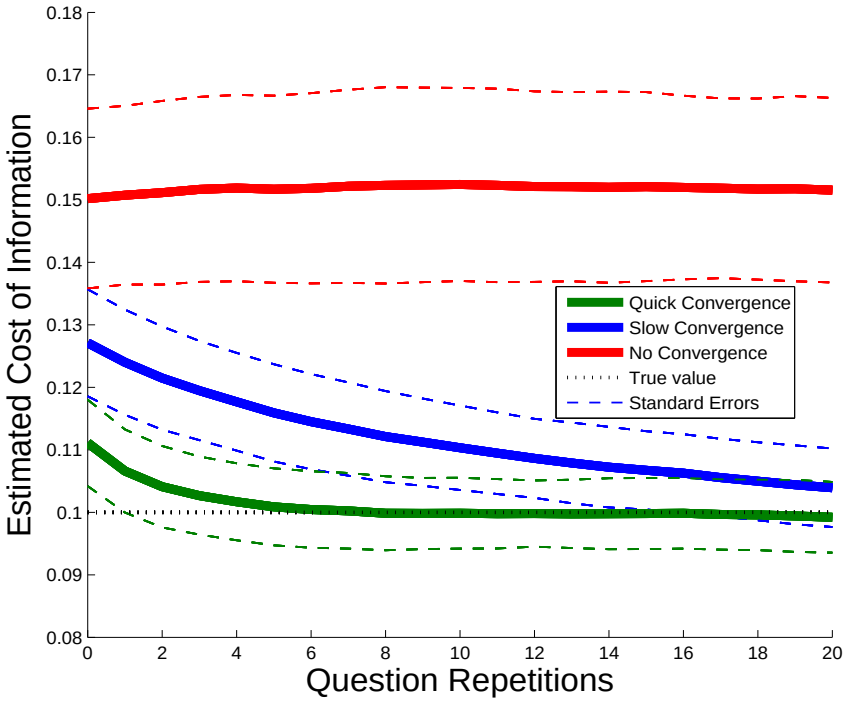

Figure 3: Monte-Carlo Simulations

Figure 3 illustrates the results of Monte-Carlo simulations. It shows three paths of rolling-window estimates of the cost of information (and confidence bounds around them), which differ only by the speed of convergence. Monte-Carlo simulations show that the changes in beliefs are captured by the dynamic behavior of rolling-window estimates of information costs. We find that if artificial beliefs converge monotonically from some initial values toward a uniform distribution, then the cost estimates decline over time, converging to the true value in the limit. In this case, the speed of convergence of cost estimates is directly related to the speed of convergence of beliefs. However, if beliefs fail to converge to the uniform distribution, then no clear dynamic pattern emerges regarding estimates of information costs. Both predictions of the dynamic model regarding estimates across sessions and within a session follow directly from our Monte-Carlo exercise.

To test the first dynamic prediction, that acquaintance with the experimental setup lowers the estimates of information costs, we estimate the parameters of the cost functions separately for the two sessions. For each individual, we find joint estimates of the parameters allowing either $\bar{\theta}$ or $\bar{\kappa}$ to differ between the two sessions, while treating the rest of the parameters as constants throughout both sessions. Columns 5 and 6 in each panel of Table 4 report the likelihood ratio test statistics for the hypotheses that the parameters are equal across two sessions: $H_0^{\theta} : \bar{\theta}_1 = \bar{\theta}_2$, and $H_0^{\kappa} : \bar{\kappa}_1 = \bar{\kappa}_2$ respectively. All reported test statistics have a chi-squared distribution with 1 degree of freedom.

Table 4 identifies cases when the null is rejected and the costs are higher (capacity is lower) in the first session with the § sign. Similarly, cases when the null is rejected and the costs are lower (capacity is higher) in the first session are marked with the $\clubsuit$ sign. We find statistically significant evidence that for 21 participant out of 40, the cost estimates fall in the second session from the first. For another 15 participants out of 40, we cannot reject the null that costs have not changed. However, for the majority of these participants, the estimates of costs also fall. For just four participants out of 40 we find that the estimates of costs increase in the second session from the first.

We conclude from this result that the vast majority of participants of our experiment acquire knowledge about the experimental environment in the first session. This knowledge allows them to be more precise in the second session. Only every tenth participant violates this dynamic prediction of our theory.

Running rolling-window estimates of costs of information provides a more nuanced way of studying the dynamic behavior of participants. We used rolling windows that include answers to 10

consecutive repetitions of each question. For each window, we estimate the costs of information $\hat{\hat{\theta}}$, while keeping the estimates of all other parameters fixed at values in Tables 4 and 5. We apply the rolling window estimation procedure to each session for each participant. For illustrative purposes, we average the estimates across: 1) all 40 participants of the experiment; 2) the 21 participants that we identified as "learners"; and 3) the 15 "consistent" participants for which we could not detect a significant change in cost estimates. The averaged dynamic estimates of costs of information are shown in Figure 4.

We find that rolling-window estimates of costs of information fall over the course of both sections for all three groups of participants. The main difference between the two sub-groups seems to be the speed of convergence. These estimates suggest that participants indeed acquire information slowly about the experimental environment, as predicted by our dynamic model.

Figure 5 shows for the same groups of participants the average switching rates, i.e. the frequencies with which participants change answers to the same questions over the course of the two sessions. The switching rates behave very similarly to estimates of costs of information, demonstrating that as participants learn about the environment over the course of the experiment, their choices become sharper and more consistent.

Our model provides a unified framework for rationalizing these empirical regularities without relying on ad hoc assumptions. In particular, we have established three empirical facts. The first fact is that participants are much more consistent in the second session, which may be several days after the first one. Our model shows that this is consistent with participants learning something important about the experiment in the first session. The second fact is that participants remain highly inconsistent after encountering each pair of gambles more than 50 times. Our model shows that this observation can be explained by cognitive limitations. The third fact is that participants remain predictably more consistent on more "valuable" questions. Our model shows that this can be accounted for by the participants' choice to vary information capacity in response to incentives. These three facts are predicted and jointly accounted for by our model.

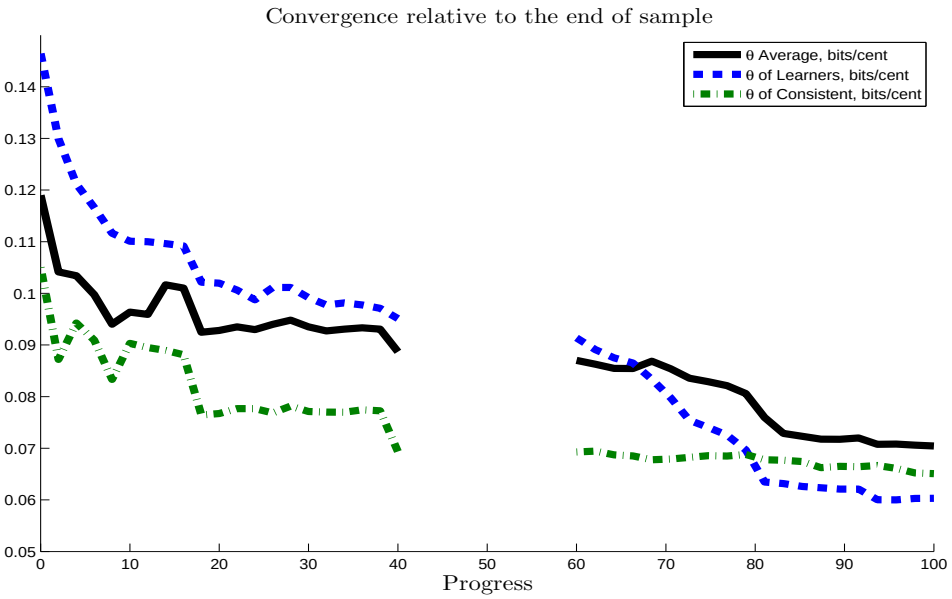

Figure 4: Information Cost Convergence

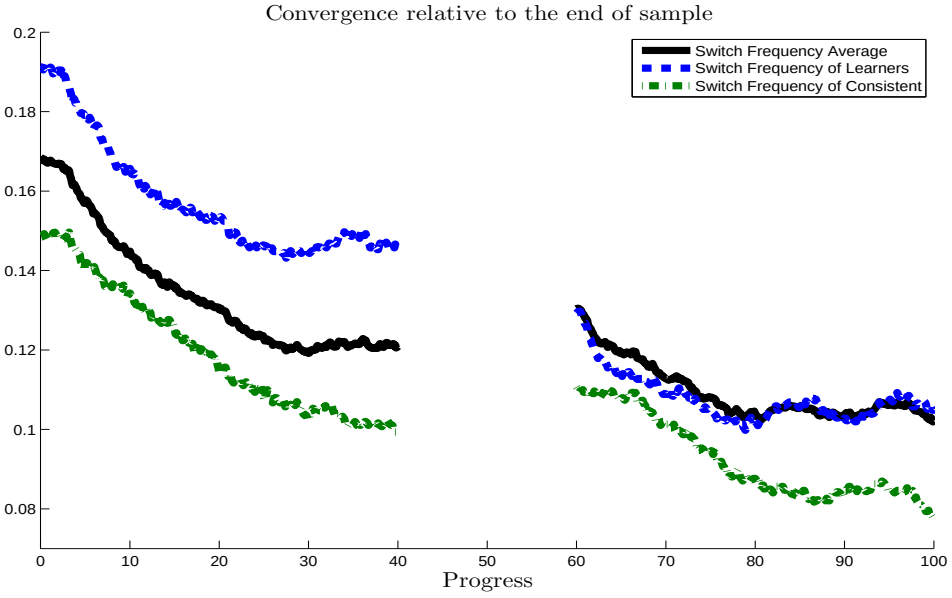

Figure 5: Switching Rate Convergence

6 Conclusion

In this paper we propose a rational inattention model as a microfoundation for stochastic choice. Our model establishes a mapping between information processing costs and probabilistic choice distributions and generates various probabilistic choice theories as special cases. The main contribution of the paper is to estimate the shape and size of information processing costs using data from a unique behavioral experiment.

Our estimates represent the first attempt to measure information costs in the laboratory and can serve as a benchmark calibration in bounded rationality models. Simultaneously, the estimates inform the behavioral choice literature on the appropriate selection of error model.

These estimates allow us to discriminate between physical bounds on the decision maker's cognitive ability and deliberate choice to disregard information. We find that experimental data reject the hypothesis that errors are driven by physical bounds.

Building on an extension of the model to a repeated setting, we draw two lessons. First, we show that individuals become less prone to error as they learn from experience. The speed of convergence is slow and it is worth having a large number of repetitions to obtain unbiased estimates. The second lesson is that aggregating across individuals is a source of large bias both for the probabilistic choice model and for the model of risky choice.

We hope that the approach proposed in this paper can be a unifying framework for modeling bounded rationality in macroeconomics and in behavioral social choice. Understanding whether stochastic choice is rooted in physical limits or is a result of conscious errors also has important policy implications. If cognitive mistakes are a result of physical limitations, then no incentive scheme could reduce them. However, if errors are a result of deliberate disregard of information, then one could design appropriate ways to motivate agents and guide them toward right decisions.

References

[1] Agranov, M., and P. Ortoleva (2013). Stochastic Choice and Hedging. Mimeo.

[2] Birnbaum, M. H. (1974). The nonadditivity of personality impressions. Journal of Experimental Psychology, 102, 543–561.

[3] Birnbaum, M. H. (2008). New paradoxes of risky decision making. Psychological Review, Vol 115(2), Apr 2008, 463-501.

[4] Birnbaum, M. H., and Chavez, A. (1997). Tests of theories of decision making: Violations of branch independence and distribution independence. Organizational Behavior and Human Decision Processes, 71, 161–194.

[5] Block, H. D., and Marschak, J. (1960). Random orderings and stochastic theories of responses. In I. Olkin, S. Ghuyre, W. Hoeffding, W. Madow, & H. Mann (Eds.) Contributions to Probability and Statistics. Stanford University Press, Pp.97-132.

[6] Busemeyer, J. and Townsend, J.T. (1993) Decision Field Theory: A dynamic-cognitive approach to decision making under uncertainty. Psychological Review, vol. 100 No.3, 432-459.

[7] Caplin, A. and M. Dean (2013a). Rational Inattention and State Dependent Stochastic Choice. Mimeo.

[8] Caplin, A. and M. Dean (2013b). Rational Inattention, Entropy, and Choice: The Posterior-Based Approach. Mimeo.

[9] Cokely, E.T., Schooler, L.J., and Gigerenzer, G. (2010). Information use for decision making. In M.N. Maack & M.J. Bates (Eds.), Encyclopedia of Library and Information Sciences, 3rd Edition (pp. 2727-2734). Taylor & Francis Group.

[10] Falmagne, J.-Cl. (1978) A Representation Theorem for Finite Random Scale Systems, Journal of Mathematical Psychology, 1978, pp. 52—72.

[11] Fechner, G.T. (1860) Elemente der Psychophysik. Breitkopf & Härtel, Leipzig (reprinted in 1964 by Bonset, Amsterdam); English translation by HE Adler (1966): Elements of psychophysics. Holt, Rinehart & Winston, New York

[12] Fudenberg, D. and T. Strzalecki (2012). Recursive Stochastic Choice. Mimeo.

[13] Gabaix, X. and Laibson, D. (2005) Shrouded attributes, Consumers Myopia and information suppression in competitive markets, Quarterly Journal of Economics 121 (2): 505-540.

[14] Gabaix, X. (2012). A sparsity-based model of bounded rationality, applied to basic consumer and equilibrium theory. Mimeo

[15] Goldstein D.G. and Gigerenzer, G. (2002). Models of Ecologica Rationality: The Recognition Heuristic. Psychological Review, vol. 109 No.1, 75-90.

[16] Goldstein, W. M. and Einhorn, H. J. (1987), Expression Theory and the Preference Reversal Phenomena. Psychological Review 94, 236—254.

[17] Gul, F. and Pesendorfer, W. (2006) Random Expected Utility, Econometrica, vol.74, issue 1, pp.121-146.

[18] Harless, D. W., and Camerer, C. F. (1994). The Predictive Utility of Generalized Expected Utility Theories, Econometrica, Econometric Society, vol. 62(6), pages 1251-89, November.

[19] Hey, J.D. and Orme, C., (1994). "Investigating Generalizations of Expected Utility Theory Using Experimental Data," Econometrica, Econometric Society, vol. 62(6), pages 1291-1326, November.

[20] Hey, J. D. (2001). "Does Repetition Improve Consistency?," Experimental Economics, Springer, vol. 4(1), pages 5-54, June.

[21] Holt, C.A. and Laury, S.K. (2002). Risk Aversion and Incentive Effects, American Economic Review, American Economic Association, vol. 92(5), pages 1644-1655, December.

[22] Kahneman, D., and Tversky, A. (1979). Prospect Theory: An Analysis of Decision under Risk. Econometrica, 47(2), 263–292.

[23] Karmarkar U.S. (1979), Subjectively weighted utility and the Allais paradox. Organizational Behavior and Human Performance, 24, 67–72.

[24] Kumaraswamy, P. (1980). A generalized probability density function for double bounded random processes. Journal of Hydrology, 46(1–2):79–88, 1980.

[25] Lattimore, P.K., Baker, J.R. and Witte, A.D. (1992), The influence of probability on risky choice, Journal of Economic Behavior and Organization 17, 377–400.

[26] Luce, R.D. (1959). Individual Choice Behaviours: A Theoretical Analysis. New York: J. Wiley.

[27] Luce, R.D. (1994) Thurstone and sensory scaling: then and now. Psychological Review, 107, 271-277.

[28] Luce, R.D. (2010) Behavioral Assumptions for a Class of Utility Theories: A Program of Experiments. Journal of Risk and Uncertainty, 41, 19-27.

[29] Mackowiack, B. and Wiederholt, M. (2009) Optimal sticky prices under rational inattention. American Economic Review 99(3): 769-803.

[30] Mankiw, N. G. and Reis, R. (2002). Sticky Information Versus Sticky Prices: A proposal to replace the New Kensian Phillips Curve. Quarterly Journal of Economics 117: 1295-1328.

[31] Manzini, P. and Mariotti, M (2014). Stochastic Chice and Consideration Sets. Econometrica, 82 (3), 1153-1176.

[32] Marley, A.A. J. and Luce R.D. (2005) Independence properties vis-à-vis several utility representations. Theory and Decision, 58, 77-143.

[33] Masatlioglu, Y., Nakajima, D., and Ozbay, E. Y. (2012). Revealed Attention. The American Economic Review, 102 (5), 2183-2205.

[34] Matějka, F. and McKay, A. (2015) Rational Inattention to Discrete Choices: A New Foundation for the Multinomial Logit Model. The American Economic Review, 105 (1), 272-98.

[35] McFadden, D. (1978) Modelling the choice of residential location. In Karlqvist, A., Lundqvist, L., Snickars, F. and Weibull, J. (eds) Spatial Interaction Theory and Residential Location. North-Holland, Amsterdam.

[36] Mosteller, F. and Nogee, P. 1951. "An Experimental Measurement of Utility," Journal of Political Economy, 59, 371-404.

[37] Moscarini, G. (2004), "Limited Information Capacity as a Source of Inertia", Journal of Economic Dynamics and Control, 28, 2003–2035.

[38] Prelec, D. (1998), "The Probability Weighting Function," Econometrica 66, 497—527.

[39] Puterman, M.L. (1994). Markov Decision Processes: Discrete Stochastic Dynamic Programming. Wiley Series in Probability and Mathematical Statistics, John Wiley and Sons, Inc.

[40] Quiggin, J. (1982), A theory of anticipated utility', Journal of Economic Behavior and Organisation 3(4), 323-43.

[41] Regenwetter, M., Dana, J. and Davis-Stober, C. (2010). "Testing Transitivity of Preferences on Two-Alternative Forced Choice Data." Frontiers in Psychology, 1.

[42] Regenwetter, M., Dana, J., and Davis-Stober, C. (2011). "Transitivity of Preferences." Psychological Review, 118, 42-56.

[43] Reis, R. A. (2006a) Inattentive Consumers, Journal of Monetary Economics, 53 (8), 1761-1800.

[44] Reis, R. A. (2006b) Inattentive Producers, Review of Economic Studies, 73, 793-821.

[45] Rieskamp, J. (2008) The probabilistic nature of preferential choice. Journal of Experimental Psychology: Learning, Memory, and Cognition, Vol 34(6), Nov 2008, 1446-1465.

[46] Rubinstein, A. (1998) Modeling bounded rationality. MIT Press.

[47] Saari, D.G. (2005) The profile structure for Luce's choice axiom. Journal of Mathematical Psychology Vol 49 pp. 226–253.

[48] Shannon, C.E. (1948) A Mathematical Theory of Communication. Bell System Technical Journal, vol. 27, pp. 379–423 and 623–656, July and October, 1948.

[49] Sims, C.A. (2003). Implications of Rational Inattention. Journal of Monetary Economics 50(3), 665-690.

[50] Sims, C.A. (2006). Rational Inattention: Beyond the Linear-Quadratic Case. American Economic Review Papers and Proceedings 96(2), 158-163.

[51] Stokey, N. L. and R. E. Lucas Jr., with Edward C. Prescott (1989) *Recursive Methods in Economic Dynamics*, Cambridge, Harvard University Press.

[52] Swait, J. and A.A.J. Marley (2013) Probabilistic choice (models) as a result of balancing multiple goals. Journal of Mathematical Psychology Vol 57 pp. 1-14.

[53] Thomas, C. and Cover, J. (1991) Elements of Information Theory. John Wiley & Sons, Inc.

[54] Thurstone, L. L. (1927). "A Law of Comparative Judgement." Psychological Review, 34: pp. 273-286.

[55] Tutino, A. (2011) Rationally Inattentive Macroeconomic Wedges. The Journal of Economic Dynamics and Control, March 2011.

[56] Tutino, A. (2012) Rationally inattentive consumption choices. Review of economic dynamics, forthcoming.

[57] Tversky, A. (1969). "Intransitivity of Preferences." Psychological Review, 76(1): pp. 31-48.

[58] Tversky, A., Kahneman, D. (1992). Advances in Prospect Theory: Cumulative Representation of Uncertainty. Journal of Risk and Uncertainty, 5(4), 297–323.

[59] Von Neumann, J.,and Morgenstern, O. (1944). Theory of Games and Economic Behavior, Princeton, NJ: Princeton University Press.

[60] Yellott, J. I., Jr. (1977). The relationship of Luce's choice axiom, Thurstone's theory of comparative judgment, and the double exponential distribution. Journal of Mathematical Psychology, 15, 109-144.

[61] Wiederholt, M. (2010) "Rational Inattention." prepared for the The New Palgrave Dictionary of Economics.

[62] Wilcox, N. T. (2010) "A Comparison of Three Probabilistic Models of Binary Discrete Choice Under Risk", mimeo.

[63] Wilcox, N. T. (2011) "Stochastically More Risk Averse: A Contextual Theory of Stochastic Discrete Choice Under Risk." Journal of Econometrics, 162(1), May 2011, 89-104.

[64] Woodford, M. (2012). Inattentive valuation and reference-dependent choice. Mimeo, 2012.

[65] Wu, G. and Gonzalez, R. (1996), Curvature of the probability weighting function, Management Science 42, 1676–1690.

(NOT FOR PUBLICATION)

A Appendix: Extensions

A.1 The Dynamic Model

Consider a game (experimental setup) that is composed of questions, denoted by q, and answers denoted by k.We begin by defining the space of questions as Ω_q which contains Q elements and the set of answers as Ω_k containing K elements. We assume that each question $q \in \Omega_q$ can be answered with each answer $k \in \Omega_k$. Answers and questions respectively are mutually exclusive. The state space contains all the combinations of questions and answers and it is denoted by $\Omega = \Omega_q \times \Omega_k$. Elements of the state space $\omega = (q, k) \in \Omega$, represent pairs of potential questions and answers to them. We shall use ω and (q, k) interchangeably.

We define the state variable $g(\omega_t)$ of the DM as a probability distribution over all pairs of questions and answers in period t. The variable $g(\omega_t)$ represents prior beliefs of the DM about the probabilities of each combination of a question and an answer occurring in the current period. These are the beliefs the DM is endowed prior to seeing the computer screen in period t. Thus, $g(\omega_t)$ is a function $g : \Omega \longrightarrow [0, 1]$ characterized by:

$$\sum_{\omega_t \in \Omega} g(\omega_t) = \sum_{q_t \in \Omega_q} \sum_{k_t \in \Omega_k} g(q_t, k_t) = 1, \qquad g(\omega_t) \geq 0, \forall t.$$

The stochastic process which governs the realizations of questions and possible answers to them is assumed to be first-order Markov, with a law of motion characterized by the transition matrix $P(\omega_{t+1}|\omega_t) : \Omega \times \Omega \longrightarrow \mathbb{R}$. Consistent with our experimental setting, we assume that answers and questions evolve independently from each other with a computer selecting the next question using a transition matrix $r(q_{t+1}|q_t)$ for the questions and $\rho(k_{t+1}|k_t)$ for the answers. Thus, the transition matrix for the system is given by $P(\omega_{t+1}|\omega_t) = r(q_{t+1}|q_t) \circledast \rho(k_{t+1}|k_t)$ where "$\circledast$" denotes Kroenecker product.

Let $\hat{\omega}$ represent the observation of the computer screen by the DM. We assume that the random variable $\hat{\omega}$ takes a value in Ω. Let the variable $p(\omega_t, \tilde{\omega}_t)$ be the joint distribution of $(\omega, \hat{\omega})$ which is chosen by the DM at time t. Let $h(\omega_t) = \sum_{\hat{\omega}_t} p(\omega_t, \hat{\omega}_t)$ be the marginal probability distribution

from which the DM draws her answers at time t. We assume that the DM chooses such a probability after seeing the computer screen.

The probability of selecting answer k given the last observation $\hat{\omega}$ equals:

$$s\left(k_t|\hat{\omega}_t\right) = \sum_{q_t \in \Omega_q} p\left(q_t, k_t|\hat{\omega}_t\right). \tag{18}$$

From the DM's perspective, the transition function is a function $R(\omega_{t+1}, \hat{\omega}_{t+1} \ |\omega_t, \hat{\omega}_t) : \Omega \times \Omega \times \Omega \times \Omega \longrightarrow \mathbb{R}$ which maps current values of $(\omega_t, \hat{\omega}_t)$ into their future values. The relationship between $R(\omega_{t+1}, \hat{\omega}_{t+1} \ |\omega_t, \hat{\omega}_t)$ and $P\left(\omega_{t+1}|\omega_t\right)$ is given by manipulating the joint distribution of $(\omega_{t+1}, \hat{\omega}_{t+1}|\omega_t, \hat{\omega}_t)$ denoted by $M\left(\omega_{t+1}, \hat{\omega}_{t+1}, \omega_t, \hat{\omega}_t\right)$. Let $N\left(\hat{\omega}_{t+1}, \hat{\omega}_{t+1}|\omega_t, \omega_t\right)$ be the distribution of current and future observations $(\hat{\omega}_{t+1}, \hat{\omega}_t)$ conditional on current and future values (ω_{t+1}, ω_t) and note that by Markovianity such a function boils down to $N\left(\hat{\omega}_{t+1}, \hat{\omega}_{t+1}|\omega_t, \omega_t\right) = f\left(\hat{\omega}_{t+1}|\omega_{t+1}\right) = \frac{p(\omega_{t+1}, \hat{\omega}_{t+1})}{\Sigma_{\hat{\omega}_{t+1}} p(\hat{\omega}_{t+1}, \omega_{t+1})}$. Recall that $h\left(\omega_t\right) = \sum_{\hat{\omega}_t} p\left(\omega_t, \hat{\omega}_t\right)$. Then the relationship between $R\left(:\right)$ and $P\left(:\right)$ is given by:

$$\begin{aligned} R\left(\omega_{t+1}, \hat{\omega}_{t+1}|\omega_t, \hat{\omega}_t\right) &= \frac{f\left(\hat{\omega}_{t+1}|\omega_{t+1}\right)\left(P\left(\omega_{t+1}|\omega_t\right) h\left(\omega_t\right)\right)}{p\left(\omega_t, \tilde{\omega}_t\right)} \\ &= \frac{f\left(\hat{\omega}_{t+1}|\omega_{t+1}\right)}{f\left(\hat{\omega}_t|\omega_t\right)} P\left(\omega_{t+1}|\omega_t\right) \end{aligned} \tag{19}$$

We can cast the DM's problem into the following Bellman equation:

$$\begin{aligned} \mathcal{W}\left(g\left(\omega_t\right)|\hat{\omega}_t\right) = \max_{p(\omega_t, \hat{\omega}_t)} & \sum_{\omega_t \in \Omega_\omega} V\left(\omega_t|\hat{\omega}_t\right) s\left(k_t|\hat{\omega}_t\right) - C\left(\kappa_t\right) \\ + \beta & \sum_{\omega_{t+1} \in \Omega} \mathcal{W}\left(g_{t+1}\left(\omega_{t+1}\right)|\hat{\omega}_{t+1}\right) R\left(\omega_{t+1}, \hat{\omega}_{t+1}|\omega_t, \hat{\omega}_t\right) s\left(k_t|\hat{\omega}_t\right) \end{aligned} \tag{20}$$

s.t.

$$\kappa_t = I\left(p\left(\omega_t, \hat{\omega}_t\right), g\left(\omega_t\right)\right) = \sum_{\omega_t \in \Omega} \sum_{\hat{\omega}_t \in \Omega} p\left(\omega_t, \hat{\omega}_t\right) \log_2 \frac{p\left(\omega_t, \hat{\omega}_t\right)}{g\left(\omega_t\right)} \tag{21}$$

$$\begin{aligned} g_{t+1}\left(\omega_{t+1}\right) &= \sum_{\hat{\omega}_{t+1} \in \Omega} \sum_{\omega_t \in \Omega} R\left(\omega_{t+1}, \hat{\omega}_{t+1}|\omega_t, \hat{\omega}_t\right) \odot p\left(\omega_t, \hat{\omega}_t\right) \\ &= \sum_{\omega_t \in \Omega} \left(P\left(\omega_{t+1}|\omega_t\right) h\left(\omega_t\right)\right) \odot \sum_{\hat{\omega}_{t+1} \in \Omega} f\left(\hat{\omega}_{t+1}|\omega_{t+1}\right) \end{aligned} \tag{22}$$

$$p\left(\omega_t, \hat{\omega}_t\right) \geq 0 \qquad \sum_{\omega_t} \sum_{\hat{\omega}_t} p\left(\omega_t, \hat{\omega}_t\right) = 1, \quad \forall t \tag{23}$$

The value function in (20) takes up as argument the distribution of the prior $g(\omega_t)$ in t. The variable $p(\omega_t, \hat{\omega}_t)$ is chosen to maximize the current expected value $V(\omega_t|\hat{\omega}_t)$ as well as the discounted continuation value $\mathcal{W}(g_{t+1}(\omega_{t+1})|\hat{\omega}_t)$ both conditional on having observed $\hat{\omega}_t$. The continuation value depends on the state one period ahead, $g_{t+1}(\omega_{t+1})$. We assume that the discount factor is bounded: $\beta \in [0, 1)$.

The cost of processing information is denoted by $C(\kappa_t)$ which is an increasing convex function of the information processing capacity, κ_t, whose functional form is described in (21). Note that the DM's information-processing capacity, κ_t, may change as t unfolds. The interpretation of capacity that varies with t is that people may choose to vary their information-processing needs as their experience progresses according to the environment they face. For instance, a choice that involves a large sum of money may call for bigger attention effort than a choice where modest amount of money is involved.

The law of motion of the state variable in equation (22) is derived using Bayesian conditioning by convoluting the transition function $R(\omega_{t+1}, \hat{\omega}_{t+1}|\omega_t, \hat{\omega}_t)$ with the choice made by the DM, $p(\omega_t, \hat{\omega}_t)$. The symbol "$\odot$" denotes such a convolution. Equations (23) describe the consistency requirement that the distribution chosen is a proper distribution.

The system (20)-(23) fully characterizes the dynamic problem of the DM. We now turn to establishing the properties of this dynamic problem and deriving testable predictions from it.

A.1.1 Properties of the Bellman program

The purpose of this subsection is twofold. First, it establishes existence and uniqueness of a solution to the system (20)-(23) and properties of the solution. Then, it derives the dynamic behavior of beliefs.

First, note that all the constraints are concave. In fact, all the constraints but (21) are linear in $p(\omega, \hat{\omega})$ and $g(\omega)$. For (21), the concavity of the problem with respect to $p(\omega, \hat{\omega})$ and $g(\omega)$ are guaranteed by Theorem (16.1.6) of Thomas and Cover (1991).

Next, we prove convexity of the value function and the fact that the value iteration is a contraction mapping. The following theorem provides the desired result. All proofs are in Appendix 2.

Theorem A.1

For the system (20)-(23), value recursion H and two given value functions V and U, it holds

that

$$||HV - HU|| \leq \beta ||V - U||,$$

with $0 \leq \beta < 1$ *and* $||.||$ *the supreme norm. That is, the value recursion* H *is a contraction mapping.*

Proof. See Appendix B. ■

The theorem can be explained as follows. The space of value functions defines a *vector space* which is closed under addition and scalar scaling and the contraction property ensures this space to be complete, in the sense that all Cauchy sequences have a limit in this space. The space of value functions together with the supreme norm form a *Banach space* and the *Banach fixed-point theorem* ensures (a) the existence of a single fixed point and (b) that the value recursion always converges to this fixed point (see Theorem 6 of Alvarez and Stockey, 1998 and Theorem 6.2.3 of Puterman, 1994).

The following theorem shows the convexity of the value function in the program *(20)-(23)*:

Theorem A.2 *If the utility is bounded and if* $p(\omega, \hat{\omega})$ satisfies (21)-(*23*) *then the recursion (20) is convex.*

Proof. See Appendix B. ■

The proof of Theorem 2 shows that the recursion for the program *(20)-(23)* is convex and can be represented as a set of $|\Omega|$ −dimensional hyperplanes. In the proof, the convex property is given by the fact that the n−step value function $\mathcal{W}_n(g)$ is defined as the supreme of a set of convex (linear) functions and thus, obtains a convex function as a result. The optimal value function $\mathcal{W}^*(g)$ is the limit for n that goes to infinity and, since all $\mathcal{W}_n(g)$ are convex functions so is $\mathcal{W}^*(g)$.

A.1.2 Long-run Behavior

We now establish the dynamic behavior of the Markovian processes governing the state variable by showing the convergence of the distribution $h(\omega_t) = \sum_{\hat{\omega}_t} p(\omega_t, \hat{\omega}_t)$ to $\overline{g}(\omega)$ where $\bar{g}(\omega)$ is defined as the limiting distribution of the prior $g(\omega_t)$. We shall proceed in three steps. First we show that the transition matrix $R(:) = R(\omega_{t+1}, \hat{\omega}_{t+1} | \omega_t, \hat{\omega}_t)$ converges to P, $P = P(\omega_{t+1} | \omega_t)$ and its unique invariant distribution $h(\omega) \to \bar{g}(\omega)$. Second, we show that the distance (21) decreases over time. These results help us with generating testable predictions for the dynamic behavior of the DM.

The first task is concerned with the long-run behavior of the transition matrixes. Let $p_0 = p(\omega_0, \hat{\omega}_0)$ be the distribution over Ω from which the initial values $(\omega_0, \hat{\omega}_0)$ are drawn and let $(\omega_t, \hat{\omega}_t)$

be a Markov chain (Ω, p_0, R) and define $\Pr(\omega_{t+l} = z, \hat{\omega}_{t+l} = x | \omega_t = i, \hat{\omega}_t = y) = \left(R^l\right)_{(z,x,i,y)}$ as an element of the matrix R to the power of l. The following result applies:

Lemma 1 *The transition function* $\frac{1}{T}\sum_{t=0}^{T-1} R^t$ *converges as* $T \to \infty$ *to the transition function* P.

Proof. See Appendix B. ■

This result shows that the variable R, which captures the dynamic beliefs of the DM about the transition process, converges to the true transition process, P. We use the lemma to prove the following:

Lemma 2 *There exists an invariant distribution. Moreover, any row of* P *is an invariant distribution and any invariant distribution is a convex combination of* P.

Proof. See Appendix B. ■

We are still left with the case where multiple invariant distributions may occur. The following lemma establishes the existence of a unique ergodic set for P.

Lemma 3 *There exists a unique ergodic set in* Ω, E, *for the transition function* P.

Proof. See Appendix B. ■

Applying theorem 11.2 of Stockey, Lucas and Prescott leads us to conclude that, given Lemma 3, R has a unique invariant distribution given by $\bar{g}(\omega)$. From the previous results, we can state the following result:

Theorem A.3 *The asymptotic distribution of* $h(\omega) = \sum_{\hat{\omega}} p(\omega, \hat{\omega})$ *converges to* $\bar{g}(\omega)$.

Proof. See Appendix B. ■

Next, we turn to the limiting dynamic behavior of (21). The following theorem shows that (21) decreases over time

Theorem A.4 *Let* $p(\omega_t, \hat{\omega}_t)$ *and* $g(\omega_t)$ *be two probability distributions of a finite state Markov chain at time* t. *Then* $d(p(\hat{\omega}_t, \omega_t) || g(\omega_t))$ *is monotonically decreasing. Moreover, the limit of this distance is positive:*

$$\lim_{t\to\infty} \left(d\left(p\left(\hat{\omega}_t, \omega_t\right) || g\left(\omega_t\right)\right)\right) > 0.$$

Proof. See Appendix B. ■

The implication of this theorem is that a decrease of information occurs as statistical equilibrium is approached. However, the distance between the two distributions does not vanish.

The results in this section make it possible to derive a static version of our problem as a special case of the dynamic program in (20)-(23). This version is expecially useful if one wishes to compare our model to the ones proposed in the literature. In fact, statistical models typically lack a dynamic dimension. As a result, these theories are silent on how people use their knowledge to sharpen future decisions as well as on the effect their choice of information has on current and future expected gains.

While the dynamic implications of the model constitute the thrust of this paper, before turning to the empirical results, we shall briefly discuss the static version of the model in order to ease the comparison between our rational inattention theory and the models in the literature.

A.2 A static version of the model

Studying the properties of a static version of our rational inattention model allows us to compare the solution of the decision maker's problem with existing static probabilistic choice models in the literature. Our static version of the problem is a special case of the dynamic program (20)-(23) when the process $\omega_t \in \Omega$ is zero-order Markov and when convergence to the stationary distribution $\bar{g}(\omega)$ has been achieved.

As time progresses, the continuation value in the dynamic problem will eventually stop depending on previous period's realizations and the prior carried over will be the same ever since. As a result, in the static version of the model the choice variable becomes $p(\omega) = p(q,k)$ and the model can be cast into:

$$\max_{p(q,k)} \sum_{k \in \Omega_k} V(q,k)\, s(k) - C(\kappa) \tag{24}$$

s.t.

$$\kappa = I\left(p(q,k), \bar{g}(q,k)\right) = \sum_{q \in \Omega_q} \sum_{k \in \Omega_k} p(q,k) \log_2 \frac{p(q,k)}{\bar{g}(q,k)} \tag{25}$$

$$p(q,k) \geq 0, \quad \sum_{q \in \Omega_q} \sum_{k \in \Omega_k} p(q,k) = 1 \tag{26}$$

where $s(k) = \sum_{q \in \Omega_q} p(q,k)$ in the objective function (24) represents the DM's chosen choice probability, observable in our experiment. As before, equation (25) is the mutual information between the distributions $\left(p(q,k), g(q,k)\right)$ and the constraint (26) limits the choice of the decision

maker to the space of proper distributions. With a slight abuse of notation, let $g(k) = \sum_{q \in \Omega_q} \bar{g}(q,k)$ denote the assymptotic prior distribution over the gambles and let $s(k)$ denote the choice of the decision-maker. Moreover recall that $I(p(q,k), g(q,k)) = I(s(k))$. This is now exactly the static problem we study in the main text.

B Appendix: Proofs

B.1 Proof of Theorem 1 in the Main Text

Proof. The problem can be conveniently rewritten into:

$$\max_{s(k)} \sum_{k=1}^{K} V(k) s(k) - C(\kappa) \tag{27}$$

s.t.

$$\kappa = I(s(k)) = \sum_{k=1}^{K} s(k) \log_2 \frac{s(k)}{g(k)} \tag{28}$$

$$s(k) \geq 0, \quad \sum_{k=1}^{K} s(q,k) = 1 \tag{29}$$

First, note that information, $I(s(k))$ is a strictly convex function of the probability distribution $\{s(k)\}$. This follows from the fact that this function is twice differentiable, and its Hessian is a diagonal matrix which contains only non-negative elements.

Second, since $C(\kappa)$ is increasing and convex in $I(s(k))$, convexity of information with respect to probabilities $\{s(k)\}$ guarantees that the composite function $C(I(s(k)))$ is also a convex function of the probability distribution $\{s(k)\}$. This in turn implies that the objective function of the decision-maker is concave in the choice variable $\{s(k)\}$.

Maximization of a concave function with respect to a linear constraint with a non-zero gradient and a set of non-negativity constraints leads to a unique solution satisfying the first-order condition:

$$V(k) - \frac{\theta}{\ln 2}\left(\ln \frac{s(k)}{g(k)} + 1\right) - \lambda = 0.$$

where $\theta = \frac{\partial C(I(s(k)))}{\partial I(s(k))}$ is the derivative of the cost function and λ is the Lagrange multiplier associated with the constraint that probabilities sum up to one. Note that this equation holds for all $k \in \Omega_k$. We can combine first-order conditions for any pair of k and $k' \in \Omega_k$to obtain:

$$\frac{s(k)}{s(k')} = \frac{g(k)}{g(k')} \exp\left(\frac{V(k) - V(k')}{\theta / \ln 2}\right).$$

By further rearranging and summing up over $s(k)$ we obtain the optimal probability (12). ■

B.2 Proof of Theorem 2 in the Main Text

Proof. Consider a pair of gambles which gives each agent a value differential $x \in R$. Denote parameter of the gamble pair $a = e^{-x \ln 2} > 0$ and inverse cost $\psi_i = \frac{1}{\theta_i} > 0$. According to predictions of rational inattention theory, the choice probability of agent i over gamble pair x is given by:

$$y_i = \frac{1}{1+e^{-\frac{x \ln 2}{\theta_i}}} = \frac{1}{1+a^{\psi_i}}.$$

The representative agent's choice probability is computed by averaging across agents:

$$y_{RA} = \frac{1}{N}\Sigma_{i=1}^{N} y_i = \frac{1}{N}\Sigma_{i=1}^{N} \frac{1}{1+a^{\psi_i}}.$$

Her inverse cost of information is then computed inverting the function:

$$y_{RA} = \frac{1}{1+a^{\psi_{RA}}}.$$

Consider the function $f(z) = \frac{1}{1+a^z}$ on $z > 0$. This function is strictly increasing and concave for $a < 1$, strictly decreasing and convex for $a > 1$, equals $\frac{1}{2}$ when $a = 1$. This last case happens only when when $x = 0$, when both sides are $\frac{1}{2}$ so θ_{RA} is undefined. Consider the case $a < 1$ first. By Jensen's inequality for any (unequal) values z_j in the domain and for any strictly positive weights a_j a concave function $f(z)$ satisfies:

$$f\left(\frac{\Sigma_j a_j z_j}{\Sigma_j a_j}\right) > \frac{\Sigma_j a_j f(z_j)}{\Sigma a_j}.$$

Hence,

$$\frac{1}{1+a^{\psi_{RA}}} = \frac{1}{N}\Sigma_{i=1}^{N} \frac{1}{1+a^{\psi_i}} < \frac{1}{1+a^{\frac{1}{N}\Sigma_{i=1}^{N}\psi_i}}.$$

Since the function $f(z)$ is strictly increasing in z it follows that

$$\frac{1}{\theta_{RA}} < \frac{1}{N}\Sigma_{i=1}^{N} \frac{1}{\theta_i}.$$

Similarly, when $a > 1$ the function $-f(z)$ is strictly increasing and concave. Hence,

$$\frac{1}{1+a^{\psi_{RA}}} = \frac{1}{N}\Sigma_{i=1}^{N} \frac{1}{1+a^{\psi_i}} > \frac{1}{1+a^{\frac{1}{N}\Sigma_{i=1}^{N}\psi_i}}.$$

Since the function $f(z)$ is now strictly decreasing in z it again follows that

$$\frac{1}{\theta_{RA}} < \frac{1}{N}\Sigma_{i=1}^{N} \frac{1}{\theta_i}.$$

Note that the bias disappears only if all agents have identical costs of information ($\theta_i = \theta_j$) or when the two options being compared are identical ($a \to 1$). ■

B.3 Proof of Theorem A.1

Proof. Let $\Gamma(\omega, \hat{\omega})$ be the constraint set containing (21)-(23). The H mapping displays:

$$H\mathcal{W}(g) = \max_{p(\omega,\hat{\omega})\in\Gamma(\omega,\hat{\omega})} H^{p}\mathcal{W}(g(\omega)),$$

with

$$\begin{aligned} H^{p}\mathcal{W}(g) = &\sum_{\omega_t\in\Omega_\omega} V(\omega_t|\hat{\omega}_t)\, s(k_t|\hat{\omega}_t) - C(\kappa_t) \\ &+ \beta \sum_{\omega_{t+1}\in\Omega} \mathcal{W}(g_{t+1}(\omega_{t+1})|\hat{\omega}_t)\, R(\omega_{t+1},\hat{\omega}_{t+1}|\omega_t,\hat{\omega}_t)\, s(k_t|\hat{\omega}_t). \end{aligned}$$

Assume that $||H\mathcal{W} - H\mathcal{U}||$ is the maximum at point $g \equiv g(\omega)$. Let $p_1 \equiv p_1(\omega,\hat{\omega})$ denote the optimal control for $H\mathcal{W}$ at g and $p_2 \equiv p_2(\omega,\hat{\omega})$ the optimal one for $H\mathcal{U}$.

$$\begin{aligned} H\mathcal{W}(g) &= H^{p_1}\mathcal{W}(g), \\ H\mathcal{U}(g) &= H^{p_2}\mathcal{U}(g). \end{aligned}$$

Then it holds

$$||H\mathcal{W}(g) - H\mathcal{U}(g)|| = H^{p_1}\mathcal{W}(g) - H^{p_2}\mathcal{U}(g),$$

assuming *WLOG* that $H\mathcal{W}(g) \geq H\mathcal{U}(g)$. Since p_2 maximizes $H\mathcal{U}$ at g, it follows that

$$H^{p_2}\mathcal{U}(g) \geq H^{p_1}\mathcal{U}(g) \tag{30}$$

When we apply the mapping H we have:

$$\begin{aligned} ||H\mathcal{W} - H\mathcal{U}|| &= \\ ||H\mathcal{W}(g) - H\mathcal{U}(g)|| &= \\ H^{p_1}\mathcal{W}(g) - H^{p_2}\mathcal{U}(g) &\leq \qquad (31) \\ H^{p_1}\mathcal{W}(g) - H^{p_1}\mathcal{U}(g) &= \\ \beta \sum_{\omega\in\Omega_\omega}\sum_{\hat{\omega}\in\Omega_{\hat{\omega}}} \mathcal{W}^{p_1}(g\prime|\hat{\omega})\, p_1 g - \beta \sum_{\omega\in\Omega_\omega}\sum_{\hat{\omega}\in\Omega_{\hat{\omega}}} [(\mathcal{U}^{p_1}(g\prime|\hat{\omega}))]\, p_1 g &\leq \\ \beta \sum_{\omega\in\Omega_\omega}\sum_{\hat{\omega}\in\Omega_{\hat{\omega}}} (||\mathcal{W} - \mathcal{U}||)\, p_1 g &= \\ \beta\, ||\mathcal{W} - \mathcal{U}|| & \end{aligned}$$

where the inequality in (31) comes from the fact that we are subtracting less given (30).

Recalling that $0 \leq \beta < 1$ completes the proof. ■

B.4 Proof of Theorem A.2

Proof. The proof is done via induction. We assume that all the operations are well-defined in their corresponding spaces. As in the previous proof, let $\Gamma(\omega, \hat{\omega})$ be the constraint set containing (21)-(23). For planning horizon $n = 0$, we have only to take into account the immediate expected rewards. Let $m(\hat{\omega}|\omega)$ be the conditional distribution of $\hat{\omega}$ given ω defined ad $m(\hat{\omega}|\omega) = \frac{p(\omega,\hat{\omega})}{g(\omega)}$. Then, we can define the contemporaneous reward as:

$$\mathcal{W}_0(g) = \max_{m(\hat{\omega}|\omega)\in\Gamma(\omega,\hat{\omega})} \left[\sum_{\omega\in\Omega} V(\omega)\, m(\hat{\omega}|\omega)\, g(\omega) - C(\kappa) \right] \tag{32}$$

and given that the cost function $C(\kappa)$ is increasing and convex, we can define the vectors

$$\left\{\alpha_0^i(\omega)\right\}_i \equiv \left(\sum_{\omega\in\Omega} V(\omega|\hat{\omega})\, m(\hat{\omega}|\omega) \right)_{m(\hat{\omega}|\omega)\in\Gamma(\omega,\hat{\omega})} \tag{33}$$

which leads to the desired

$$\mathcal{W}_0(g) = \max_{\left\{\alpha_0^i(\omega)\right\}_i} \left\langle \alpha_0^i, g \right\rangle \tag{34}$$

where $\langle ., . \rangle$ denotes the inner product $\left\langle \alpha_0^i, g \right\rangle \equiv \sum_{\omega\in\Omega} \alpha_0^i(\omega), g(\omega)$.For the general case, :

$$\mathcal{W}_n(g) = \max_{m(\hat{\omega}|\omega)\in\Gamma(\omega,\hat{\omega})} \left[\begin{array}{c} \sum_{\omega\in\Omega} V(\omega|\hat{\omega})\, m(\hat{\omega}|\omega)\, g(\omega) - C(\kappa) + \\ +\beta \sum_{\omega,\omega'\in\Omega} \sum_{\hat{\omega},\hat{\omega}'\in\Omega} \mathcal{W}(g'(\omega\prime)|\hat{\omega})\, R(\omega', \hat{\omega}'|\omega, \hat{\omega})\, m(\hat{\omega}|\omega)\, g(\omega) \end{array} \right] \tag{35}$$

by the induction hypothesis

$$\mathcal{W}_{n-1}(g(\cdot)|_{\hat{\omega}}) = \max_{\left\{\alpha_{n-1}^i\right\}_i} \left\langle \alpha_{n-1}^i, g'_{\hat{\omega}}(\cdot) \right\rangle \tag{36}$$

Plugging into the above equation and by definition of $\langle ., . \rangle$,

$$\mathcal{W}_{n-1}(g'_{\hat{\omega}}(\cdot)) = \max_{\left\{\alpha_{n-1}^i\right\}_i} \sum_{\omega,\omega'\in\Omega} \sum_{\hat{\omega},\hat{\omega}'\in\Omega} \alpha_{n-1}^i(g(\omega')) \left(R(\omega', \hat{\omega}'|\omega, \hat{\omega}) \frac{p(\omega, \hat{\omega})}{f(\hat{\omega})} \right) \tag{37}$$

where $f(\hat{\omega})$ is the marginal distribution of $\hat{\omega}$.

With the above:

$$\begin{aligned}\mathcal{W}_n(g) &= \max_{m\in\Gamma}\left[\begin{array}{c}\sum_{\omega\in\Omega_\omega} V(\omega|\hat{\omega})\, m(\hat{\omega}|\omega)\, g(\omega) - C(\kappa) + \\ +\beta \max_{\left\{\alpha^i_{n-1}\right\}_i} \sum_{\hat{\omega}\in\Omega} \frac{1}{f(\hat{\omega})} \sum_{\omega,\omega'\in\Omega} \sum_{\hat{\omega}'\in\Omega} \alpha^i_{n-1}(\omega') \left(R(\omega',\hat{\omega}'|\omega,\hat{\omega})\, m(\hat{\omega}|\omega)\right) g(\omega)\end{array}\right] \\ &= \max_{m\in\Gamma}\left[\langle V\cdot m,\ g(\omega)\rangle - C(\kappa) + \beta\sum_{\hat{\omega}\in\Omega}\frac{1}{f(\hat{\omega})} \max_{\left\{\alpha^i_{n-1}\right\}_i} \left\langle \sum_{\omega'\in\Omega} \alpha^i_{n-1}(w')\, R(\omega',\hat{\omega}'|\omega,\hat{\omega})\cdot m,\ g\right\rangle\right]\end{aligned} \tag{38}$$

At this point, it is possible to define

$$\alpha^j_{m,\hat{\omega}}(\omega) = \sum_{\omega'\in\Omega} \alpha^i_{n-1}(\omega')\, R(\omega',\hat{\omega}'|\omega,\hat{\omega})\cdot m. \tag{39}$$

Note that these hyperplanes are independent on the prior g for which the value function $\mathcal{V}_n$ is computed. Thus, the value function amounts to

$$\mathcal{W}_n(g) = \max_{m\in\Gamma}\left[\langle V\cdot m,\ g\rangle + \beta\sum_{\hat{\omega}\in\Omega}\frac{1}{f(\hat{\omega})} \max_{\left\{\alpha^j_{m,\hat{\omega}}\right\}_j} \left\langle \alpha^j_{m,\hat{\omega}}, g\right\rangle\right], \tag{40}$$

and define:

$$\alpha_{m,\hat{\omega},g} = \arg\max_{\left\{\alpha^j_{m,\hat{\omega}}\right\}_j} \left\langle \alpha^j_{m,\hat{\omega}}, g\right\rangle. \tag{41}$$

Note that $\alpha_{m,\hat{\omega},g}$ is a subset of $\alpha^j_{m,\hat{\omega}}$ and using this subset results into

$$\begin{aligned}\mathcal{W}_n(g) &= \max_{m\in\Gamma}\left[\langle V\cdot m,\ g\rangle + \beta\sum_{\hat{\omega}\in\Omega}\frac{1}{f(\hat{\omega})}\langle\alpha_{m,\hat{\omega},g}, g\rangle\right] \\ &= \max_{m\in\Gamma}\left\langle V\cdot m + \beta\sum_{\hat{\omega}\in\Omega}\frac{1}{f(\hat{\omega})}\alpha_{m,\hat{\omega},g},\ g\right\rangle.\end{aligned} \tag{42}$$

Now

$$\left\{\alpha^i_n\right\}_i = \bigcup_{\forall g}\left\{V\cdot m + \beta\sum_{\hat{\omega}\in\Omega}\frac{1}{f(\hat{\omega})}\alpha_{m,\hat{\omega},g}\right\}_{m\in\Gamma} \tag{43}$$

is a finite set of linear functions parametrized in the action set.

The final step entails the proof that the $\left\{\alpha^i_n\right\}_i$ sets are finite and discrete for all n. The finite cardinatility of these sets is an important step since it proves that we can represent $\mathcal{W}_n(g)$ with a finite set of supporting $\alpha-$functions. Again, we proceed via induction. For discrete actions, $\left\{\alpha^i_{\hat{\omega}}\right\}_i$ is discrete from its definition in (36). For the general case, we have to observe that for discrete

actions and observation $\hat{\omega}$ and assuming $M = \left|\left\{\alpha_{n-1}^{j}\right\}_{i}\right|$, the sets $\left\{\alpha_{m,\hat{\omega}}^{i}\right\}_{i}$ are finite and discrete: for a given action m and observation $\hat{\omega}$ we can generate at most M $\alpha_{m,\hat{\omega}}^{j}$ functions. Note that fixing the action, we can select one of the M $\alpha_{p,0}^{j}$ functions for each one of the observation and, thus, the $\left\{\alpha_{n}^{i}\right\}_{i}$ set is of finite cardinality. ■

B.5 Proof of Lemma 1

Proof. We need to evaluate $\lim_{t\to\infty}\left(\frac{f(\hat{\omega}_{t+1}|\omega_{t+1})(P(\omega_{t+1}|\omega_t)h(\omega_t))}{p(\omega_t,\tilde{\omega}_t)}\right)$. Given that the optimal distribution is ergodic, letting $t\to\infty$ leads to $p\left(\omega_t,\tilde{\omega}_t\right)=p\left(\omega_{t+1},\tilde{\omega}_{t+1}\right)=p\left(\omega,\tilde{\omega}\right)$, $h\left(\omega_t\right)=h\left(\omega_{t+1}\right)=h\left(\omega\right)$ and $P\left(\omega_{t+1}|\omega_t\right)=P$. Then:

$$\lim_{t\to\infty}\left(\frac{f\left(\hat{\omega}_{t+1}|\omega_{t+1}\right)h\left(\omega_t\right)}{p\left(\omega_t,\tilde{\omega}_t\right)}\left(P\left(\omega_{t+1}|\omega_t\right)\right)\right)=\left(\frac{f\left(\hat{\omega}|\omega\right)}{f\left(\hat{\omega}|\omega\right)}P\right)=P$$

■

B.6 Proof of Lemma 2

Proof. From the previous lemma, we know that $\frac{1}{T}\sum_{t=0}^{T-1}R^t\to P$ and that $PR=P$. Writing the equality of these matrixes as the equlity of the row vetors, we have $p_{-s}=[p_{\omega_1},..,p_{\omega_\Omega}]=p_\omega\times R$, so each row p_ω is an invariant distribution. Moreover, an invariant distribution $\bar{g}\left(\omega\right)$ satisfies

$$\forall n:\ \bar{g}\left(\omega\right)=\sum_{i\in\Omega}\left(\frac{1}{T}\sum_{t=0}^{T-1}R^t\right)\left(i,\omega\right)\bar{g}\left(i\right)\to\sum_{i\in\Omega}P\left(i,\omega\right)\bar{g}\left(\omega\right)$$

■

B.7 Proof of Lemma 3

Proof. Let us suppose that $\mathcal{E}$ and $\mathcal{E}^*$ are two ergodic sets for the transition function P. Proving that there exists a subset $a\in\mathcal{E}\cap\mathcal{E}^*$ such that $\bar{g}\left(a\right)>0$, then $\mathcal{E}$ and $\mathcal{E}^*$ are not distinct ergodic sets. That is, if $P\left(a,\mathcal{E}\right)=1$ and $P\left(a,\mathcal{E}^*\right)=1$, then $\mathcal{E}$ is equal to $\mathcal{E}^*$. Since there is a positive probability of asking question $\omega_1\in\Omega$ in the experiment, $g\left(\omega_1\right)>0$. If $\mathcal{E}$ is an ergodic set in Ω, then $P\left(\omega_1,\mathcal{E}\right)=1$ which implies $\omega_1\in\mathcal{E}$. If $\mathcal{E}^*$ were another ergodic set of Ω, we would get $\omega_1\in\mathcal{E}^*$ using the same argument. Thus, $\omega_1\in\mathcal{E}\cap\mathcal{E}^*$. ■

B.8 Proof of Theorem A.3

Proof. Combining the fact that the long-run transition function $R(:)$ converges to P (Lemma 1) and $R(:)$ has a unique invariant distribution (Lemma 3 and Theorem 11.2 of Lucas, Stockey and Prescott), it follows that $h(\omega) = \sum_{\hat{\omega}} p(\omega, \hat{\omega})$ eventually converges to the steady state distribution, $\bar{g}(\omega)$. ■

B.9 Proof of Theorem A.4

Proof. Let $\psi(\omega_t, \omega_{t+1})$ denote the joint distribution of ω_t and ω_{t+1} under the prior, i.e., $\psi(\omega_t, \omega_{t+1}) = g(\omega_t) P(\omega_{t+1}|\omega_t)$ and let $\upsilon(\hat{\omega}_{t+1}, \hat{\omega}_t, \omega_{t+1}, \omega_{t+1}) = p(\hat{\omega}_t, \omega_t)(R(\omega_{t+1}, \hat{\omega}_{t+1}|\omega_t, \hat{\omega}_t))$ be the corresponding joint probability under the distribution selected by the decision maker. The chain rule for relative entropy implies

$$\begin{aligned} & d(\upsilon(\hat{\omega}_{t+1}, \hat{\omega}_t, \omega_{t+1}, \omega_{t+1}) || \psi(\omega_t, \omega_{t+1})) \\ & \overset{(a)}{=} d(p(\hat{\omega}_t, \omega_t) || g(\omega_t)) + d(R(\omega_{t+1}, \hat{\omega}_{t+1}|\omega_t, \hat{\omega}_t) || P(\omega_{t+1}|\omega_t)) \\ & \overset{b}{=} d(p(\hat{\omega}_{t+1}, \omega_{t+1}) || g(\omega_{t+1})) + d(R(\omega_t, \hat{\omega}_t|\omega_{t+1}, \hat{\omega}_{t+1}) || P(\omega_t|\omega_{t+1})) \end{aligned} \tag{44}$$

where (a) comes from the chain rule for entropy and (b) comes from the time symmetry of the Markov process.

The conditional probability distributions are given by: $p(\hat{\omega}_{t+1}, \omega_{t+1}|\hat{\omega}_t, \omega_t) = (R(\omega_{t+1}, \hat{\omega}_{t+1}|\omega_t, \hat{\omega}_t)) = \frac{h(\omega_t) f(\hat{\omega}_{t+1}|\omega_{t+1})}{p(\omega_t, \hat{\omega}_t)} P(\omega_{t+1}|\omega_t) = \frac{f(\hat{\omega}_{t+1}|\omega_{t+1})}{f(\hat{\omega}_t|\omega_t)} P(\omega_{t+1}|\omega_t)$ and $g(\omega_{t+1}|\omega_t) = P(\omega_{t+1}|\omega_t)$. Using the non-negativity of $d(R(\omega_t, \hat{\omega}_t|\omega_{t+1}, \hat{\omega}_{t+1}) || P(\omega_t|\omega_{t+1}))$ from Corollary to Theorem 2.6.3 in Cover and Thomas, it has to be the case that:

$$d(p(\hat{\omega}_t, \omega_t) || g(\omega_t)) \geq d(p(\omega_{t+1}, \hat{\omega}_{t+1}) || g(\omega_{t+1})) \tag{45}$$

and consequently the distance between these two probability functions is decreasing in time.

Note that as time progresses $\lim_{t\to\infty} \frac{1}{t} \left(\frac{f(\hat{\omega}_{t+1}|\omega_{t+1})}{f(\hat{\omega}_t|\omega_t)} \right) = \lim_{t\to\infty} \frac{1}{t} \left(\frac{f(\hat{\omega}_t|\omega_t)}{f(\hat{\omega}|\omega)} \right) = 1$ and thus $\lim_{t\to\infty} \frac{1}{t} d\left(\left(\frac{f(\hat{\omega}_{t+1}|\omega_{t+1})}{f(\hat{\omega}_t|\omega_t)} P(\omega_{t+1}|\omega_t) || P(\omega_{t+1}|\omega_t) \right) \right) = 0$.

This implies that the quantity $d(R(\omega_{t+1}, \hat{\omega}_{t+1}|\omega_t, \hat{\omega}_t) || P(\omega_{t+1}|\omega_t))$ vanishes over time.

Let us focus now on the limiting distributions. If we let $\bar{g}(\omega_t)$ be any stationary distribution, the sequence $d(p(\hat{\omega}_t, \omega_t) | \bar{g}(\omega_t))$ is a monotonically non-increasing non-negative sequence and must therefore have a non-negative limit. Note that this limit is non-zero since we can further decompose

$p(\omega_t, \hat{\omega}_t) = h(\omega_t) m(\omega_t|\hat{\omega}_t)$ and by Theorem 3 we know that $\lim_{t\to\infty} \frac{1}{t}(h(\omega_t)) = \bar{g}_0(\omega_t) = g_0$ implying that $\lim_{t\to\infty}(d(h(\omega_t)||\bar{g}_0(\omega_t))) = \lim_{t\to\infty}(d(h(\omega_{t+1})||\bar{g}_0(\omega_{t+1}))) = 0$. Then, by the definition of relative entropy:

$$\begin{aligned} d(p(\hat{\omega}_t, \omega_t)||\bar{g}_0(\omega_t)) &= \sum_{\omega}\sum_{\hat{\omega}} p(\omega_t, \hat{\omega}_t) \log\left(\frac{p(\omega_t, \hat{\omega}_t)}{\bar{g}_0(\omega_t)}\right) \\ &= -\sum_{\omega}\sum_{\hat{\omega}} h(\omega_t) m(\omega_t|\hat{\omega}_t) \log\left(\frac{\bar{g}_0(\omega_t)}{h(\omega_t) m(\omega_t|\hat{\omega}_t)}\right) \\ &= -\sum_{\omega}\sum_{\hat{\omega}} h(\omega_t) m(\omega_t|\hat{\omega}_t) \left[\log\left(\frac{\bar{g}_0(\omega_t)}{h(\omega_t)}\right) + \log\left(\frac{1}{m(\omega_t|\hat{\omega}_t)}\right)\right] \end{aligned}$$

Let $\bar{m}(\omega|\hat{\omega})$ denote the stationary distribution of $m(\omega_t|\hat{\omega}_t)$. Then, taking the $\lim_{t\to\infty}$ for the above expression results into:

$$\begin{aligned} \lim_{t\to\infty}(d(p(\hat{\omega}_t, \omega_t)|g_0)) &\to -\left[\sum_{\omega}\sum_{\hat{\omega}} \bar{g}_0(\omega) \bar{m}(\omega|\hat{\omega}) \log\left(\frac{1}{\bar{m}(\omega|\hat{\omega})}\right)\right] \\ &\overset{(c)}{>} \log\left(\sum_{\omega} \bar{g}_0(\omega)\right) = 0 \end{aligned}$$

where (c) follows from Jensen's inequality and the inequality is strict since $\bar{m}(\omega|\hat{\omega}) = \frac{p(\omega,\hat{\omega})}{\bar{g}_0(\omega)} \neq \bar{g}_0(\omega)$. ■

C Appendix: Coding, knowledge, models of heuristics and coalescing

Our model based on rational inattention theory focuses on a decision maker's ability to act in an uncertain environment with limited processing capacity. Our model postulates that the decision maker, aware of her limited processing capacity, selects the information structure that conveys the highest utility. As a result, our model predicts that a rationally inattentive decision maker optimally chooses the amount of uncertainty that he is willing to tolerate by evaluating costs and benefits of processing information. This subsection makes four observations on the rational inattention model we proposed.

First, one important assumption of our model is that the only source of uncertainty faced by the decision maker is in the distribution of options ("states of the world"). We do not address any form of cognitive bias that might emerge from presenting the options as made of different numbers

of branches (e.g. three instead of two) or of different probability representations (e.g. pie charts as opposite to percentages). These cognitive biases are treated in Shannon's information theory as coding and decoding problems. As the example in the previous subsection makes clear, the way options are presented is one potential source of inefficient coding. That is, when evaluating the capacity of a channel -human brain, in our case-, a prominent branch of information theory is concerned about the optimal design and compression of inputs and outputs of the channel. Albeit we recognize that such a cognitive bias may be sizeable in experimental studies, we choose not to model this bias explicitly. In the main body of the paper we assume that the coding is always efficient.

Second, we want to highlight the difference between information and knowledge in our model. Some studies have interpreted information as equivalent to knowledge. For instance, Gigerenzer and Goldstein (2011) describe recognition and evaluation as the two processes that constitute information use for decision making. They describe recognition as the process of accessing memory, -i.e., previous knowledge-, and evaluation as the process of comparing choice options to objects in the knowledge base. The decision maker does not acquire new information or produce new knowledge when using this heuristic process. In our model, recognition corresponds to the prior of the participant about the gamble she faces. Before processing any information, this prior knowledge is measured by the uncertainty (or entropy) of the gambles. Then, evaluation corresponds to processing information about the gambles in order to reduce uncertainty. Thus, in our model, evaluation is the process of acquiring information and forming new knowledge.

Third, we want to emphasize the difference between rational inattention models and models of heuristics as advocated by, inter alia, Cokely, Schooler and Gigerenzer (2010), as well as models based on Decision Field Theory, as advocated by Busemeyer and Townsend (1993). The reason why we use rational inattention theory to describe people's behavior is due to the fact that its statistical foundations make the model general and universally applicable. So long as we can characterize the distribution of the state variables, we can measure ex-ante uncertainty. So long as we can postulate a decision theory, we can predict and measure the optimal reduction of uncertainty of the decision maker. We are concerned about the ability of the model to produce predictions consistent with observed behavior. We do not take a stand on whether this modeling strategy replicates the cognitive process that occurs in people's brains when they make decisions. This area of research goes beyond the scope of our paper.

Forth, we want to address the relationship between the information processing constraint and experimental design, with particualr emphasis on the phenomenon of coalescing. In Section 3, we pointed out that the technological constraint is independent of the objective probabilities p_{jk} and the J possible outcomes since by assumption the decision maker cannot influence the experimental set-up of the gambles proposed: she can only choose which gamble to pick. A word of caution is in order here. While experiment participants take the format of the game as given, the experiment designer needs to be mindful of the way the gambles are set-up.

Experimental evidence[15] suggests that varying the number of possible outcomes per gamble influences the decision-maker's choice. To make the discussion concrete, we illustrate the point with the following example:

Example 1 *[Birnbaum (2008)]Consider gamble A presented as follows*

$$\begin{array}{lll} A: & X_1 & \textbf{.1 probability to win \$100} \\ & X_2 & \textbf{.1 probability to win \$100} \\ & Y_2 & \textbf{.8 probability to win \$10} \end{array}$$

and define $p(X_1) = p_{x_1} = 0.1$, $p(X_2) = p_{x_2} = 0.1$ *and* $p(Y_2) = p_2 = .8$. *Now consider gamble A' where* (X_1, X_2) *have been combined as follows*

$$\begin{array}{lll} A': & Y_1 & \textbf{.2 probability to win \$100} \\ & Y_2 & \textbf{.8 probability to win \$10} \end{array}$$

Gamble A' is defined as the coalesced form of gamble A. Birbaum (2008) finds that people choose differently if presented with gamble A compared to their choice if presented with gamble A'. Rational inattention based Shannon's information theory suggests that the transformation of gamble A into gamble $A'$is not entropy-neutral, i.e., the uncertainty intrisic to gamble A is different from that of gamble A' since the event space in gamble A' is coarser than that in gamble A. Thus, the decision-maker's choice when presented with gamble A and gamble A' would encompass the difference in costs per bit involved in processing information about gamble A and gamble A'. The following lemma formalize the statement.

[15]See, e.g., Birnbaum (2003).

Lemma 4 *Given a partition $\alpha = [X_1, X_2, Y_2]$ we form the partition $\beta = [Y_1, Y_2]$ obtained by merging (X_1, X_2) into Y_1 where $p(X_1) = p_{x_1}$ and $p(X_2) = p_{x_2}$ and $p_i = P(Y_i)$. Then*

$$H(\beta) \leq H(\alpha) \tag{46}$$

Proof. The function $\varphi(p) = -p \log p$ is convex. Therefore for $\lambda > 0$ and $p_1 - \lambda < p_1 < p_2 < p_2 + \lambda$ we have that

$$\varphi(p_1 + p_2) < \varphi(p_1 - \lambda) + \varphi(p_2 + \lambda) < \varphi(p_1) + \varphi(p_2)$$

Then,

$$H(\alpha) - \varphi(p_{x_1}) - \varphi(p_{x_2}) = H(\beta) - \varphi(p_{x_1} + p_{x_2}) \tag{47}$$

because each side equals the contribution to $H(\alpha)$ and $H(\beta)$ respectively due to the common elements of α and β. Hence, (46) follows from (47). ■

Transforming the event space α into β implies moving probability mass from a state with low probability to a state with high probability. Whenever this move occurs, the system becomes less uniform and thus entropy decreases. This is the case for the example offered by Birnbaum (2008).

Example 2 *[Birnbaum (2008) con't.]The entropy of the gamble A is larger than the entropy of gamble A':*

$$H(A) = 0.92 > 0.72 = H(A').$$

Thus, the first gamble has more uncertainty than the second gamble and thus requires higher capacity to be processed.