Inattentive Valuation and Reference-Dependent Choice[*]

Michael Woodford
Columbia University

January 1, 2012

Abstract

In rational choice theory, individuals are assumed always to choose the option that will provide them maximum utility. But actual choices must be based on subjective perceptions of the attributes of the available options, and the accuracy of these perceptions will always be limited by the information-processing capacity of one's nervous system. I propose a theory of valuation errors under the hypothesis that perceptions are as accurate as possible on average, given the statistical properties of the environment to which they are adapted, subject to a limit on processing capacity. The theory is similar to the "rational inattention" hypothesis of Sims (1998, 2003, 2011), but modified for closer conformity with psychophysical and neurobiological evidence regarding visual perception. It can explain a variety of aspects of observed choice behavior, including the intrinsic stochasticity of choice; focusing effects; decoy effects in consumer choice; reference-dependent valuations; and the co-existence of apparent risk-aversion with respect to gains with apparent risk-seeking with respect to losses. The theory provides optimizing foundations for some aspects of the prospect theory of Kahneman and Tversky (1979).

PRELIMINARY

[*] I would like to thank Tom Cunningham, David Laibson, Andrei Shleifer, and the participants in the Columbia University MBBI neuroscience and economics discussion group for helpful comments; Dmitriy Sergeyev for research assistance; and the Institute for New Economic Thinking and the Taussig Visiting Professorship, Harvard University, for supporting this research.

Experiments by psychologists (and experimental economists) have documented a wide range of anomalies that are difficult to reconcile with the model of rational choice that provides the foundation for conventional economic theory. This raises an important challenge for economic theory. Can standard theory be generalized in such a way as to account for the anomalies, or must one start afresh from entirely different foundations?

In order for a theory consistent with experimental evidence to count as a generalization of standard economic theory, it would need to have at least two properties. First, it would still have to be a theory which explains observed behavior as optimal, given people's goals and the constraints on their behavior — though it might specify goals and constraints that differ from the standard ones. And second, it ought to nest standard theory as a limiting case of the more general theory.

Here I sketch the outlines of one such theory, that I believe holds promise as an explanation for several (though certainly not all) well-established experimental anomalies. These include stochastic choice, so that a given subject will not necessarily make the same choice on different occasions, even when presented with the same choice set, and so may exhibit apparently inconsistent preferences; focusing effects, in which some attributes of the choices available to a decisionmaker are given disproportionate weight (relative to the person's true preferences), while others (that do affect true utility) may be neglected altogether; choice-set effects, in which the likelihood of choosing one of two options may be affected by the other options that are available, even when the other options are not chosen; reference-dependence, in which choice among options depends not merely upon the final situation that the decisionmaker should expect to reach as a result of each of the possible choices, but upon how those final situations compare to a reference point established by a prior situation or expectations; and the co-existence of risk-aversion with respect to gains with risk-seeking with respect to losses, as predicted by the prospect theory of Kahneman and Tversky (1979).

There are three touchstones for the approach that I propose to take to the explanation of these phenomena. The first is the observation by McFadden (1999) that many of the best-established behavioral anomalies relate to — or can at least potentially be explained by — errors in *perception,* under which heading he includes errors in the retrieval of memories of past experiences. Because of the pervasiveness of the evidence for perceptual errors, McFadden argues that economic theory should be extended to allow for them. But he suggests that "if the cognitive anomalies that do appear in economic behavior arise mostly from perception errors, then much of the conventional apparatus of economic analysis survives, albeit in a form in which history and experience are far more important than is traditionally allowed" (p. 99).

Here I seek to follow this lead, by examining the implications of a theory in which economic choices are optimal, subject to the constraint that they must be based on subjective perceptions of the available choices. I further seek to depart from standard theory as minimally as possible, while accounting for observed behavior, by postulating that the perceptions of decisionmakers are themselves optimal, subject to a constraint on the decisionmaker's information-processing capacity. Standard rational choice theory is then nested as a special case of the more general theory proposed here, the one in which available information-processing capacity is sufficient to allow for accurate perceptions of the relevant features of one's situation.

A second touchstone is the argument of Kahneman and Tversky (1979) that key postulates of prospect theory are psychologically realistic, on the ground that they are "compatible with basic principles of perception and judgment" in other domains, notably perceptions of "attributes such as brightness, loudness, or temperature" (pp. 277-278). Here I pursue this analogy further, by proposing an account of the relevant constraints on information-processing that can also explain at least some salient aspects of the processing of sensory information in humans and other organisms. This has the advantage of allowing the theory to be tested against a much larger body of data, as perception has been studied much more thoroughly (and in quantitatively rigorous ways), both by experimental psychologists and by neuroscientists, in sensory domains such as vision.

More specifically, the theory proposed here seeks to develop an idea stressed by Glimcher (2011) in his discussion of how a neurologically grounded economics would differ from current theory: that judgements of value are necessarily reference-dependent, because "neurobiological constraints ... make it clear that the hardware requirements for a reference point-free model ... cannot in principle be met" (p. 274). I do not here consider constraints that may result from specific structures of the nervous system, but I do pursue the idea that reference-dependence is not simply an arbitrary fact, but may be necessary, or at least an efficient solution, given constraints on what it is possible for brains to do, given fundamental limitations that result from their being finite systems.

The third touchstone is the theory of "rational inattention" developed by Sims

(1998, 2003, 2011). Sims proposes that the relevant constraint on the precision of economic decisionmakers' awareness of their circumstances can be formulated using the quantitative measure of information transmission proposed by Shannon (1948), and extensively used by communications engineers. An advantage of information theory for this purpose is the fact that it allows a precise quantitative limit on the accuracy of perceptions to be defined, in a way that does not require some single, highly specific assumption about what might be perceived and what types of errors might be made in order for the theory to be applied. This abstract character of the theory means that it is at least potentially relevant across many different domains.[1] Hence if any general theory of perceptual limitations is to be possible — as opposed to a large number of separate studies of heuristics and biases in individual, fairly circumscribed domains — information theory provides a natural language in which to seek to express it. Here I do not adopt the precise quantitative formulation of the relevant constraint on information processing proposed by Sims; instead, I propose a modification of rational inattention theory that I believe conforms better to findings from empirical studies of perception. But the theory proposed here remains a close cousin of the one proposed by Sims.

The paper proceeds as follows. In section 1, I review some of the empirical evidence regarding visual perception that motivates the particular quantitative limit on the accuracy of perceptions that I use in what follows. Section 2 then derives the implications for perceptual errors in the evaluation of economic choices that follow from the hypothesis of an optimal information structure to the particular kind of constraint that is motivated in the previous section. Section 3 discusses several ways in which this theory can provide interpretations of apparently anomalous aspects of choice behavior in economic contexts, that have already received considerable attention in the literature on behavioral economics, and compares the present theory to other proposals that seek to explain some of the same phenomena. Section 4 concludes.

[1] Indeed, a number of psychologists and neuroscientists have already sought to characterize limits to human and animal perception using concepts from information theory. See, for example, Attneave (1954) and Miller (1956) from the psychology literature, or Barlow (1961), Laughlin (1981), Rieke *et al.* (1997), or Dayan and Abbott (2001), chap. 4, for applications in the neurosciences.

1 What Do Perceptual Systems Economize?

I shall begin by discussing the form of constraint on the degree of precision of people's awareness of their environment that is suggested by available evidence from experimental psychology and neurobiology. I wish to consider a general class of hypotheses about the nature of perceptual limitations, according to which the perceptual mechanisms that have developed are optimally adapted to the organism's circumstances, subject to certain limits on the degree of precision of information of *any* type that it would be feasible for the organism to obtain. And I am interested in hypotheses about the constraints on information-processing capacity that can be formulated as generally as possible, so that the nature of the constraint need not be discovered independently for each particular context in which the theory is to apply.

If high-level principles exist that determine the structure of perception across a wide range of contexts, then we need not look for them simply by considering evidence regarding perceptions in the context of economic decisionmaking. In fact, the nature of perception, and the cognitive and neurobiological mechanisms involved in it, has been studied much more extensively in the case of sensory perception, and of visual and auditory perception particularly. I accordingly start by reviewing some of the findings from the literatures in experimental psychology and neuroscience about relations between the objective properties of sensory stimuli and the subjective perception or neural representation of those stimuli, in the hope of discovering principles that may also be relevant to perception in economic choice situations.

I shall review this literature with a specific and fairly idiosyncratic goal, which is to consider the degree to which the experimental evidence provides support for either of two important general hypotheses about perceptual limitations that have been proposed by economic theorists. These are the model of partial information as an optimally chosen partition of the states of the world, as proposed in Gul *et al.* (2011), and the theory of "rational inattention" proposed by Sims (1998, 2003, 2011).

1.1 The Stochasticity of Perception

Economic theorists often model partial information of decisionmakers about the circumstances under which they must choose by a partition of the possible states of the world; it is assumed that a decisionmaker (DM) is correctly informed about which element of the partition contains the current state of the world, but that the DM has no ability to discriminate among states of the world that belong to the same element of the partition. This is not the only way that one might model partial awareness, but it has been a popular one; Lipman (1995) argues that limited information must be modeled this way in the case of "an agent who is fully aware of how he is processing his information" (p. 43).

In an approach of this kind, more precise information about the current state corresponds to a finer partition. One might then consider partial information to nonetheless represent a constrained-optimal information structure, if it is optimal (from the point of view of expected payoff that it allows the DM to obtain) subject to an upper bound on the *number of states* that can be distinguished (*i.e.*, the number of elements that there can be in the partition of states of the world), or to an information-processing cost that is an increasing function of the number of states. For example, Neyman (1985) and Rubinstein (1986) consider constrained-optimal play of repeated games, when the players' strategies are constrained not to require an ability to distinguish among too many different possible past histories of play; Gul *et al.* (2011) propose a model of general competitive equilibrium in which traders' strategies are optimal subject to a bound on the number of different states of the world that may be distinguished. This way of modeling the constraint on DMs' awareness of their circumstances has the advantage of being applicable under completely general assumptions about the nature of the uncertainty. The study of optimal information structures in this sense also corresponds to a familiar problem in the computer science literature, namely the analysis of "optimal quantization" in coding theory (Sayood, 2005).

However, it does not seem likely that human perceptual limitations can be understood as optimal under any constraint of this type. Any example of what Lipman (1995) calls "partitional" approaches to modeling information limitations implies that the DM's subjective representation of the state of the world is a *deterministic* function of the true state: the DM is necessarily aware of the unique element of the information partition to which the true state of the world belongs. And different states of the world can either be *perfectly* discriminated from one another (because they belong to separate elements of the partition, and the DM will necessarily be aware of one element or the other), or cannot be distinguished from one another *at all* (because they belong to the same element of the partition, so that the DM's awareness will always be identical in the two cases): there are no *degrees* of discriminability.

Yet one of the most elementary findings in the area of psychophysics — the study by experimental psychologists of the relation between subjective perceptions and the objective physical characteristics of sensory stimuli — is that subjects respond *randomly* when asked to distinguish between two relatively similar stimuli. Rather than mapping the boundaries of disjoint sets of stimuli that are indistinguishable from one another (but perfectly distinguishable from all stimuli in any other equivalence class), psychophysicists plot the way in which the *probability* that a subject recognizes one stimulus as brighter (or higher-pitched, or louder, or heavier...) than another varies as the physical characteristics of the stimuli are varied; the data are generally consistent with the view that the relationship (called a "psychometric function") varies continuously between the values of zero and one, that are approached only in the case of stimuli that are sufficiently different.[2] Thus, for example, Thurstone (1959) reformulates Weber's Law as: "The stimulus increase which is correctly discriminated in any specified proportion of attempts (except 0 and 100 percent) is a constant fraction of the stimulus magnitude." How exactly and over what range of stimulus intensities this law actually holds has been the subject of a considerable subsequent literature; but there has been no challenge to the idea that any lawful relationships to be found between stimulus intensities and discriminability must be *stochastic* relations of this kind.

Under the standard paradigm for interpretation of such measurements, known as "signal detection theory" (Green and Swets, 1966), the stochasticity of subjects' responses is attributed to the existence of a *probability distribution* of subjective perceptions associated with each objectively defined stimulus.[3] The probability of error in identifying which stimulus has been observed is then determined by the degree to which the distributions of possible subjective perceptions overlap;[4] stimuli that are objectively more similar are mistaken for one another more often, because

[2] See, for example, Gabbiani and Cox (2010), chap. 25; Glimcher (2011), chap. 4; Green and Swets (1966); or Kandel, Schwartz, and Jessel (2010), Box 21-1.

[3] This interpretation dates back at least to Thurstone (1927), who calls the random subjective representations "discriminal processes," and postulates that they are Gaussian random variables.

[4] Of course, even given a stochastic relationship between the objective stimulus and its subjective representation, there remains the question of how the subject's response is determined by the subjective representation. In "ideal observer theory," the response is the one implied to be optimal under statistical decision theory: the response function maximizes the subject's expected reward, given some prior probability distribution over the set of stimuli that are expected to be encountered.

the probabilities of occurrence of the various possible subjective perceptions are quite similar (though not identical) in this case. Interestingly, the notion that the subjective representation is a random function of the objective characteristics is no longer merely a conjecture; studies such as that of Britten *et al.* (1992) — who record the electrical activity of a neuron in the relevant region of the cortex of a monkey trained to signal perceptual discriminations, while the stimulus is presented — show that random variation in the neural coding of particular stimuli can indeed explain the observed frequency of errors in perceptual discriminations.

In order to explain the actual partial ability of human (or animal) subjects to discriminate between alternative situations, then, one needs to posit a stochastic relationship between the objective state and the subjective representation of the state. A satisfactory formalization of a constraint on the degree of precision of awareness of the environment that is possible — or of the cost of more precise awareness — must accordingly be defined not simply for partitions, but for arbitrary information structures that specify a set of possible subjective representations R and a conditional probability distribution $p(r|x)$ for each true state of the world x. It should furthermore be such that it is more costly for an information structure to discriminate more accurately between different states, by making the conditional distributions $p(\cdot|x)$ more different for different states x. But in order to decide which type of cost function is more realistic, it is useful to consider further experimental evidence regarding perceptual discriminations.

1.2 Experimental Evidence on the Allocation of Attention

While the studies cited above make it fairly clear that subjective perceptions are stochastically related to the objective characteristics of the environment, it may not be obvious that there is any scope for *variation* in this relationship, so as to make it better adapted to a particular task or situation. Perhaps the probability distribution of subjective perceptions associated with a particular objective state is simply a necessary consequence of the way the perceptual system is built, and will be the same in all settings. In that case, the nature of this relationship could be an object of study; but it might be necessary to make a separate study of the limits of perception of every distinct aspect of the world, with little expectation of finding any useful high-level generalizations.

There is, however, a certain amount of evidence indicating that people are able to vary the amount of attention that they pay to different aspects of their surroundings. Some aspects of this are commonplace; for example, we can pay more attention to a certain part of our surroundings by looking in that direction. The eye only receives light from a certain range of angles; moreover, the concentration of the light-sensitive cone cells in the retina is highest at a particular small area, the fovea, so that visual discrimination is sharpest for that part of the visual field that is projected onto the fovea. This implies opportunities for (and constraints upon) the allocation of attention that are very relevant to certain tasks (such as the question of how one should move about a classroom in order to best deter cheating on an exam), but that do not have obvious implications for more general classes of information processing problems. Of greater relevance for present purposes is evidence suggesting that even given the information reaching the different parts of the retina, people can vary the extent to which they attend to different parts of the visual field, through variation in what is done with this information in subsequent levels of processing.[5]

1.2.1 The Experiment of Shaw and Shaw (1977)

A visual perception experiment reported by Shaw and Shaw (1977) is of particular interest. In the experiment, a letter (either E, T, or V) would briefly appear on a screen, after which the subject had to report which letter had been presented. The letter would be chosen randomly (independently across trials, with equal probability of each of the three letters appearing on each trial), and would appear at one of eight possible locations on the screen, equally spaced around an imaginary circle; the location would also be chosen randomly (independently across trials, and independently of the letter chosen). The probability of appearance at the different locations was not necessarily uniform across locations; but the subjects were told the probability π_i of appearance at each location i in advance. The question studied was the degree to which the subjects' ability to successfully discriminate between the appearances of the different letters would differ depending on the location at which the letter appeared, and the extent to which this difference in the degree of attention paid to each location would vary with the likelihood of observing the letter at that location.

[5] See, for example, Kahneman (1973) and Sperling and Dosher (1986) for general discussions of this issue.

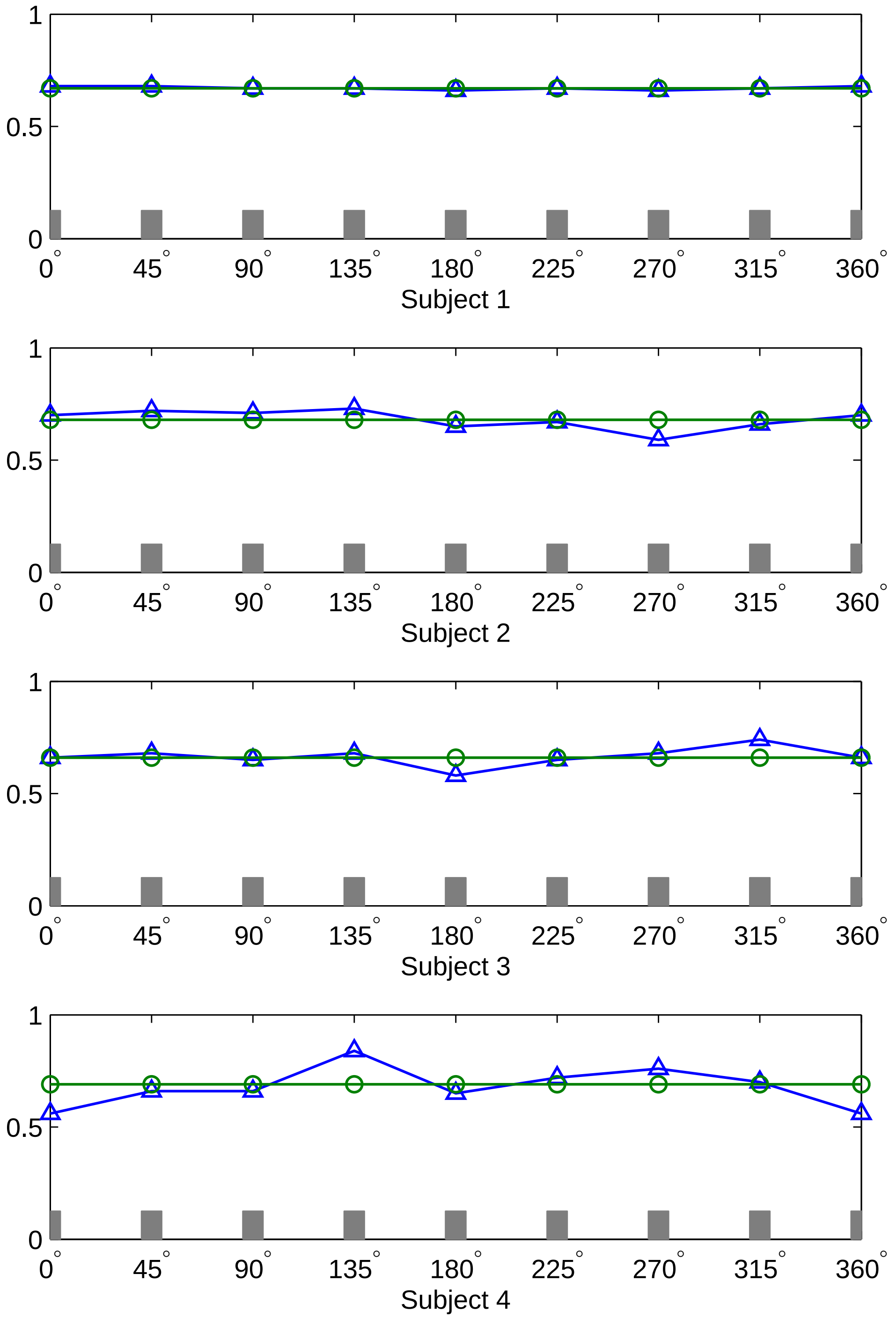

Figure 1: The experimental results of Shaw and Shaw (1977), when the letters appear with equal frequency at all 8 locations. Solid grey bars show the frequency of appearance at each location. Triangles show the fraction of correct identifications at each location, for each subject. Circles show the predicted fraction of correct identifications at each location, under the hypothesis of uniform attention allocation, and an individual-specific processing capacity. Data from Table 1, Shaw and Shaw (1977).

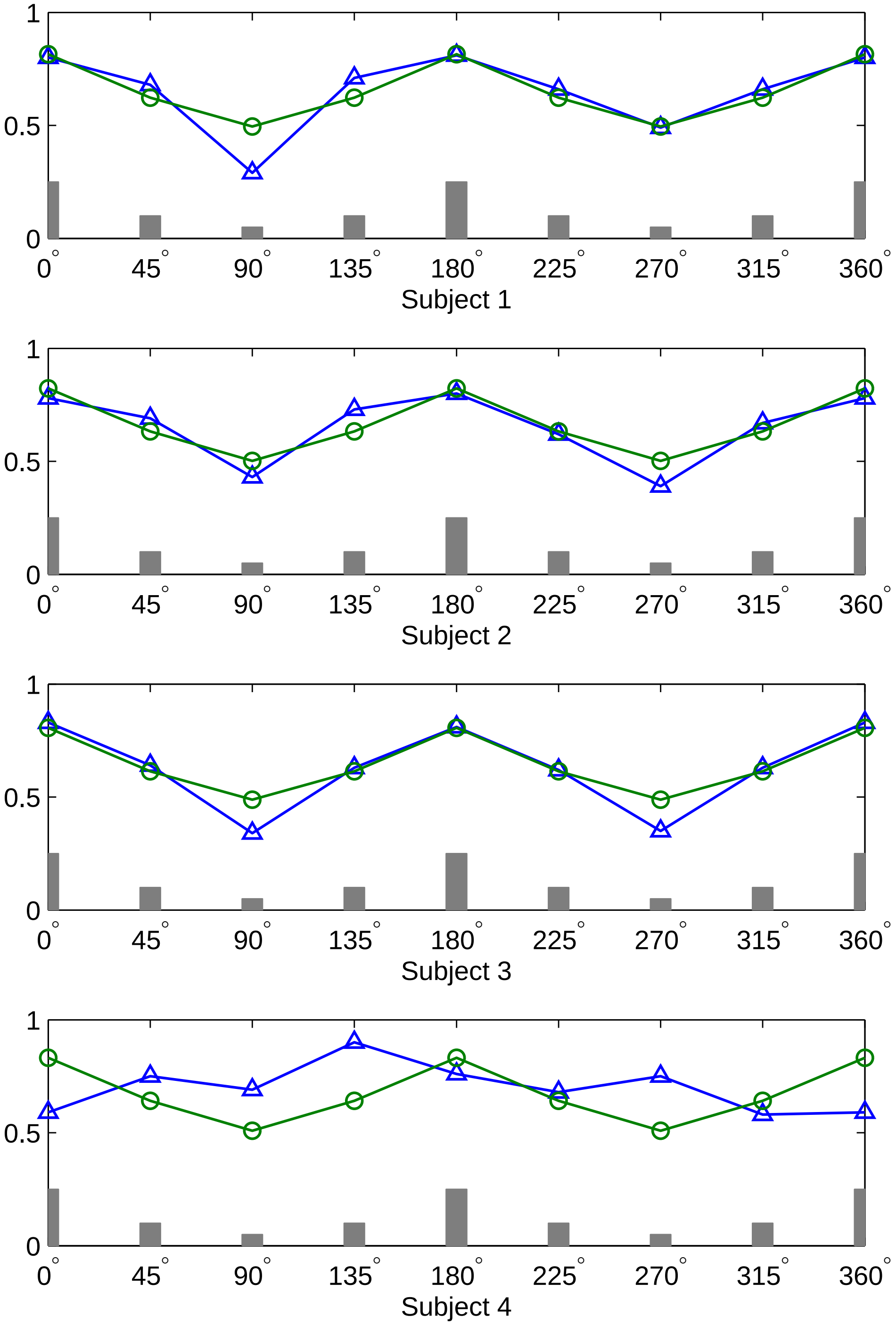

Figure 2: The experimental results of Shaw and Shaw (1977), when the letters appear with different frequencies at different locations. Solid grey bars show the frequency of appearance at each location. Triangles show the fraction of correct identifications at each location, for each subject. Circles show the predicted fraction of correct identifications at each location, under the criterion for optimal allocation of attention discussed in the text. Data from Table 2 , Shaw and Shaw (1977).

The experimental data are shown in Figures 1 and 2 for two different probability distributions $\{\pi_i\}$. Each panel plots (with triangles) the fraction of correct responses as a function of the location around the circle (indicated on the horizontal axis) for one of the four subjects.[6] In Figure 1, the probabilities of appearance at each location (indicated by the solid grey bars at the bottom of each panel) are equal across locations. In this case, for subjects 1-3, the frequency of correct discrimination is close to uniform across the eight locations; indeed, Shaw and Shaw report that one cannot reject the hypothesis that the error probability at each location is identical, and that the observed frequency differences are due purely to sampling error. (The behavior of subject 4 is more erratic, involving apparent biases toward paying more attention to certain locations, of a kind that do not represent an efficient adaptation to the task.)

Figure 2 then shows the corresponding fraction of correct responses at each location when the probabilities of the letters' appearing at the different locations are no longer uniform; as indicated by the grey bars, in this case the letters are most likely to appear at 0° or 180°, and least likely to appear at either 90° or 270°. The probabilities for the non-uniform case are chosen so that there are two locations, distant from one another, at *each* of which it will be desirable to pay particularly close attention; in this way the experiment is intended to test whether attention is *divisible* among locations, and not simply able to be *focused* on alternative directions. In fact, the non-uniform distribution continues to be symmetric with respect to reflections around both the vertical and horizontal axes; the symmetry of the task thus continues to encourage fixation of the subject's gaze in the exact center of the circle, as in the uniform case. Any change in the capacity for discrimination at the different locations should then indicate a change in the mental processing of visual information, rather than a simple change in the orientation of the eye.

As shown in Figure 2, the data reported by Shaw and Shaw indicate that in the case of all except subject 4, the frequency of correct discrimination does not remain constant across locations when the frequency of appearance at the different locations ceases to be uniform; instead, the frequency of correct responses rises at the locations that are used most frequently (0° and 180°) and falls at the locations that are used least frequently (90° and 270°). Thus subjects do appear to be able to reallocate

[6] The location labeled 0°, corresponding to the top of the circle, is shown twice (as both 0° and 360°), to make clear the symmetry of the setup.

their attention within the visual field, and to multiple locations, without doing so by changing their direction of gaze; and they seem to do this in a way that serves to increase their efficiency at letter-recognition, by allocating more attention to the locations where it matters more to their performance.

These results indicate that the nature of people's ability to discriminate between alternative situations is not a fixed characteristic of their sensory organs, but instead adapts according to the context in which the discrimination must be made. Nor are the results consistent with the view (as in the classic "signal detection theory" of Green and Swets, 1966) that each objective state is associated with a fixed probability distribution of subjective perceptions, and that is only the cutoffs that determine which subjective perceptions result in a particular behavioral response that adjust in response to changes in the frequency with which stimuli are encountered. For in moving between the first experimental situation and the second, the probability of presentative of an E as opposed to a T or V at any given location does not change; hence there is no reason for a change in a subject's propensity to report an E when experiencing a subjective perception that might represent either an E or a T at the 0° location. Evidently, instead, the degree of overlap between the probability distributions of subjective perceptions conditional upon particular objective states *changes* — becoming greater in the case of the different letters appearing at 90° and less in the case of the different letters appearing at 0°. But how can we model this change, and under what conception of the possibilities for such adaptation might the observed adaptation be judged an optimal response to the changed experimental conditions?

1.2.2 Sims's Hypothesis of Rational Inattention

Sims (1998, 2003, 2011) proposes a general theory of the optimal allocation of limited attention that might appear well-suited to the explanation of findings of this kind. Sims assumes that a DM makes her decision (*i.e.,* chooses her action) on the basis of a subjective perception (or mental representation) of the state of the world r, where the probability of experiencing a particular subjective perception r in the case that the true state of the world is x is determined by a set of conditional probabilities $\{p(r|x)\}$. The formalism is a very general one, that makes no general assumption about the kind of sets to which the possible values of x and r may belong. There is no assumption, for example, that x and r must be vectors of the same dimension;

indeed, it is possible that the set of possible values for one variable is continuous while the other variable is discrete. The hypothesis of rational inattention (RI) asserts that the set of possible representations r and the conditional probabilities $\{p(r|x)\}$ are precisely those that allow as high as possible a value for the DM's performance objective (say, the expected number of correct decisions), subject to an upper bound on the information that the representation conveys about the state.

The quantity of information conveyed by the representation is measured by Shannon's (1948) *mutual information,* defined as

$$I = \mathrm{E}\left[\log \frac{p(r|x)}{p(r)}\right] \tag{1.1}$$

where $p(r)$ is the frequency of occurrence of representation r (given the conditional probabilities $\{p(r|x)\}$ and the frequency of occurrence of each of the objective states x), and the expected value of the function of r and x is computed using the joint distribution for r and x implied by the frequency of occurrence of the objective states and the conditional probabilities $\{p(r|x)\}$. This can be shown (see, *e.g.,* Cover and Thomas, 2006) to be the average amount by which observation of r reduces uncertainty about the state x, if the ex ante uncertainty about x is measured by the entropy

$$H(X) \equiv -\mathrm{E}\left[\log \pi(x)\right],$$

where $\pi(x)$ is the (unconditional) probability of occurrence of the state x, and the uncertainty after observing r is measured by the corresponding entropy, computed using the conditional probabilities $\pi(x|r)$. Equivalently, the mutual information is the average amount by which knowledge of the state x would reduce uncertainty (as measured by entropy) about what the representation r will be.[7] Not only is this concept defined for stochastic representations; the proposed form of constraint implies that there is an advantage to stochastic representations, insofar as a "fuzzier" relation between x and r reduces the mutual information, and so relaxes the constraint.

[7] The formula (1.1) for mutual information follows directly from the definition of entropy and this second characterization. While the first characterization provides better intuition for why this should be a reasonable measure of the informativeness of the representation r, I have written the formula (1.1) in terms of the conditional probabilities $\{p(r|x)\}$ rather than the $\{\pi(x|r)\}$, because this expression makes it more obvious how the choice of the conditional probabilities $\{p(r|x)\}$, which are a more natural way of specifying the design problem, is constrained by a bound on the mutual information.

Rather than assuming that some performance measure Π is maximized subject to an upper bound on I, one might alternatively suppose that additional information-processing capacity can be allocated to this particular task at a cost, and that the information structure and decision rule are chosen so as to maximize $\Pi - \theta I$, where $\theta > 0$ is a unit cost of information-processing capacity.[8] This latter version of the theory assumes that the DM is constrained only by some bound on the sum of the information processing capacity used in each of some large number of independent tasks; if the information requirements of the particular task under analysis are small enough relative to this global constraint, the shadow cost of additional capacity can be treated as independent of the quantity of information used in this task. A constrained-optimal information structure in any given problem can be equally well described in either of the two ways (as maximizing Π given the quantity of information used, or as maximizing $\Pi - \theta I$ for some shadow price θ); the distinction matters, however, when we wish to ask how the information structure should change when the task changes, as in the movement from the first experimental situation to the second in experiment of Shaw and Shaw. We might assume that the bound on I remains unchanged when the probabilities $\{\pi_i\}$ change, or alternatively we might assume that the shadow price θ should remain unchanged across the two experiments. The latter assumption would imply not only that attention can be reallocated among the different locations that may be attended to in the experiment, but that attention can also be reallocated between this experiment and other matters of which the subject is simultaneously aware.

Because Sims's measure of the cost of being better informed implies that allowing a greater degree of overlap between the probability distributions of subjective representations associated with different objective states reduces the information cost, it might seem to be precisely the sort of measure needed to explain the results obtained by Shaw and Shaw (for their first three subjects) as an optimal adaptation to the change in the experimental setup. But in fact it makes no such prediction.

Suppose that (as in the pure formulation of Sims's theory) there are no other constraints on what the set of possible representations r or the conditional probabilities $\{p(r|x)\}$ may be. In the experiment of Shaw and Shaw, the state x (the objective properties of the stimulus on a given trial) has two dimensions, the location i at which the stimulus appears, and the letter j that appears, and under the prior these

[8] This is the version of the theory used, for example, in Woodford (2009).

two random variables are distributed independently of one another. In addition, only the value of j is payoff-relevant (the subject's reward for announcing a given letter is independent of the location i, but depends on the true letter j). Then it is easy to show that an optimal information structure will provide no information about the value of i: the conditional probabilities $p(r|x) = p(r|ij)$ will be functions only of j, and so can be written $p(r|j)$.

The problem then reduces to the choice of a set of possible representations r and conditional probabilities $\{p(r|j)\}$ so as to maximize the probability of a correct response subject to an upper bound on the value of

$$I = \mathrm{E}\left[\log \frac{p(r|j)}{p(r)}\right],$$

where the expectation $\mathrm{E}[\cdot]$ now represents an integral over the joint distribution of j and r implied by the conditional probabilities. This problem depends on the prior probabilities of appearance of the different letters j, but does not involve the prior probabilities of the different locations $\{\pi_i\}$. Since the prior probabilities of the three letters are *the same* across the two experimental designs, the solution to this optimum problem is the same, and this version of RI theory implies that the probability of correct responses at each of the eight locations should be identical across the two experiments. This is of course not at all consistent with the experimental results of Shaw and Shaw.

Why is this theory inadequate? Under the assumption that the DM could choose to pay attention solely to the letter that appears and not to its location, it would clearly be optimal to ignore the latter information; and there would be no reason for the subject's information-processing strategy to be location-dependent, as it is evidently is under the second experimental design. It appears, then, that it is *not* possible (or at any rate, not costlessly possible) to first classify stimuli as Es, Ts or Vs, and then subsequently decide how much information about that summary statistic to pass on for use in the final decision. It is evidently necessary for the visual system to *separately* observe information about what is happening at each of the eight different locations in the visual field, and at least some of the information-processing constraint must relate to the separate processing of these individual information streams — as opposed to there being only a constraint on the rate of information flow to the final decision stage, after the information obtained from the different streams has been optimally combined.[9]

Let us suppose, then, that the only information structures that can be considered are ones under which the subject will necessarily be aware of the location i at which the letter has appeared (though not necessarily making a correct identification of the letter that has appeared there). One way of formalizing this constraint is to assume that the set of possible representations R must be of the form

$$R = \bigcup_{i=1}^{8} R_i, \tag{1.2}$$

and that the conditional probabilities must satisfy

$$p(R_i|ij) = 1 \qquad \forall i, j \tag{1.3}$$

We then wish to consider the choice of an information structure and decision rule to maximize the expected number of correct responses, subject to constraints (1.2)–(1.3) and an upper bound on the possible value of the quantity I defined in (1.1).

As usual in problems of this kind, one can show that an optimal information structure reveals only the choice that should be made as a result of the signal; any additional information would only increase the size of the mutual information I with no improvement in the probability of a correct response.[10] Hence we may suppose that the subjective representation is of the form ik, where i indicates the location at which a letter is seen (necessarily revealed, by assumption) and k is the response that the subject gives as a result of this representation. We therefore need only to specify the conditional probabilities $\{p(ik|ij)\}$ for $i = 1, \ldots, 8$, and $j, k = 1, 2, 3$. Moreover, because of the symmetry of the problem under permutations of the three letters, it is easily seen that the optimal information structure must possess the same symmetry.

[9] The issue is one that arises in macroeconomic applications of RI theory, whenever there is a possibility of observing more than one independent aspect of the state of the world. For example, Mackowiak and Wiederholt (2009) consider a model in which both aggregate and idiosyncratic shocks have implications for a firm's optimal price, and assume a form of RI theory in which firms must observe separate signals (each more or less precise, according to the firm's attention allocation decision) about the two types of shocks, rather than being able to observe a signal that is a noisy measurement of an optimal linear combination of the two state variables. This is effectively an additional constraint on the set of possible information structures, and it is of considerable importance for their conclusions.

[10] See the discussion in Woodford (2008), in the context of a model with a binary choice.

Hence the conditional probabilities must be of the form

$$\begin{aligned} p(ij|ij) &= 1 - e_i \qquad \forall i, j, & (1.4) \\ p(ik|ij) &= e_i/2 \qquad \forall i, j, \;\; \text{any} \;\; k \neq j, & (1.5) \end{aligned}$$

where e_i is the probability of error in the identification of a letter that appears at location i.

With this parameterization of the information structure, the mutual information (1.1) is equal to

$$I = -\sum_i \pi_i \log \pi_i + \log 3 - \sum_i \pi_i h(e_i), \tag{1.6}$$

where

$$h(e) \equiv -(1-e)\log(1-e) - e\log(e/2)$$

is the entropy of a three-valued random variable with probabilities $(1-e, e/2, e/2)$ of the three possible outcomes.[11] The optimal information structure subject to constraints (1.2)–(1.3) and an upper bound on I will then correspond to the values $\{e_i\}$ that minimize

$$\sum_i \pi_i e_i + \theta I(e), \tag{1.7}$$

where $I(e)$ is the function defined in (1.6), and $\theta \geq 0$ is a Lagrange multiplier associated with the upper-bound constraint. (Alternatively, if additional information-processing capacity can be allocated to this task at a cost, θ measures that cost.)

Note that the objective (1.7) is additively separable; this means that for each i, the optimal value of e_i is the one that minimizes

$$e_i - \theta h(e_i),$$

[11] The derivation of (1.6) is most easily understood as a calculation of the average amount by which knowledge of the state ij reduces the entropy of the subjective representation ik. The unconditional entropy (before knowing the state) of the subjective representation is given by the sum of the first two terms on the right-hand side, which represent the entropy of the location perception (8 possibilities with ex ante probabilities $\{\pi_i\}$) and the entropy of the letter perception (3 possibilities, equally likely ex ante) respectively. The final term on the right-hand side subtracts the average value of the entropy conditional upon the state; the conditional entropy of the location perception is zero (it can be predicted with certainty), while the conditional entropy of the letter perception is $h(e_i)$ if the location is i.

regardless of the values chosen for the other locations. Since this function is the same for all i, the minimizing value $\bar{e}$ is the same for all i as well. (One can easily show that the function $h(e)$ is strictly convex, so that the minimum is unique for any value of θ.) Thus we conclude once again that under this measure of the cost of more precise awareness, non-uniformity of the location probabilities should *not* make it optimal for subjects to make fewer errors at some locations than others. If the shadow cost of additional processing capacity is assumed to be the same across the two experiments, then constancy of the value of θ would imply that the value of $\bar{e}$ should be the same for each subject in the two experiments. If instead it is the upper bound on I that is assumed to be the same across the two experiments, then the reduction in the entropy of the location in the second experiment (because the probabilities are no longer uniform, there is less uncertainty ex ante about what the location will be) means that more processing capacity should be available for transmission of more accurate signals about the identity of the letter, and the value of $\bar{e}$ should be substantially lower (the probability of correct identifications should be higher) in the second experiment. (This prediction is also clearly rejected by the data of Shaw and Shaw.) But in either case, the probability of correct identification should be the same across all locations, in the second experiment as much as in the first, a prediction that is not confirmed by the data.

Why does the mutual information criterion not provide a motive for subjects to reallocate their attention when the location probabilities are non-uniform? Mutual information measures the average degree to which the subjective representation reduces entropy, weighting each possible representation by the probability with which it is used. This means that arranging for available representations that will be highly informative about low-probability states when they occur is *not* costly, except in proportion to the probability of occurrence of those states. And while the expected benefit of being well-informed about low-probability states is small, there remain benefits of being informed about those states — proportional to the probability that the states will occur. Hence the fact that some states occur with much lower probability than others *does not* alter the ratio of cost to benefit of a given level of precision of the subjective representation of those states.

But this means that theory of rational inattention, as formulated by Sims, cannot account for reallocation of attention of the kind seen in the experiment of Shaw and Shaw. We need instead a measure of the cost of more precise awareness that implies that it is costly to be *able* to discriminate between low-probability states (say, an E as opposed to a T at the 90° location), even if one's *capacity* to make such a discrimination is not exercised very frequently.

1.2.3 An Alternative Information-Theoretic Criterion

One possibility is to assume that the information-processing capacity required in order to arrange for a particular stochastic relation $\{p(r|x)\}$ between the subjective representation and the true state depends not on the *actual* amount of information about the state that is transmitted on average, given the frequency with which different states occur, but rather on the *potential* rate of information transmission by this system, in the case of *any* probabilities of occurrence of the states x. Under this alternative criterion, it is costly to arrange to have precise awareness of a low-probability state in the case that it occurs; because even though the state is not expected to occur very often, a communication channel that can provide such precise awareness when called upon to do so is one that could transmit information at a substantial rate, in a world in which the state in question occurred much more frequently. We may then suppose that the information-processing capacity required to implement such a stochastic relation will be substantial.

Let the mutual information measure defined in (1.1) be written as $I(p;\pi)$, where p refers to the set of conditional probabilities $\{p(r|x)\}$ that specify how subjective representations are related to the actual state, and π refers to the prior probabilities $\{\pi(x)\}$ with which different states are expected to occur. (The set of possible subjective representations R is implicit in the specification of p.) Then the proposed measure of the information-processing capacity required to implement a given stochastic relation p can be defined as[12]

$$C = \max_{\pi} I(p;\pi). \tag{1.8}$$

This measure of required capacity depends only on the stochastic relation p. I propose to consider a variant of Sims's theory of rational inattention, according to which any stochastic relation p between subjective representations and actual states is possible, subject to an upper bound on the required information-processing capacity C.

[12] Note that this is just Shannon's definition of the *capacity* of a communication channel that takes as input the value of x and returns as output the representation r, with conditional probabilities given by p.

Alternatively, we may suppose that there is a cost of more precise awareness that is proportional to the value of C, rather than to the value of I under the particular probabilities with which different states are expected to be encountered.

Let us consider the implications of this alternative theory for the experiment of Shaw and Shaw (1977). I shall again suppose that possible information structures must respect the restrictions (1.2)–(1.3), and shall also again consider only symmetric structures of the form (1.4)–(1.5). Hence the information structure can again be parameterized by the 8 coefficients $\{e_i\}$. But instead of assuming that these coefficients are chosen so as to minimize the expected fraction of incorrect identifications subject to an upper bound on I, I shall assume that the expected fraction of incorrect identifications is minimized subject to an upper bound on C. Alternatively, instead of choosing them to minimize (1.7) for some $\theta \geq 0$, they will be chosen to minimize

$$\sum_i \pi_i e_i + \theta C(e) \tag{1.9}$$

for some $\theta \geq 0$, where $C(e)$ is the function defined by (1.8) when the conditional probabilities p are given by (1.4)–(1.5).

For an information structure of this form, the solution to the optimization problem in (1.8) is given by

$$\pi_i^* = \frac{\exp\{-h(e_i)\}}{\sum_j \exp\{-h(e_j)\}}$$

for all i. Substituting these probabilities into the definitition of mutual information, we obtain

$$C(e) = I(p; \pi^*) = \log 3 + \log\left(\sum_i \exp\{-h(e_i)\}\right).$$

The first-order conditions for the problem (1.9) are then of the form

$$\pi_i = \tilde{\theta} \exp\{-h(e_i)\} h'(e_i) \tag{1.10}$$

for each i, where $\tilde{\theta} \equiv \theta / \sum_j \exp\{-h(e_j)\}$ will be independent of i. Because the right-hand side of (1.10) is a monotonically decreasing function of e_i, the solution for e_i will vary inveresely with π_i. That is, under the optimal information structure, the probability of a correct identification will be highest at those locations where the letter is most likely to occur, as in the results of Shaw and Shaw.

Indeed, the proposed theory makes very specific quantitative predictions about the experiment of Shaw and Shaw. Let us suppose that the shadow value θ of additional information-processing capacity remains constant across the two experiments.[13] Then the observed frequencies of correct identification in the case of the uniform location probabilities can be used to identify the value of θ for each subject. Given this value, the theory makes a definite prediction about each of the e_i in the case of non-uniform location probabilities. For the parameter values of the Shaw and Shaw experiment, these theoretical predictions are shown by the circles in each panel of Figure 2.[14] For each of the first three subjects (i.e., the ones with roughly optimal allocation of attention in the first experiment), the predictions of the theory are reasonably accurate.[15] Hence the reallocation of attention reported by Shaw and Shaw is reasonably consistent with a version of the theory of rational inattention, in which the only two constraints on the possible information structure are (i) the requirement that the subject be aware of the location of the letter, and (ii) an upper bound on the channel capacity C.

1.3 Visual Adaptation to Variations in Illumination

One of the best-established facts about perception is that the subjective perception of a given stimulus depends not just on its absolute intensity, but on its intensity relative to some background or reference level of stimulation, to which the organism has become accustomed.[16] Take the example of the relation between the *luminance* of objects in one's visual field — the intensity of the light emanating from them, as measured by photometric equipment — and subjective perceptions of their brightness. We have all experienced being temporarily blinded when stepping from a dark area

[13] The numerical results shown in Figure 2 are nearly identical in the case that the upper bound on C is assumed to be constant across the two experiments, rather than the shadow cost θ.

[14] The value of θ used for each subject is the one that would imply a value of $\bar{e}$ in the first experiment equal to the one indicated in Table 1 of Shaw and Shaw (1977).

[15] They are certainly more accurate than the predictions of the alternative theory according to which the information structure minimizes (1.7), with the value of θ again constant across the two experiments. The likelihood ratio in favor of the new theory is greater than 10^{21} in the case of the data for subject 1, greater than 10^{15} for subject 2, and greater than 10^{30} for subject 3. The likelihood is instead higher for the first theory in the case of subject 4, but the data for subject 4 are extremely unlikely under either theory. (Under a chi-squared goodness-of-fit test, the p-value for the new theory is less than 10^{-14}, but it is on the order of 10^{-11} for the first theory as well.)

[16] See, *e.g.*, Gabbiani and Cox (2010), chap. 19; Glimcher (2011), chap. 12; Kandel, Schwartz and Jessel, 2000, chap. 21; or Weber (2004).

into bright sunlight. At first, visual discrimination is difficult between different (all unbearably bright) parts of the visual; but one's eyes quickly "adjust," and it is soon possible to see fairly "normally." Similarly, upon first entering a dark room, it may be possible to see very little; yet, after one's eyes adjust to the low illumination, one finds that different objects in the room can be seen after all. These observations indicate that one's ability to discriminate between different levels of luminance is not fixed; the contrasts between different levels that are perceptible depend on the mean level of luminance (or perhaps the distribution of levels of luminance in one's environment) to which one's eyes have adapted.

It is also clear that the subjective perception of a given degree of luminance changes in different environments. The luminance of a given object — say, a white index card — varies by a factor of 10^6 between the way it appears on a moonlit night and in bright sunlight (Gabbiani and Cox, 2010, Figure 19.1). Yet one's subjective perception of the brightness of objects seen under different levels of illumination does not vary nearly so violently. The mapping from objective luminance to the subjective representation of brightness evidently varies across environments. It is also not necessarily the same for all parts of one's visual field at a given point in time. Looking at a bright light, then turning away from it, results in an *after-effect,* in which part of one's visual field appears darkened for a time. After one has gotten used to high luminance in that part of the visual field, a more ordinary level of luminance seems dark — but this is not true of the other parts of one's visual field, which have not similarly adjusted. Similarly, a given degree of objective luminance in different parts of one's visual field may simultaneously appear brighter or darker, depending on the degree of luminance of nearby surfaces in each case, giving rise to a familiar optical illusion.[17]

Evidence that the sensory effects of given stimuli depend on how they compare to prior experience need not rely solely on introspection. In the case of non-human organisms, measurements of electrical activity in the nervous system confirm this, dating from the classic work of Adrian (1928). For example, Laughlin and Hardie (1978) graph the response of blowfly and dragonfly photoreceptors to different intensities of light pulses, when the pulses are delivered against various levels of background lumi-

[17] For examples, see Frisby and Stone (2010), Figures 1.12, 1.13, 1.14, 16.1, 16.9, and 16.11. Kahneman (2003) uses an illusion of this kind as an analogy for reference-dependence of economic valuations.

nance. The higher the background luminance, the higher the intensity of the pulse required to produce a given size of response (deflection of the membrane potential). Laughlin and Hardie point out that the effect of this adaptation is to make the signal passed on to the next stage of visual processing more a function of *contrast* (*i.e.*, of luminance relative to the background level) than of the absolute level of luminance (p. 336).

An important recent literature argues that the neural coding of stimuli depends not merely on some average stimulus intensity to which the organism has been exposed, but on the complete *probability distribution* of stimuli encountered in the organism's environment. For example, Laughlin (1981) records the responses (changes in membrane potential) of the large monopolar cell (LMC) in the compound eye of the blowfly to pulses of light that are either brighter or darker than the background level of illumination to varying extents. His experimental data are shown in Figure 3 by the black dots with whiskers. The change in the cell membrane potential in response to the pulse is shown on the vertical axis, with the maximum increase normalized as +1 and the maximum decrease as -1.[18] The intensity of the pulse is plotted on the horizontal axis in terms of contrast,[19] as Laughlin and Hardie (1978) had already established that the LMC responds to contrast rather than to the absolute level of luminance.

Laughlin also plots an empirical frequency distribution for levels of contrast in the visual environment of the blowflies in question. The cumulative distribution function (cdf) is shown by the solid line in the figure.[20] Laughlin notes the similarity between the graph of the cdf and the graph of the change in membrane potential. They are not quite identical; but one sees that the potential increases most rapidly — allowing sharper discrimination between nearby levels of luminance — over the range of contrast levels that occur most frequently in the natural environment, so

[18] For each level of contrast, the whiskers indicate the range of experimental measurements of the response, while the dot shows the mean response.

[19] This is defined as $(I - I_0)/(I + I_0)$, where I is the stimulus luminance and I_0 is the background luminance. Thus contrast is a monotonic function of relative luminance, where 0 means no difference from the background level of illumination, +1 is the limiting case of infinitely greater luminance than the background, and -1 is the limiting case of a completely dark image.

[20] The cdf is plotted after a linear transformation so that it varies from -1 to +1 rather than from 0 to 1.

that the cdf is also rapidly increasing.[21] Thus Laughlin proposes not merely that the visual system of the fly responds to contrast rather than to the absolute level of luminance, but that the degree of response to a given variation in contrast depends on the degree of variation in contrast found in the organism's environment. This, he suggests, represents an efficient use of the LMC's limited range of possible responses: it "us[es] the response range for the better resolution of common events, rather than reserving large portions for the improbable" (p. 911).

The adaptation to the statistics of the natural environment suggested by Laughlin might be assumed to have resulted from evolutionary selection or early development, and not to be modified by an individual organism's subsequent experience. However, other studies find evidence of adaptation of neural coding to statistical properties of the environment that occurs fairly rapidly. For example, Brenner *et al.* (2000) find that a motion-sensitive neuron of the blowfly responds not simply to motion relative to a background rate of motion, but to the difference between the rate of motion and the background rate, *rescaled* by dividing by a local (time-varying) estimate of the standard deviation of the stimulus variability. Other studies find that changes in the statistics of inputs change the structure of retinal receptive fields in predictable ways.[22]

These studies all suggest that the way in which stimuli are coded can change with changes in the distribution of stimuli to which a sensory system has become habituated. But can such adaptation be understood as the solution to an optimization problem? The key to this is a correct understanding of the relevant constraints on the processing of sensory information.

1.4 Adaptation as Optimal Coding

Let us suppose that the frequency distribution of degrees of luminance in a given environment is log-normally distributed; that is, log luminance is distributed as $N(\mu, \sigma^2)$ for some parameters μ, σ.[23] We wish to consider the optimal design of a perceptual

[21] It is worth recalling that the probability density function (pdf) is the derivative of the cdf. Thus a more rapid increase in the cdf means that the pdf is higher for that level of contrast.

[22] See Dayan and Abbott (2001), chap. 4; Fairhall (2007); or Rieke *et al.* (1997), chap. 5, for reviews of this literature.

[23] The histograms shown in Figure 19.4 of Gabbiani and Cox (2010) for the distribution of luminance in natural scenes suggest that this is not an unreasonable approximation.

system, in which a subjective perception (or neural representation) of brightness r will occur with conditional probability $p(r|x)$ when the level of log luminance is x. By optimality I mean that the representation is as accurate as possible, on average, subject to a constraint on the information-processing requirement of the system. Let us suppose that the relevant criterion for accuracy is minimization of the mean squared error of an estimate $\hat{x}(r)$ of the log luminance based on the subjective perception r.[24]

... Note that it is important to distinguish between the subjective perception r and the estimate of the luminance that one should make, given awareness of r. For one thing, r need not itself be assumed to be commensurable with luminance (it need not be a real number, or measured in the same units), so that it may not be possible to speak of the closeness of the representation r itself to the true state x. But more importantly, I do not wish to identify the subjective representation r with the optimal inference that should be made from it, because the mapping from r to $\hat{x}(r)$ should change when the prior and/or the coding system changes. Experiments that measure electrical potentials in the nervous system associated with particular stimuli, like those discussed above, are documenting the relationship between x and r, rather than between x and an optimal estimate of x. Similarly, the observation that the subjective perception of the brightness of objects in different parts of the visual field can be different depending on the luminance of nearby objects in each region is an observation about the context-dependence of the mapping from x to r, and not direct evidence about how an optimal estimate of luminance in different parts of the visual field should be formed.

The solution to this optimization problem depends on the kind of constraint on information-processing capacity one assumes. Suppose, for example, that we assume an upper bound on the *number* of distinct representations r that may be used, and no other constraints, as in Gul *et al.* (2011). In this case, it is easily shown that

[24] Our criteria for the accuracy of perceptions would be possible, of course. This one has the consequence that, under any of the possible formulations of the constraint on the information content of subjective representations considered below, the optimal information structure will conform to Weber's Law, in the formulation given by Thurstone (1959) cited above in section 1.1. That is, for any threshold $0 < p < 1$, the probability that a given stimulus S will be judged brighter than a stimulus with the mean level of luminance will be less than p if and only if the luminance of S is less than some multiple of the mean luminance, where the multiple depends on p and σ, but is independent of μ — *i.e.,* independent of the mean level of luminance to which the perceptual system is adapted.

an optimal information structure partitions the real line into a N intervals (each representing a range of possible levels of luminance), each of which is assigned a distinct subjective representation r. The optimal choice of the boundaries for these intervals is a classic problem in the theory of optimal coding; the solution is given by the algorithm of Lloyd and Max (Sayood, 2005, chap. 9).

This sort of information structure does not, however, closely resemble actual perceptual processes. It implies that while varying levels of luminance over some range should be completely indistinguishable from one another, it should be possible to find two levels of luminance x_1, x_2 that differ only infinitesimally, and yet are perfectly discriminable from one another (because they happen to lie on opposite sides of a boundary between two intervals that are mapped to different subjective representations). This sort of discontinuity is, of course, never found in psychophysical or neurological studies.

If we instead assume an upper bound I on the mutual information between the state x and the representation r, in accordance with the "rational inattention" hypothesis of Sims, this is another problem with a well-known solution (Sims, 2011). One possible representation of the optimal information structure is to suppose that the subjective perception is a real number, equal to the true state plus an observation error,

$$r = x + \epsilon, \tag{1.11}$$

where the error term ϵ is an independent draw from a Gaussian distribution $N(0, \sigma_\epsilon^2)$, where

$$\frac{\sigma_\epsilon^2}{\sigma^2} = \frac{e^{-2I}}{1 - e^{-2I}}.$$

Thus the signal-to-noise ratio of the noisy percept is an increasing function of the bound I, falling to zero as I approaches zero, and growing without bound as I is made unboundedly large.

In this model of imperfect perception, there is no problem of discontinuity: the probability that the subjective representation will belong to any subset of the set R of possible representations is now a continuous function of x. But this model fails to match the experimental evidence in other respects. Note that the optimal information structure (1.11) is independent of the value of μ. Thus the model implies that the discriminability of two possible levels of luminance x_1, x_2 should be *independent* of the mean level of luminance in the environment to which the visual system has adapted;

but in that case there should be no difficulty in seeing when abruptly moving to an environment with a markedly different level of illumination. Similarly, it implies that the degree of discriminability of x_1 and x_2 should depend only on the distance $|x_1 - x_2|$, and not on where x_1 and x_2 are located in the frequency distribution of luminance levels. But this is contrary to the observation of Laughlin (1981) that finer discriminations are made among the range of levels of illumination that occur more frequently.

Moreover, according to this model, there is no advantage to responding to contrast rather than to the absolute level of illumination: a subjective representation of the form (1.11), which depends on the absolute level of illumination x and *not* on contrast $(x - \mu)$, is fully optimal.[25] This leaves it a mystery why response to contrast is such an ubiquitous feature of perceptual systems. Moreover, since the model implies that there should be no need to recalibrate the mapping of objective levels of luminance into subjective perceptions when the mean level of luminance in the environment changes, it provides no explanation for the existence of after-effects or lightness illusions.

The problem with the mutual information criterion seems, once again, to be the fact that there is no penalty for making fine discriminations among states that seldom occur: such discriminations make a small contribution to mutual information as long as they are infrequently used. Thus the information structure (1.11) involves not only an extremely large set of different possible subjective representations (one with the cardinality of the continuum), but nearly all of them (all $r << \mu$ and all $r >> \mu$) are subjective representations that are mainly used to distinguish among different states that are far out in the tails of the frequency distribution. As a consequence, the observation of Laughlin (1981) that it would be inefficient for neural coding to leave "large parts of the response range [of a neuron] underutilized because they correspond

[25] It is true that the representation given in (1.11) is not *uniquely* optimal; one could also have many other optimal subjective representations, including one in which $r = (x - \mu) + \epsilon$, so that the representation depends only on contrast. The reason is that Sims' theory does not actually determine the representations r at all, only the degree to which the distributions $p(r|x)$ for different states x overlap one another. However, the theory provides no reason for the representation of contrast to be a superior approach. Furthermore, if one adds to the basic theory of rational inattention a supposition that there is even a tiny cost of having to code stimuli differently in different environments, as surely there should be, then the indeterminacy is broken, and the representation (1.11) is found to be uniquely optimal.

to exceptionally large excursions of input" (p. 910) is completely inconsistent with the cost of information precision assumed in RI theory.

As in the previous section, the alternative hypothesis of an upper bound on the capacity requirement C defined in (1.8) leads to predictions more similar to the experimental evidence. The type of information structure that minimizes mean squared error subject to an upper bound on C involves only a finite number of distinct subjective representations r, which are used more to distinguish among states in the center of the frequency distribution than among states in the tails. Figure 4 gives, as an example, the optimal information structure in the case that the upper bound on C is equal to only one-half of a binary digit.[26] In this case, the optimal information structure involves three distinct possible subjective representations (labeled 1, 2, and 3), which one may think of as subjective perceptions of the scene as "dark," "moderately illuminated," and "bright" respectively. The lines in the figure indicate the conditional probability of the scene being perceived in each of these three ways, as a function of the objective log luminance x.[27]

In panel (a), the prior distribution has a mean of -2 and a standard deviation of 1, while in panel (b), the mean is 2 and the standard deviation is again equal to 1. One observes that the shift in the mean luminance between the two cases shifts the functions that indicate the conditional probabilities. In the high-average-luminance environment, a log luminance of zero has a high probability of being perceived as "dark" and only a negligible probability of being perceived as "bright," while in the low-average-luminance environment, the same stimulus has a high probability of being perceived as "bright" and only a negligible probability of being perceived as "dark." Thus the theory predicts that perceptions of brightness are recalibrated depending on the mean luminance of the environment. In fact, the figure shows that for a fixed value of σ, subjective perceptions of brightness are predicted to be functions *only of contrast*, $x - \mu$, rather than of the absolute level of luminance. Hence the theory is consistent both with the observed character of neural coding and with subjective

[26] If the logarithm in (1.1) is a natural logarithm, then this corresponds to a numerical value $C = 0.5 \log 2$. For those readers who may have difficulty imagining "half of a binary digit": a communication channel with this capacity can transmit the same amount of information, on average, in each two transmissions as can be transmitted in each individual transmission using a channel which can send the answer to one yes/no question with perfect precision.

[27] The equations that are solved to plot these curves are stated in section 2, and the numerical algorithm used to solve them is discussed in the Appendix.

experiences of after-effects and lightness illusions.

The theory also predicts that finer discriminations will be made among levels of luminance that occur more frequently, in the environment to which the perceptual system has adapted. One way to discuss the degree of discriminability of nearby levels of luminance is to plot the Fisher information,

$$I^{Fisher}(x) \equiv -\sum_r p(r|x) \frac{\partial^2 \log p(r|x)}{(\partial x)^2},$$

as a function of the objective state x, where the sum is over all possible subjective representations r in the case of that state.[28] This function is shown in the two panels of Figure 5, for the two information structures shown in the corresponding panels of Figure 4. In each panel, the solid line plots the Fisher information for the information structure shown in Figure 4 (the optimal structure subject to an upper bound on C), while the dashed line plots the Fisher information for the optimal information structure in the case of the same prior distribution, but where the structure is optimized subject to an upper bound on the mutual information I (also equal to one-half a binary digit).

As discussed above, when the relevant constraint is the mutual information (Sims's RI hypothesis), the optimal structure discriminates equally well among nearby levels of luminance over the entire range of possible levels: in fact, $I^{Fisher}(x)$ is constant in this case. In the theory proposed here instead (an upper bound on C), the optimal information structure implies a greater ability to discriminate among alternative states within an interval concentrated around the mean level of log luminance μ, but almost no ability to discriminate among alternative levels of luminance when these are either all more than one standard deviation below the mean, or all more than one standard deviation above the mean. Hence the theory predicts that someone moving from one of these two environments to the other should have very poor vision, until their visual system adapts to the new environment. The theory is also reasonably consistent with Laughlin's (1981) observations about the visual system of the fly: not only that only contrast is perceived, but that sharper discriminations are made among nearby levels of contrast in the case of those levels of contrast that occur most frequently in the environment.

[28] For the interpretation of this as a measure of the discriminability of nearby states in the neighborhood of a given state x, see, *e.g.*, Cox and Hinkley (1974).

Both this application and the one in the previous section, then, suggest that the hypothesis of an optimal information structure subject to an upper bound on the channel capacity C required to implement it can explain at least some important experimental findings with regard to the nature of visual perception. Since the hypothesis formulated in this way is of a very general character, and not dependent on special features of the particular problems in visual perception discussed above, it may be reasonable to conjecture that the same principle should explain the character of perceptual limitations in other domains as well.

2 A Model of Inattentive Valuation

I now wish to consider the implications of the theory of partial awareness proposed in the previous section for the specific context of economic choice. I shall consider the hypothesis that economic decisionmakers, when evaluating the options available to them in a situation requiring them to make a choice, are only partially aware of the characteristics of each of the options. But I shall give precise content to this hypothesis by supposing that the particular imprecise awareness that they have of each of their options represents an optimal allocation of their scarce information-processing capacity. The specific constraint that this imposes on possible relations between subjective valuations and the objective characteristics of the available options is modeled in a way that has been found to explain at least certain features of visual perception, as discussed in the previous section.

2.1 Formulation of the Problem

As an example of the implications of this theory, let us suppose that a DM must evaluate various options x, each of which is characterized by a value x_a for each of n distinct *attributes*. I shall suppose that each of the n attributes must be observed *separately,* and that it is the capacity required to process these separate observations that represents the crucial bottleneck that results in less than full awareness of the characteristics of the options. As a consequence, the subjective representation of each option will also have n components $\{r_a\}$, though some of these may be *null* representations — in the sense that the value of component r_a for some a may be the same for all options, so that there is no awareness of differences among the options on this attribute. The DM's partial awareness can then be specified by a collection of conditional probabilities $\{p_a(r_a|x_a)\}$ for $a = 1, \ldots, n$. Here it is assumed that the probability of obtaining a particular subjective representation r_a of attribute a depends only on the true value x_a of this particular attribute; this is the meaning of the assumption of independent observations of the distinct attributes.[29]

The additional constraint that I shall assume on possible information structures is an upper bound on the required channel capacity (1.8). Because of the assumed decomposability of the information structure into separate signals about each of the attributes a, the solution for the optimal prior probabilities π^* in problem (1.8) can be obtained by separately choosing prior probabilities π^*_a for each attribute a that solve the problem

$$\max_{\pi_a} I(p_a; \pi_a), \tag{2.1}$$

which depends only on the type of signal that is obtained about attribute a. (This follows from the standard result that the total capacity, in Shannon's sense, of a set of n independent communication channels that can be operated in parallel is equal to the sum of the capacities of the individual channels.)

The total capacity C is then just the sum $\sum_a C_a$, where C_a is the maximized value of problem (2.1). I shall suppose that there is an upper bound on the possible capacity C that can be used to evaluate options of this kind, or alternatively that additional processing capacity C can be allocated to this task only at some positive unit cost θ. In either case, the *shadow cost* of allocating additional capacity C_a to the evaluation of attribute a will have to be the same for all attributes, and this is what makes it possible to draw conclusions about the relative amount of attention that will be paid to different attributes by the DM. It is simplest to consider the case in which there is also a common shadow cost of additional processing capacity for other tasks about which the same DM is concerned, and that this cost θ is little affected by variations in the amount of capacity allocated to awareness of the particular kind of options under consideration. In this case, the value of θ can be taken as given in solving for the optimal degree of partial awareness of the options in this particular decision problem.

Finally, I shall suppose that the DM values options according to the sum $u =$

[29] As discussed in the previous section, additional constraints of this kind on the set of possible information structures are necessary in order for there to be any possibility of understanding actual perceptual systems as optimal subject to an information-processing constraint.

$\sum_a x_a$ of their values along the various dimensions. (I assume that the measures of the various attributes have been scaled so that a unit increase in any attribute increases the DM's utility by an equal amount; this is purely a notational convenience.[30]) And I shall assume that the information structure is optimal in the sense of minimizing the mean squared error of the DM's estimate $\hat{u}$ of the value of the option, where $\hat{u}$ is the optimal estimate of u given the subjective representation r.

The mean squared error associated with an estimate $\hat{u}(r)$ can be defined only using a particular prior probability distribution $\pi(x)$ over the possible objective characteristics x of the options that the DM may encounter. Hence optimality can only be *relative* to a particular prior distribution, to which a particular information structure (described by the conditional probabilities $p(r|x)$) may be better or worse adapted. My goal here is to derive implications of the hypothesis of optimal adaptation of the information structure to a particular prior; I shall not seek to provide a theory of how the prior should be determined. A standard assumption in economic models of choice under imperfect information, of course, is that the prior coincides with the *objective* probabilities with which different options may be encountered by the DM. The prior may be interpreted that way here, as well; but the conclusions derived would equally follow in the case of a subjective prior that does not necessarily coincide with any objective probabilities.[31]

The characterization of the optimal estimate is simplest in the case that the prior assumes that the value of each attribute is distributed independently of the others. Then the optimal estimate $\hat{x}_a$ of each attribute is a function only of component r_a of the subjective representation, and the mean squared error of the estimate $\hat{u} = \sum_a \hat{x}_a$

[30] The assumption of additive separability of utility in the different attributes is instead an important restriction, here adopted for the sake of simplicity. The assumption of linearity is not in itself a restriction, as we may suppose that the value of the attribute is subject to a nonlinear transformation that makes utility linear in the transformed value. Linearity only becomes a restriction in conjunction with the further assumption that the prior probability distribution of values for x_a is Gaussian: I am then assuming Gaussianity of the distribution of transformed values, after the transformation required to make utility linear in x_a.

[31] There is a certain amount of experimental evidence indicating that it is possible to manipulate the choices of experimental subjects among options with unchanged objective characteristics, simply by providing additional options (that are not chosen), or directing the subject's attention to particular possibilities in other ways. I would propose to interpret these findings as examples of manipulation of the subject's prior by changes in the experimental setting; these changes do not always involve the provision of objectively relevant information.

is just the sum of the mean squared errors of the component estimates $\hat{x}_a$. We can then solve separately for the optimal information structure $p_a(r_a|x_a)$ for a given attribute, as the one that minimizes the mean squared error of the optimal estimate $\hat{x}_a(r_a)$, given the (marginal) prior distribution π_a over possible values of x_a, for a given processing capacity C_a allocated to the perception of this attribute. The value of C_a depends only on the information structure chosen for the perception of attribute a. If the shadow cost θ of additional capacity is given independently of the total capacity C used to evaluate the option, then it is possible to solve for the optimal information structure for an individual attribute a, without any reference to the information structures chosen for the other attributes.

The optimal information structure for any attribute can be defined as the choice of a set of possible representations R, conditional probabilities $p(r|x)$ for each possible objective state x, and an estimation rule $\hat{x}(r)$[32] so as to minimize the mean squared error

$$\mathrm{E}\left[(x-\hat{x})^2\right], \tag{2.2}$$

subject to an upper bound on the required channel capacity C defined in (1.8). (Here the expectation $\mathrm{E}[\cdot]$ means an integral using the joint distribution for (r, x) implied by the prior $\pi(x)$ for the objective states and the conditional probabilities $p(r|x)$.) The notation used here assumes that the estimate $\hat{x}$ is a deterministic function of the representation r, and one can show that this is optimal. One also easily sees that the identity of the representations R is irrelevant to the solution: it only matters how many different representations belong to the set, and the extent to which the probability distributions over representations associated with different objective states x are similar or different. Hence one can equivalently state the problem in terms of a direct choice of conditional probabilities $p(\hat{x}|x)$ of different estimates $\hat{x}$ of the value of the attribute. In this case, the required channel capacity C is the maximum value of the mutual information between x and $\hat{x}$, where the maximization is again over all possible prior distributions $\{\pi(x)\}$.

I show in the Appendix that this problem is equivalent to the choice of a measure $p_x = p(\cdot|x)$ over the set of possible estimates $\hat{x}$ for each objective state x, and another measure q over that same set of possible estimates that is independent of the objective

[32] Here I omit the subscripts a associated with the particular attribute under consideration. More precisely, one should refer to a set R_a, conditional probabilities $p_a(r_a|x_a)$, and an estimation rule $\hat{x}_a(r_a)$.

state, so as to minimize (2.2) subject to the constraint that

$$D(p_x||q) \leq C \tag{2.3}$$

for each possible state x.[33] Here the expectation $\mathrm{E}[\cdot]$ in (2.2) is an integral using the joint distribution over values $(\hat{x}, x)$ implied by the prior $\pi(x)$ and the conditional probabilities $\{p_x\}$; and for any measures p, q over the set of possible estimates,

$$D(p||q) \equiv \mathrm{E}_p \left[\log \frac{p(\hat{x})}{q(\hat{x})} \right]$$

is the *relative entropy* (or Kullback-Leibler divergence) between the two measures, in which expression $\mathrm{E}_p[\cdot]$ denotes an integral using the measure p over the possible values of $\hat{x}$. The relative entropy is only defined in the case that p_x is *absolutely continuous* with respect to the measure q; this means that p_x cannot assign positive probability to any events that have zero probability under q. Constraint (2.3) thus assumes that the measure p_x has this property for each possible state x.

Here the measure q is not part of the definition of the information structure — that is fully specified by the conditional probabilities $\{p_x\}$ — but only a convenient mathematical representation of the constraint upon possible choices of the $\{p_x\}$. The collection of measures p_x for different objective states x must all be such that they satisfy a bound of the form (2.3) for *some* common measure q. The optimal choice of q in the mathematical problem stated above is the one that relaxes the constraints (2.3) in the way that is most helpful in allowing the mean squared error of the estimate $\hat{x}$ to be reduced.

The relative entropy is a measure of how different the probability distribution p_x is from the "typical" probabilities q of occurrence of the various representations (or estimates).[34] The complexity of the signal that must be sent from one's senses (or more generally, from earlier stages of processing of information about this feature of one's situation) in state x is greater the more different the distribution of subjective representations is required to be in state x than what it typically is. The constraint

[33] More precisely, the problem is to choose a joint distribution of possible values $(\hat{x}, x)$ with marginal distribution corresponding to the prior $\pi(x)$, subject to a constraint that (2.3) hold almost surely (*i.e.,* for values of x that occur with probability 1 under the prior).

[34] Note that the relative entropy is necessarily non-negative, and equal to zero only if the two distributions are identical (almost surely). For further discussion, see *e.g.*, Cover and Thomas (2006).

(2.3) indicates that under this conception of the cost of information processing, the available capacity C constrains how different the distribution of subjective representations associated with *any* particular state can be from the "typical" distribution.

Under the Sims version of RI, instead, the corresponding constraint can be written[35]

$$\mathrm{E}[D(p_x||q)] \leq C, \tag{2.4}$$

where the expectation $\mathrm{E}[\cdot]$ integrates over possible values of x using the prior $\pi(x)$. According to that theory, processing capacity constrains only the *average* complexity of the signal that may be sent about the environment to later stages of processing; sending highly complex signals (to produce subjective representations quite unlike the typical ones) is not a problem as long as the states in which they are sent occur with sufficiently low frequency. Under the theory proposed here, instead, processing capacity limits the degree to which subjective representations can differ in *any* state.

The capacity C_a allocated to the DM's awareness of a given attribute a is itself assumed to represent an optimal allocation of some total processing capacity. Hence in deriving the optimal information structure for a given attribute, C should also be chosen, along with q and the $\{p_x\}$, so as to minimize

$$\mathrm{E}\left[(x - \hat{x})^2\right] + \theta C \tag{2.5}$$

for some unit cost of capacity $\theta > 0$, subject to the constraint that (2.3) hold for all x. The value of θ may either be taken as externally given, or may be solved for so as to imply capacity demands $\{C_a\}$ for the several attributes that sum to some given total capacity available for evaluation of this type of option.

2.2 Optimal Information Structures

Even when x (the value of a particular attribute) is a continuous variable and the prior is described by a density function $\pi(x)$, the solution to the above problem will generally involve only a discrete set of possible subjective representations, and a corresponding discrete set of possible estimates $\{\hat{x}_i\}$, where i indexes the particular

[35] This is equivalent to a requirement that the conditional probabilities $\{p_x\}$ be chosen so that $I \leq C$, where I is the mutual information defined in (1.1). Note that the measure q that minimizes $\mathrm{E}[D(p_x||q)]$ is $p \equiv \mathrm{E}p_x$, and that $\mathrm{E}[D(p_x||p)] = I$.

subjective representation of the attribute. The measure q is correspondingly a discrete measure (that can be specified by a set of probabilities $\{q_i\}$), and the optimal information structure is given by a set of functions $\{p_i(x)\}$, where $p_i(x)$ indicates the probability of observing subjective representation i (and hence of making estimate $\hat{x}_i$) when the objective state is x. We may assume without loss of generality that $q_i > 0$ for all i, since the absolute continuity requirement implies that the function $p_i(x)$ would have to equal zero (almost surely) in the case of any representation i with a zero probability under the measure q. The functions $p_i(x)$ take values between zero and one for all x, and satisfy $\sum_i p_i(x) = 1$ for all x. Figure 4 above presents an example of an optimal information structure, for a case in which the prior $\pi(x)$ is Gaussian, and θ is large enough for the optimal allocation of processing capacity to this attribute to be only one-half a binary digit per observation.[36]

As shown in the Appendix, an optimal information structure is characterized by a system of equations of the following form. For each objective state x, let $\hat{\mathcal{I}}(x)$ be the subset of the discrete set of possible representations $\mathcal{I}$) for which the squared error $|\hat{x}_i - x|^2$ implied by the representation is minimized. (Typically, there will be a *single* minimum-distortion representation, but for certain special states x, there will be multiple equidistant representations.) Then let

$$\hat{q}(x) \equiv \sum_{i \in \hat{\mathcal{I}}(x)} q_i$$

be the total probability of occurrence of representations in the minimum-distortion set for state x, under the measure q. Then for any state x such that

$$\hat{q}(x) \geq e^{-C}, \tag{2.6}$$

it is possible to find a measure p_x over the set $\mathcal{I}$ with a support contained in $\hat{\mathcal{I}}(x)$ that satisfies (2.3). In this case, the constraint (2.3) does not bind, and $p_i(x) = 0$ for any representation $x \notin \hat{\mathcal{I}}(x)$: only minimum-distortion representations ever occur in this state. If $\hat{\mathcal{I}}(x)$ consists of a single representation $\hat{\imath}(x)$,[37] then $p_{\hat{\imath}(x)}(x) = 1$. More generally, p_x will be some measure with support on the discrete set $\hat{\mathcal{I}}(x)$ that satisfies the bound (2.3).[38]

[36] That is, $C_a = 0.5 \log 2$. Given that the standard deviation σ_a is equal to 1 in the figure, this corresponds to a value $\theta = 0.997$, using the result from Table 1 below.

[37] Note that this is the generic case.

[38] Except when (2.6) holds with equality, the solution for p_x is indeterminate in this case.

When (2.6) does not hold,[39] constraint (2.3) binds, and the probability of occurrence of each possible subjective representation i will be given by

$$p_i(x) = \frac{q_i \exp\{-\theta(x)^{-1}|x - \hat{x}_i|^2\}}{\sum_j q_j \exp\{-\theta(x)^{-1}|x - \hat{x}_j|^2\}}, \tag{2.7}$$

where $\theta(x) \geq 0$ is a Lagrange multiplier associated with the constraint (2.3) for that state x. This implies that there is at least some probability of occurrence of *every* possible subjective representation, in the case of each objective state x. However, those representations that lead to a more accurate estimate of the state are relatively more likely to occur.

In particular, if the state x is a single real variable (as in the examples plotted in Figure 4), then (2.7) implies that for any pair of representations i, j, and any two states x, y for which (2.6) does not hold,

$$\log \frac{p_i(x)}{p_j(x)} - \log \frac{p_i(y)}{p_j(y)} = 2(\hat{x}_i - \hat{x}_j) \left[\frac{x - \bar{x}_{i,j}}{\theta(x)} - \frac{y - \bar{x}_{i,j}}{\theta(y)} \right],$$

where $\bar{x}_{i,j} \equiv (\hat{x}_i + \hat{x}_j)/2$ is the mean of the estimates implied by the two representations. If the shadow value of relaxing the constraint (2.3) is the same for both states, then representation i will have a greater relative likelihood of occurrence in the case of state x if and only if the sign of $x - y$ is the same as the sign of $\hat{x}_i - \hat{x}_j$ (a monotone likelihood ratio property). But even if $\hat{x}_i > \hat{x}_j$ while $x < y$, it is possible for representation i to be relatively more likely in the case of state x, if both x and y exceed $\bar{x}_{i,j}$ (so that i is the more accurate representation of either state) and $\theta(y) > \theta(x)$ to a sufficient extent (so that the capacity constraint prevents sharp discriminations from being made as to which subjective state will arise, to a greater extent in the case of state y), or if instead both x and y are less than $\bar{x}_{i,j}$ and $\theta(x) > \theta(y)$ (so that j is more accurate for both states, and less sharp discriminations are made in state x).

Given the state-independent probabilities q_i and the estimates $\hat{x}_i$ associated with the various subjective representations $i \in \mathcal{I}$, the value of $\theta(x)$ can be determined for

[39] Often, the condition does not hold for any x; this is true of the various numerical examples discussed in this paper. Condition (2.6) cannot hold unless the measure q assigns a sufficiently large probability to the small number of representations (generally only one) that are closest to the true state x, while at the same time C must be a sufficiently large positive quantity. But if C is not small, an optimal information structure will allow for many different representations, and the measure q will not assign too much probability to any one of them.

any state x for which (2.6) does not hold, by increasing the assumed value of θ until the measure p_x defined by (2.7) satisfies (2.3) with equality.[40] In this way, assumed values $\{q_i\}$ and $\{\hat{x}_i\}$ give rise to uniquely defined[41] conditional probabilities $\{p_i(x)\}$ and a Lagrange multiplier $\theta(x)$ for each objective state x.

The probabilities q are then given by

$$q_i = \frac{\int p_i(x)\theta(x)\pi(x)dx}{\int \theta(x)\pi(x)dx} \tag{2.8}$$

for each possible representation i. (This expression is well-defined and positive for each i, as long as (2.6) is violated for some set of states with positive measure under the prior $\pi(x)$. If this is not the case, the measure q is undefined, but is not needed to describe the optimal information structure, since (2.7) applies at most to a set of states assigned measure zero under the prior.) Thus the measure q indicates the overall frequency of occurrence of the various subjective representations, if the weights placed on the various possible objective states are given by the prior $\pi(x)$ adjusted by a factor $\theta(x)$ reflecting the value of additional capacity in that state.

Finally, for each subjective representation i, the estimate $\hat{x}_i$ is just the optimal (minimum mean-squared-error) estimate of x when representation i is observed, if a posterior $\pi(x|i)$ is formed using Bayes' Rule. That is,

$$\hat{x}_i = \frac{\int x p_i(x)\pi(x)dx}{\int p_i(x)\pi(x)dx} \tag{2.9}$$

for each possible representation i. (These conditional expectations will be well-defined for each i, as long as i occurs with positive probability under the prior $\pi(x)$.)

The optimal information structure for any capacity constraint C is therefore given by a measure q and estimates $\{\hat{x}_i\}$ that imply conditional expectations $\{p_i(x)\}$ and multipliers $\theta(x)$ for all x in the support of $\pi(x)$, such that (2.8) and (2.9) yield

[40] The geometry of this is illustrated in the Appendix.

[41] This is subject to the qualification that the measure p_x is indeterminate in the case that (2.6) holds as a strict inequality and there is more than one element in the set $\hat{\mathcal{I}}(x)$. Note that the indeterminacy of p_x in this case does not affect the calculation of q in (2.8). Nor does it matter for the calculation of the $\{\hat{x}_i\}$ in (2.9), unless the set of x for which p_x is indeterminate is of positive measure under the prior. In the relatively special case that the prior $\pi(x)$ assigns an atom to a value x that is equidistant from two or more of the estimates $\{\hat{x}_i\}$, then the indeterminacy of the measure p_x is resolved precisely by the requirement that (2.9) yield estimates for the representations $i \in \hat{\mathcal{I}}(x)$ with the property that x is equidistant from them.

precisely this measure and these estimates.[42] Associated with this solution will be a Lagrange multiplier for the capacity constraint C, given by

$$\theta = \int \theta(x)\pi(x)dx. \tag{2.10}$$

If we wish to solve for the optimal information structure for a given attribute a, taking as given not the capacity C_a allocated to processing information about that attribute, but rather the shadow value θ of additional capacity, which is equalized across all attributes under an optimal allocation of attention across attributes, this can be done by varying the assumed value of C_a in the constraints (2.3) until the solution for the multipliers $\{\theta(x)\}$ implies the desired value of θ when substituted into (2.10).

When C is chosen endogenously, the optimal value may or may not be positive. As C is reduced to zero, the implied value of θ may well remain finite. For example, in the case of a Gaussian prior $N(\mu, \sigma^2)$, the upper bound on θ is given by $1.32\sigma^2$, as shown in Table 1 below. This means that there are finite shadow values of processing capacity for which it will be optimal to pay *no attention at all* to a given attribute. The optimal information structure involves measures p_x which are independent of x, and hence convey no information, and as a consequence the optimal action will be independent of the signal received about this attribute. One may then without loss of generality assume that there is a single representation that is the same at all times — which is the same as saying, there is no representation of this attribute at all.

2.3 Effects of the Prior on Discriminations of Value

A key implication of the theory just proposed is that the probability of occurrence of different subjective representations, and hence of different estimates of the value of the various attributes of a given option, depend not only on the objective values of those attributes, but also upon the *prior distribution* for each of the attributes, for which the DM's perceptual system has been optimized. This implies an essential *reference-dependence* of perceptions of value, though the importance of the reference-dependence (in terms of the degree to which it alters the choices that the DM will

[42] The numerical algorithm used to solve these equations in the numerical examples presented in the paper is discussed in the Appendix.

make) will be greatest when the processing capacity allocated to the perception of a particular attribute is small.

Here I illustrate the way in which changes in the prior change the predicted nature of the discriminations that should be made among options that differ in a particular attribute, through examples of numerical solutions of the equations stated above. In these examples, I shall assume that the attribute can be measured by a single real variable x, and I shall assume that the prior $\pi(x)$ is a Gaussian distribution with mean μ and standard deviation $\sigma > 0$. The optimal information structures for priors within this family have an important invariance property. For a fixed value of the bound C on processing capacity, the number of distinct subjective representations in the set $\mathcal{I}$ is independent of μ and σ. Moreover, the conditional probability of occurrence of each subjective representation i is of the form $p_i((x-\mu)/\sigma)$, where the function $p_i(z)$ is independent of μ and σ, when written in terms of the normalized state (i.e., the number of standard deviations by which the state x differs from the prior mean). This invariance is illustrated in Figure 4, where the effect of a shift in μ is simply a horizontal translation of the graphs of each of the functions $\{p_i(x)\}$. The optimal estimate of the state $\hat{x}_i$ associated with a given subjective representation will be of the form

$$\hat{x}_i = \mu + \sigma \hat{z}_i,$$

where $\hat{z}_i$ (the estimated value of the normalized state) will be independent of μ and σ. It follows that the mean subjective valuation (MSV) of this attribute,

$$\mathrm{E}[\hat{x}|x] \equiv \sum_i p_i(x)\hat{x}_i,$$

will be of the form

$$\mathrm{E}[\hat{x}|x] = \mu + \sigma\phi((x-\mu)/\sigma),$$

where the function

$$\phi(z) \equiv \mathrm{E}[\hat{z}|z] \tag{2.11}$$

is independent of both μ and σ, for a given value of C.

The shadow value θ of additional capacity will also be independent of μ, though it does depend on σ; specifically, when C is held fixed, the value of θ will grow in proportion to σ^2. Hence if the cost of capacity θ is given and C is endogenously determined, the optimal capacity C will be a decreasing function of θ/σ^2, or alternatively, an increasing function of σ^2/θ, but independent of μ. This inverse relationship is illustrated for various illustrative values of C in Table 1. Note that the optimal value of C falls to zero, and the function $\phi(z)$ becomes a constant function, equal to zero for all z, for all values of θ/σ^2 above a finite critical value, approximately equal to 1.32.

C (bits)	θ/σ^2	$\phi(1)$
0.01	1.317	0.011
0.25	1.165	0.283
0.50	0.997	0.531
0.75	0.801	0.742
1.00	0.580	0.918
1.25	0.424	1.007
1.50	0.315	1.031
2.50	0.092	0.987
3.50	0.026	0.996
4.50	0.009	0.999

Table 1: The relation between optimal capacity C and the cost of capacity θ, in the case of a Gaussian prior. The value of $\phi(1)$ measures the average slope of the graph of mean subjective value as a function of objective value, over the range of objective values ± 1 standard deviation around the mean.

In order to further illustrate the consequences of variations in the prior for the stochastic subjective representation of value, it may be useful to present additional numerical computations for the case of an information structure with $C = 0.5$ bits (binary digits), shown in Figure 4. In this case, the optimal information structure involves three possible subjective representations, as discussed earlier, and can be completely specified by a single function $p(z)$, in terms of which the the conditional probabilities of the three representations are

$$\begin{aligned} p_1(z) &= p(-z), \\ p_2(z) &= 1 - p(z) - p(-z), \\ p_3(z) &= p(z). \end{aligned}$$

The optimal estimates in the case of each subjective representation are similarly of the form

$$\hat{z}_1 = -\hat{z}, \quad \hat{z}_2 = 0, \hat{z}_3 = \hat{z},$$

for a certain quantity $\hat{z} \approx 0.84$ (Both of these simplifications follow from the symmetry of the problem.) One can further show analytically that the function $p(z)$ satisfies

$$\begin{aligned} p'(z) < 0 \quad &\text{for all} \quad z < -\hat{z}/2, \\ p'(z) > 0 \quad &\text{for all} \quad z > -\hat{z}/2, \end{aligned}$$

as illustrated in Figure 4.

Suppose that there are two options A and B, and that in the attribute currently under consideration, $x^A > x^B$, so that A would be preferred (at least on this dimension) if accurately observed. What is the likelihood of its being viewed as superior on this dimension by a DM with limited processing capacity? The answer depends not only on the value of C, but on the prior to which the DM's perceptual system is adapted.

Let us first consider a prior with respect to which the options A and B are "typical," namely, a Gaussian distribution with $\mu_0 = (x^A + x^B)/2$ and $\sigma_0 = (x^A - x^B)/2$, so that the average value of the attribute for options A and B is equal to the mean value under the prior, and both A and B are an "average" distance from the prior mean (*i.e.,* one standard deviation). With respect to this prior, the normalized values are $z^A = 1, z^B = -1$. When $C = 0.5$ bits, $p(1) = 0.668, p(-1) = 0.037$. Hence A is perceived as superior ($r = 3$) on this attribute with probability 0.668, as ordinary ($r = 2$) with probability 0.295, and as inferior ($r = 1$) with probability 0.037; correspondingly, B is perceived as superior with probability 0.037, as ordinary with probability 0.295, and as inferior with probability 0.667. Let us say that A is perceived as *strongly superior to* B ($A >> B$) if A is perceived as superior *and* B is perceived as inferior, while A is simply *superior to* B ($A > B$) if either A is perceived as superior and B is not, or B is perceived as inferior and A is not. Then with this baseline prior, A is perceived as superior to B with probability 0.840, and as strongly superior with probability 0.446.

These probabilities shift if the prior changes. For a general Gaussian prior $N(\mu, \sigma^2)$, and a capacity of 0.5 bits, the probability that A is perceived as strongly superior to B will equal

$$Pr(A >> B) = p(z^A)p(-z^B) = p\left(\frac{\mu_0 + \sigma_0 - \mu}{\sigma}\right) p\left(\frac{\mu + \sigma_0 - \mu_0}{\sigma}\right)$$

This is plotted in panel (a) of Figure 6, as a function of $(\mu - \mu_0)/\sigma_0$ on the horizontal axis and σ/σ_0 (on a logarithmic scale) on the vertical axis. Note that $Pr(A >> B)$ depends only on these two ratios; the two axes of the figure show the effects of variation in the mean and standard deviation of the prior respectively. The probability that A is perceived as at least somewhat superior to B is also a function only of those two ratios, and this function is plotted in panel (b) of the figure.

One observes from panel (a) that when the prior mean is midway between x^A and x^B, increases in the standard deviation of the prior lower the probability that A will be perceived to be strongly superior. This is because the normalized value z^A falls, as a consequence of which $p(z^A)$ falls; and $p(-z_B) = p(z_A)$ falls for the same reason. If μ increases (so that the prior mean is closer to x^A than to x^B) for a given standard deviation, the probability also falls. In this case, z^A again falls, while $-z^B$ rises by the same amount. Hence $p(z^A)$ falls while $p(-z^B)$ rises, and because $\log p(z)$ is strictly concave in this region, $\log p(z^A)$ falls by more than $\log p(-z^B)$ rises, and $Pr(A >> B)$ falls. The same is true, however (simply switching the roles of z^A and $-z^B$) if μ decreases. Hence $Pr(A >> B)$ is maximized, for a given value of σ, when $\mu = \mu_0$ — that is, when the prior mean is precisely midway between x^A and x^B.

The effects of a change in the prior on $Pr(A > B)$ are less easily summarized. For values $\sigma \geq \sigma_0$ (that is, when the distance between x^A and x^B is no more than two standard deviations), the qualitative effects of variations in μ and σ are the same as those just discussed. However, for lower values of σ (so that x^A and x^B are farther apart, relative to the range of variation expected under the prior), the calculations are more complicated. Consider again the case in which $\mu = \mu_0$. If σ is sufficiently smaller than σ_0, an increase in σ again causes $z^A = -z^B$ to fall, but this can *increase* the probability that A is perceived to be superior, as shown in panel (b). The reason is that when $\sigma_0/\sigma > \hat{z}/2 \approx 0.42$, $p(z)$ is a *decreasing* function of z for z near $z^B = -z^A = -\sigma_0/\sigma$. This implies that as σ rises (while still in this range),

$p(z^B) = p(-z^A)$ falls. This means that both the probability that $r^A > 1$ (so that A is not perceived to be inferior) and the probability that $r^B < 3$ (so that B is not perceived to be superior) increase.

Thus far we have considered the effects of changing μ and σ while holding the upper bound C on processing capacity fixed. But as noted above, if θ is given and C is endogenous, then a change in σ will require a change in C. The required change can be determined by lining up the new value of σ^2/θ with the appropriate row of Table 1 (note that the reciprocal of this quantity is given in the second column of the table), and observing the value of C in that row. We must therefore also consider the effects of such a change in C. An increase in C allows both an increase in the number of distinct subjective representations of this attribute, and a reduction in the degree of randomness of the subjective representations associated with given objective values of the attribute. This allows the estimate $\hat{x}(r)$ to track the true value of the attribute with greater accuracy.

For example, Figure 7 plots the function $\phi(z)$ defined in (2.11), for each of several values of C, again for the case of a Gaussian prior. The function gives the mean normalized subjective value as a function of the true normalized value of the attribute. One sees that in each case, $\phi(z)$ is an increasing, antisymmetric function ($\phi(-z) = -\phi(z)$), which is concave for $z > 0$ and correspondingly convex for $z < 0$. In fact, for any finite positive C, $\phi(z)$ approaches a finite positive asymptote as $z \to +\infty$, and a corresponding finite negative asymptote as $z \to -\infty$. The function necessarily reaches an asymptote because the number of subjective representations in the optimal information structure is finite (for any finite C), and there is accordingly some largest (and some smallest) value $\hat{x}_i$. Hence it is a general prediction of this theory that there should be *diminishing sensitivity* to larger departures of the value of the attribute — in *either* direction — from its mean value under the prior. Eventually, there is no further sensitivity at all to further increases in the magnitude of the departure.

It is worth noting that the diminishing sensitivity result depends on the assumption of a constraint on the required capacity C, rather than on the mutual information I. Under an optimal information structure subject to an upper bound on I (discussed in section 1.3), $\phi(z)$ increases by less than the increase in z, but for each value of the upper bound I, $\phi(z)$ is a *linear* function of z ($\phi(z) = kz$, for some constant $0 < k < 1$ that is increasing in I). Thus the sensitivity of the estimate of the attribute to changes in the actual value of the attribute would be the same over the entire range of possible values of the attribute. In the present theory, instead, $\phi'(z)$ is high for small values of $|z|$ (it may be near or even above 100 percent, if C is not extremely small), but substantially lower for large values of $|z|$, and approaches zero for extreme values of the attribute. This reflects the fact that under the criterion proposed here, it is optimal to concentrate the perceptual system's capacity for discrimination on those parts of the range of possible values that occur with high probability, while under the Sims criterion, there is no reason not to maintain equal discriminability over the entire range.

The point at which the diminishing sensitivity becomes important in the present theory depends on the value of C. The value of $\phi(1)$ reported in the final column of Table 1 indicates the average value of the marginal sensitivity $\phi'(z)$ over the range $-1 \leq z \leq 1$; while this is much less than 1 for values of C equal to 0.75 bits or fewer, it is close to 1 for all values equal to 1.25 bits or higher. (This can also be seen from the plots in the top panel of Figure 7.) On the other hand, the bottom panel of Figure 7 shows that for larger values of $|z|$, the marginal sensitivity remains much less than 1 even for considerably larger values of C. For any given range of values of $|z|$, there is a minimum capacity C such that for any capacity limit of that size or greater, the marginal sensitivity over that range will be fairly close to 1; but the capacity C that is required is larger the wider the range of values of $|z|$ in question.

3 Implications for Choice Behavior

I now discuss some of the implications of the model of inattentive valuation proposed above for observed choice behavior. I shall assume that a DM chooses among available options on the basis of the subjective evaluation of each option, choosing on any given occasion the option that has the highest subjective evaluation at that time. (This does not imply that the same choice will be made each time that a given DM is presented with the same set of choices, since the subjective evaluation of the option on each occasion for choice will be drawn from a certain probability distribution.)

The assumption that choice maximizes subjective value at the time of choice is not the only reasonable one, given our assumption of limited information-processing capacity; one might instead suppose that the *information about the subjective values of the various options* that is used in the decision should also be limited, rather than assuming that a precise ranking of options' subjective values is possible. It is entirely possible to extend the theory in that direction, with the result that the *probability* of choosing a particular option would be predicted to be higher when its subjective evaluation is higher, though even precise knowledge of the current subjective valuation would not suffice to yield a definite prediction about the option that would be chosen on that occasion. For simplicity, however, I shall not pursue such an extension of the theory here; the analysis below assumes that available information-processing capacity is not a quantitatively relevant constraint at the stage at which a final decision is made on the basis of the current subjective evaluations. This is a limiting case of a more general theory that deserves further study.[43] The qualitative conclusions reached below as consequences of assuming non-trivial capacity constraints on earlier-stage processing of information about the individual attributes of individual options should continue to be relevant under that more general theory, even if the precise quantitative implications of the simpler model examined here may require modification.

3.1 Stochasticity of Choice

One immediate respect in which the predictions of this theory differ from standard economic decision theory is that choice is predicted to be *stochastic,* even from the standpoint of an observer who can measure the values of all relevant attributes of the options available to the DM. This is because the theory proposed above implies that there will be a *distribution* of subjective valuations, rather than a single numerical value, associated with each option. This is not simply a theoretical possibility; it is a nearly inevitable feature of the optimal information structure, according to the above theory.

As long as the shadow value of processing capacity θ is positive, then for each attribute i to which the DM pays *any attention at all* (*i.e.,* the information structure

[43] The models of information-constrained choice in Woodford (2008) or Matejka and McKay (2011) instead consider the polar opposite case: precise observations of the characteristics of all of the available options are assumed to be available as inputs to the final decision, but the amount of information that can be used in that decision is subject to a (single, global) capacity constraint. As between the two simplifications of the complete problem, it seems likely that abstraction from the consequences of capacity constraints on earlier-stage processing will make an even greater quantitative difference. Certainly the other type of theory is less consistent with observations such as the focusing effects discussed in section 3.2 below.

involves more than one possible subjective representation r_a of that attribute), the probability distribution of subjective representations will be non-degenerate in the case of at least some objective values x_a, that occur with positive probability (under the prior to which the information structure is optimally adapted). Equation (2.10) implies that if $\theta > 0$, there must be a set of values x with positive measure for which $\theta(x) > 0$; and for any such value x, equation (2.7) then implies that there is a positive probability of occurrence of *each* of the subjective representations r_i that occurs under *any* conditions. The randomness of the evaluation of this attribute then implies randomness of the subjective ranking of options for at least some pairs of options, that are encountered with positive probability according to the prior.

This non-standard prediction of the theory accords with a considerable amount of empirical evidence. The stochasticity of choice in experimental situations has long been noted by psychologists.[44] In addition to the evidence from laboratory experiments on individual decision problems, the hypothesis that choice is a random function of the measured characteristics of goods is a standard specification in econometric models of discrete choice (McFadden, 1974, 2001). The hypothesis that choices are random, with the probabilities of different discrete options varying continuously with the payoffs associated with each option — so-called "quantal response equilibria" — has been argued by game theorists to provide a better explanation of observed behavior in a variety of strategic situations (Goeree *et al.*, 2008; Kets, 2007). And the hypothesis of similarly quantal responses — captured by the econometric hypothesis of a smooth "hazard function" rather than a sharp threshold for adjustment — has been argued to better explain the timing of discrete adjustments by firms in a variety of contexts, including adjustments of prices, of the firm's labor force, and of its capital stock (Caballero and Engel, 1993, 1999; Caballero *et al.*, 1997).

In the economics literature, the preferred interpretation of stochastic observed behavior has been to suppose that choices are *deterministic* functions of an individual's preferences and constraints, but that preferences and/or costs of acting are subject to (unobserved) random variation. Econometric models of discrete consumer decisions

[44] For example, Luce (1959) writes that his "presupposition ... that choice behavior is best described as a probabilistic, not an algebraic phenomenon, ... is by now a commonplace in much of psychology," though "a comparatively new and unproven point of view in utility theory." He also suggests that "economists when pressed will admit that the psychologist's assumption is probably the more accurate, but they have argued that the resulting simplicity warrants an algebraic idealization" (p.2), a view that he seeks to rebut.

are commonly justified in terms of a model of optimal choice in which the utility associated with each good is random (McFadden, 1974); a popular interpretation of "generalized S-s models" of discrete adjustments by firms is that each firm's fixed cost of adjustment is a random draw from a probability distribution, each time the firm again considers whether to adjust (Caballero and Engel, 1999; Dotsey *et al.,* 1999).

While these are possible interpretations of stochastic choice behavior, I believe that there are important advantages of an interpretation in which the randomness results from noise in the DM's perceptions of the objects among which he must choose. First of all, the assumptions about the random variations in the determinants of choice that must be made to justify typical econometric specifications on conventional grounds are often implausible. For example, they typically imply an independent draw of the random preferences and/or random adjustment cost each time a decision is made. But while random variation in these determinants of choice over time is plausible enough, it is hardly obvious that a completely independent draw should occur each time a choice is made (which should be extremely frequently, in a model of the timing of discrete adjustments without information frictions). The independence of the random component over time is instead a natural consequence of a model with noisy observations of the objects of choice, if one supposes that the attributes of the available options must be observed again (through the same capacity-constrained perceptual system) each time a choice is made.

Similarly, the kind of "hazard function" for the occurrence of discrete adjustments that is found to best fit the data is often one in which the probability of adjustment at any time remains bounded away from zero even when the current gains from adjustment (conditional on knowledge of the true state) fall to zero.[45] This is possible under the assumption of choice under full information with a random cost of adjustment, but one must assume that the distribution from which the adjustment costs are independently drawn at each choice point contains an atom of probability mass at a cost of zero, as discussed in Woodford (2008). This is hardly a natural assumption about the nature of random variation in adjustment costs. Under the assumption that the choice situation is imperfectly observed, instead, it is easy to explain why there is

[45] For example, the hazard function that is non-parametrically estimated in Caballero *et al.* (1997) has this property. As discussed in Woodford (2009), the large number of small price changes observed in the size distribution of individual price changes, as documented by Midrigan (2010), can only be explained by a hazard function for price adjustments of this kind.

some probability of adjustment even when the (correctly evaluated) gains from adjustment are zero: the situation is with some probability mistakenly perceived to be one in which adjustment would be worthwhile. This is not simply a possible form of imperfect perception, but a consequence of an optimal information structure under a theory of the cost of more precise information of the kind proposed here: the degree of awareness of the choice situation that would be required to reduce the probability of such mistakes to zero is much too costly (in terms of the required processing capacity) relative to the savings that would result from the elimination of unnecessary adjustments.[46]

The hypothesis of imperfect awareness of the characteristics of the options in the choice set is also a much more specific hypothesis, at least as formulated here, than the hypothesis of random utility and/or random costs of action. While econometric implementations of models of fully-informed choice with random fundamentals often assume some specification of the random fundamentals with only a small number of free parameters to be estimated, there is seldom a clear theoretical justification for adoption of that specification; once one admits the hypothesis that the fundamentals may vary randomly, the possible types of random variation are very many. The hypothesis of random perception, in itself, is similarly subject to a very wide range of possible interpretations and hence has similarly little predictive content. On the other hand, the hypothesis of an *optimal* information structure given the statistics of the set of choices that the DM expects potentially to encounter makes a much more definite prediction. If it turns out that the predictions of a theory developed along these lines are accurate — not simply in predicting that choices will be stochastic, but in predicting the character of the stochastic relationship between objective conditions and the choices that are made — then this will be an important reason to prefer such a theory over the hypothesis of random fundamentals.

[46] This is particularly likely to be the case if the value of adjusting depends on a vector of attributes of the firm's current situation, each of element of which must be separately observed. Then awareness of whether the complete vector is near one of the points at which the value of adjusting is zero will not generally be possible without precise awareness of the value of each of the individual attributes over a large range of possible values.

3.2 Focusing Effects

The theory proposed above also implies that the probability of choosing a particular option from a particular choice set will not depend solely on the true values $u = \sum_a x_a$ of the various options. The entire vector of values of the various attributes x_a, and not simply their sum, is relevant, because the sensitivity of choice to variations in value along the various dimensions will not generally be the same, even though (under my choice of units) the true value of the option is equally sensitive to changes along any dimension. Even if we consider only the MSV of each option,[47] the MSV will be a sum of the form $\sum_a \mu_a(x_a)$, where the functions $\mu_a(\cdot)$ for the various attributes are not generally linear functions with identical slopes for each a. Indeed, even the mean slopes $\mathrm{E}[\mu_a'(x_a)]$ will generally be generally different for different attributes, as illustrated in Table 1 above.

As a particularly extreme example of differential salience of the various attributes of the available options, we have seen that it is possible for some attributes to be completely ignored (in the notation used above, the subjective representation of options along that dimension is the same for all options), even though the available options vary to some extent along this dimension and the attribute is of some relevance to the DM's utility. (In the normal-prior case discussed above, this occurs if and only if $\sigma_a^2 \leq 0.75\ \theta$ for a particular attribute.) More generally, Table 1 shows that decisions are predicted to be more sensitive to variation in a particular attribute when the set of potential options (according to the prior) includes a greater range of variation in the contribution to utility from this attribute.

"Focusing effects" of this kind are frequently observed. For example, Schade and Kahneman (1998) document a "focusing illusion" in judgments of life satisfaction under alternative circumstances: they find that subjects judging the quality of life of others living in California as opposed to the Midwest exaggerate the extent to which people are happier in California, relative to subjects' reports about their own life satisfaction in the two locations, because of an excessive focus on the difference in climate. Similarly, Kahneman *et al.* (2006) argue that people exaggerate the degree to which greater wealth would contribute to their degree of happiness, both in life choices and in certain kinds of survey questions, relative to the degree of effect that

[47] Under the choice rule proposed here, of course, the probability of choosing one option over another is not solely a function of the means of the two distributions of subjective valuations.

it actually has on experienced happiness, as measured by surveys that require people to report their feelings in real time.

Koszegi and Szeidl (2011) point out that these and many other focusing effects reported in the literature can be explained by a general hypothesis according to which people place excessive weight on attributes in which their available options *differ more.*[48] Thus the weight placed on climate in judgments about hypothetical life satisfaction in different locations is excessive because the options do obviously differ significantly along this dimension; the fact that they also differ (to a smaller extent) on many other dimensions that, cumulatively, result in people's being no happier on average in California is taken insufficiently into account. Similarly, people may choose a higher-paying job despite other features that reduce their happiness (such as longer commuting time, longer hours, and so on), because the difference between the possible job options is not as great along any of the other individual dimensions. This type of theory can simultaneously explain an apparent "present bias" in some decisions (such as whether to exercise on an individual day), where a large difference in present enjoyment must be balanced against smaller consequences for satisfaction on many future days, and an apparent bias toward excessively future-oriented choices in some other contexts (such as career decisions), where small costs over many nearer-term dates must be balanced against a single large benefit that may be far in the future. The same hypothesis offers an explanation for why firms can apparently increase demand by splitting the prices of their products into multiple components that consumers treat as separate attributes (Morwitz, Greenleaf, and Johnson, 1998).

The theory proposed here predicts the existence of focusing effects of these kinds. As shown in Table 1, an optimal information structure will increase the average sensitivity of MSV to a particular attribute when the expected range of variation from the utility contribution from that attribute is greater. Suppose that the prior distribution of values for each attribute a is a normal distribution $N(\mu_a, \sigma_a^2)$. Then the optimal information structure for each attribute will be of the kind computed in

[48] Bordalo *et al.* (2010, 2011) and Bordalo (2010) propose a more complex model of the determinants of the relative "salience" of particular attributes of the options among which one must choose, but one of their key ideas is similar: they argue that the salience of a particular attribute (payoffs in a particular state of the world, in the case of choice between lotteries) is increasing in the degree to which the options differ in that attribute.

the numerical results described in section 2, but with a different amount of processing capacity C allocated to the awareness of each attribute. If the total available capacity is optimally allocated across the attributes, then the value of θ will be the same for each a, and the value of $\sigma/\theta^{1/2}$ for each a will be proportional to σ_a. Thus the values of C for the different attributes will be ordered in the same way as the values of σ_a are ordered; and as shown in the table, a higher value of C means a higher average sensitivity of MSV to the objective value of that attribute.

More precisely, the theory proposed here implies that the MSV of an option specified by a vector of attributes x will be given by

$$\sum_a [\mu_a + \sigma_a \phi(z_a; \sigma_a/\theta^{1/2})],$$

where $\phi(z; \sigma/\theta^{1/2})$ is a function defined in section 2.3, and $z_a \equiv (x_a - \mu_a)/\sigma_a$ is the normalized value of attribute a. Hence the partial derivative of the MSV with respect to increases in the value x_a of any attribute is equal to $\phi(z_a; \sigma_a/theta^{1/2})$. Because of the properties of the $\phi(z; \sigma)$ function shown in Figure 7, we see that this derivative is decreasing in the absolute value $|z_a|$ for any given value of σ_a, and increasing in σ_a for any given value of z_a.

More specifically, suppose that a DM must choose between two options x^1 and x^2, and that the DM's perceptual mechanism is furthermore adapted to a prior for each attribute with mean and variance equal to the mean and variance of the sample consisting of these two options only, so that $\mu_a = (x_a^1 + x_a^2)/2, \sigma_a = |x_a^1 - x_a^2|/2$. Then x^2 will be valued more than x^1 on average if and only if

$$\sum_a \lambda_a (x_a^2 - x_a^1) > 0, \tag{3.1}$$

where for each a, $\lambda_a \equiv \phi(1; \sigma_a/\theta^{1/2})$ is precisely the quantity reported in column 3 of Table 1, when $\sigma_a/\theta^{1/2}$ is equal to the square root of the reciprocal of the value reported in column 2. Thus in this case (for a given cost of processing capacity $\theta > 0$), the table can be used to directly read off the weights λ_a that indicate the degree of distortion involved in the comparison of MSVs rather than true values.

The predicted focusing effect can easily lead to preference reversals. Suppose that option 2 is substantially superior in one attribute ($x_0^2 - x_0^1 = b > 0$), but inferior to the same small extent ($x_a^2 - x_a^1 = -c < 0$) for each of a large number of other attributes,

$a = 1, \ldots, N$. If $b > 1.74\ \theta^{1/2} > c$,[49] then it is optimal for the DM's perceptual system to *ignore attributes 1 through N altogether* ($\lambda_a = 0$ for $1 \leq a \leq N$), while $\lambda_0 > 0$. It follows that option 2 will be valued more on average, and will be chosen more often, because the subjective representation of the options pays attention only to the attribute in which option 2 is substantially superior.[50] On the other hand, if $Nc > b$, option 1 is *objectively* preferable to option 2; that is, the DM would prefer it if enough attention were paid to the evaluation of the two options.

The most important difference between the theory proposed here and the one proposed by Koszegi and Szeidl (2011) is that in their theory, the sensitivity to a particular attribute is an increasing function of the range of variation in the value of that attribute across the options in the DM's *current choice set,*[51] whereas here it depends on the range of variation in that attribute *under the prior* to which the information structure has been adapted.[52] Of course, it makes sense to suppose that the amount of variation present in the choice set will have some relation to the degree

[49] This condition depends on the value of θ. One can show, however, that θ necessarily falls in the range where these bounds are satisfied, if and only if the total capacity C allocated to perception of these $N+1$ attributes is positive, but less than the value of C corresponding to a value of $1.32(c/b)^2$ in the second column of Table 1. For example, if b is 7.1 times as large as c, then it is optimal to allocate all processing capacity to the perception of attribute $a = 0$ as long as C is no greater than 3.5 binary digits.

[50] Note that even in this case, the model predicts only that option 2 will be chosen *more often,* not that it will always be chosen. While option 2 is *objectively* superior with regard to attribute 0, the subjective representation of this attribute of each of the options will be stochastic, and there will remain a positive probability (less than 1/2) of option 1 being perceived to be superior to option 2 in this attribute. If the difference b is sufficiently great, this probability may be small.

[51] To be more precise, Koszegi and Szeidl assume that what is relevant is the set of options in a DM's "consideration set," which may be only a subset of the full choice set, for reasons that are exogenous to their theory of focus-dependent choice. But in any event, only the options currently under consideration are assumed to determine the weight given to a particular attribute.

[52] There are a number of other technical differences, though these are less obviously crucial to the qualitative predictions made about focusing effects. In Koszegi and Szeidl (2011), the subjective valuation of a particular attribute is a deterministic linear function of the true utility contribution of that attribute of the option, whereas in the present theory the relation is nonlinear and stochastic. Also, in the theory of Koszegi and Szeidl, it is only the difference between the maximum and minimum values of the attribute that are represented in the choice set that matters for the weight placed on that attribute, whereas other features of the prior distribution also matter in the present theory (which does not even assume that bounded minimum and maximum values necessarily exist or differ across attributes).

of variation in that attribute to which the DM's perceptions have adapted; indeed, one might reasonably suppose that marketers should have some ability to influence the perceptual discriminations of their customers by varying the set of choices with which they are presented (as is in the case of the well-documented "decoy effects" discussed below). The assumption that the distribution of possibilities to which the perceptual system has adapted is always precisely the set of options in the current choice set would be one possible specification of the present model.

But while this would allow the model to make extremely sharp predictions (an obvious appeal of the assumption of Koszegi and Szeidl), I find it implausible that adaptation to changing choice situations should always be that rapid. I would suggest instead that the prior should reflect experience with some *class* of decision situations, to which the current choice set belongs,[53] but it hardly seems efficient (given information-processing constraints) for each individual choice situation that is ever encountered to be treated as its own category with its own, optimally tailored information structure. Nor is it obvious that it is only one's actual choices on different occasions that contribute to the distribution of possibilities in the prior. It is well known from experiments that choices from among a particular set of options can be influenced by getting the subject to *think* about something, even if it is not part of the choice set.[54] This suggests that it may be most realistic to assume that the prior distribution reflects the frequency with which one has considered options with this particular attribute, including reasons for consideration of the option when one may not have actually had an opportunity to choose it. Such an assumption about the prior would be consistent with the interpretation proposed in section 1.3 for adaptation phenomena in perceptual experiments.[55]

Gabaix (2010) proposes a theory of focusing effects that is even more closely re-

[53] More precisely, one should say, to which the current choice situation is *perceived* to belong — as this classification should itself be subject to an information-processing capacity constraint, and hence be only stochastically related to the actual characteristics of the current choice situation.

[54] The "anchoring effects" documented by Tversky and Kahneman (1974) are a celebrated example.

[55] Kahneman (1992) draws an analogy between anchoring effects in choice behavior and the observation in psychophysics experiments that judgments about the intensity of a stimulus are influenced by other stimuli to which the subject has recently been exposed, even if not asked to express a judgment about them. The explanation proposed here would be the same in both cases: merely having paid attention to something to something that possesses the attribute in question influences the prior distribution of values for that attribute to which a subject's perceptions of that attribute are then adapted.

lated to the one proposed here. In Gabaix's theory, like that of Koszegi and Szeidl, decisions are based on a subjective valuation of the form $\sum_a \lambda_a x_a$, where the weights λ_a may differ from 1, the weight in the DM's true utility function. But Gabaix proposes that the weights $\{\lambda_a\}$ are chosen so as to minimize the expected discrepancy between true and subjective valuations, under a prior distribution for the values of the attributes that will be encountered, subject to an upper bound on $\sum_a |\lambda_a|$. The only important difference between Gabaix's theory and the one proposed here is the cost function (or constraint) that is assumed for possible mappings of the vector of objective characteristics into subjective representations. Unlike the theory proposed here, Gabaix assumes (at least in the leading example of his theory) that the subjective representation must be a *deterministic, linear* function of the vector x, and proposes a cost that depends on the coefficients of the linear function. The theory proposed here is much less restrictive as to the relations between objective characteristics and subjective representations that are considered, and indeed under the cost function that I propose, the optimal relation is neither linear nor deterministic, even in the quadratic-Gaussian case.

The particular cost function proposed by Gabaix is not derived from an underlying theory of information-processing constraints, but is chosen to reflect the idea that more "sparse" representations should be preferable. The theory proposed here does not directly assume that sparsity is a goal in itself, and the optimal information structure characterized above can easily imply that some attention will be paid to all attributes. (This will typically be the case when there are not extreme differences in the contributions of the different attributes to the overall degree of variation in utility across possible options.) Nonetheless, the cost function proposed here also implies the possibility of corner solutions, in which some attributes will be ignored altogether though they contribute non-negligibly to variations in true utility across options. (One such example has just been discussed.) It is surely true that sometimes aspects of a choice situation are ignored altogether even when they are of some modest relevance to the DM's objective; but it is not clear that this occurs more often than can be accounted for by the kind of theory proposed here, or as pervasively as Gabaix's theory would imply. And Gabaix's theory gives no explanation for the stochasticity of choice, or to phenomena connected with "diminishing sensitivity" to further increases in a particular attribute, of the kind discussed below.

3.3 Context Effects

Another difference between the theory proposed here and standard rational choice theory is that choice that will be made between particular options are not predicted to depend solely on the attributes of those options. Not only is the subjective representation of each option different from the objective vector of attributes, but it is not predicted to be a function solely of that vector of attributes, either. In the theory proposed here, the DM's perceptual system is optimized for a particular prior distribution for each of the attributes of potential options; hence the probability of choosing one option over another depends on what that (joint) prior distribution is like. Presumably it is derived from experience (real or imagined) with particular options that involve particular values of these attributes; this means that the probability that the DM will choose a given option over another can vary depending on the context in which the choice between these options is presented to her. "History and experience" become important determinants of choice, as McFadden (1999) comments in the passage quoted earlier.

In fact, context effects are often noted in the experimental psychology and behavioral economics literatures. An important class of context effects are *choice-set effects,* in which the probability of choosing one option over another is affected by which *other* options are also in the DM's choice set. An especially well-documented example is the "decoy effect," often employed by marketers to manipulate consumer behavior. Suppose that consumers may choose between two competing brands, and that two brands differ in two different attributes, let us say quality and price. The location of the two brands in this two-dimensional space are represented by points A and B in Figure 8. Here the horizontal axis represents quality (increasing from left to right) and the vertical axis represents price (increasing from top to bottom, so that higher points are preferable on this dimension as well). In the case indicated in the figure, neither brand dominates the other: one (brand B) is of higher quality, but is also more expensive. It is well known to marketers that sales of brand B can be increased by introducing a third option (the "decoy") with characteristics at a point such as C in the figure. The decoy is not intended to attract purchasers, as it is dominated by brand B, though not by brand A. Nonetheless, a large number of studies have confirmed that introduction of an "asymmetrically dominated" decoy of this kind increases the number of buyers who will choose brand B and reduce the number who choose brand A (Huber *et al.*, 1982; Heath and Chatterjee, 1995).

Such results are inconsistent with the standard theory of rational choice; in particular, they violate the usual axiom of "independence of irrelevant alternatives," according to which choice should be unaffected by the removal from the choice set of options that are not selected even when included. They may, however, be consistent with a model of inattentive valuation of the kind proposed here, if the addition of the decoy changes the probability distributions for the various attributes to which the DM's perceptual system is adapted.

Suppose, for simplicity, that the probability distribution for each attribute to which the perceptual system is adapted is a normal distribution with the same mean and variance as the distribution of values of that attribute in the DM's choice set. (The assumption that the prior distribution is Gaussian allows us to use the results from section 2 for numerical solution of the Gaussian case.) Then when A and B are the only two brands available, the prior distribution is characterized by $\mu_a = (x_a^A + x_a^B)/2$ and $\sigma_a = |x_a^B - x_a^A|/2$ for each attribute. In this case, the normalized values of the two goods are $+1$ and -1 for each attribute, though good B has the greater value ($+1$) of the quality attribute, while good A is superior (a normalized value of $+1$) with regard to price.

When the decoy C is added to the price set, the distribution of values of each attribute now consists of $\{x_a^A, x_a^B, x_a^C\}$, where on the price dimension $x_a^C < x_a^B < x_a^A$, while on the quality dimension $x_a^A < x_a^C < x_a^B$. This means that the mean value μ_a of both attributes falls, but the mean decreases by *more* (as a fraction of the difference between B and A on that attribute) in the case of the price attribute (*i.e.*, the attribute for which the addition of the decoy extends the range of values reflected in the distribution). Furthermore, the standard deviation σ_a of the quality attribute decreases, but the standard deviation of the price attribute does not decrease as much (in percentage terms), and may even increase.[56]

If we suppose that the capacity C_a allocated to perception of each of these attributes remains the same after the addition of the decoy (though perceptual coding is re-optimized for the new distribution of values for each attribute, subject to the capacity constraint), and that that capacity happens to equal one-half of one binary

[56] A little algebra shows that the standard deviation of the price attribute increases if and only if x_a^C is enough less than x_a^B for the distance of x_a^C from the midpoint between x_a^B and x_a^A to be more than $\sqrt{3/2}$ times the distance of x_a^B from the midpoint.

digit, then we have constructed an example in which the consequences of introduction of the decoy can be read off from Figure 6. Note that under the initial distribution of values for the attributes, brands A and B are located at normalized values z_a equal to +1 and -1, corresponding to the situation represented by the point labeled I in Figure 6. The consequences of changes in the distributions after introduction of brand C for the relative valuation of A and B are then shown in the two panels of the figure.

In these plots, the new distribution for the quality attribute will necessarily be somewhere to the southeast of the point labeled I,[57] while the new distribution for the price attribute will be to the left of I — and even farther to the left of I than the new quality distribution is to the right of I — and higher on the vertical axis than the point representing the new quality distribution.

In panel (a) of Figure 6, we see that a simultaneous movement to a higher value of σ and a greater horizontal distance from the $\mu = \mu_0$ axis always reduces the probability that the superior good on a given dimension will be regarded as *strongly* superior. This means that, while in the absence of the decoy, there should be an equal probability of recognition of the low-price good as strongly superior on the price dimension and recognition of the high-quality good as strongly superior on the quality dimension, with the decoy, there should be a smaller probability of recognition of the low-price good as strongly superior in price than the probability of recognizing the high-quality good as strongly superior in quality. *Both* the fact that the prior is more diffuse with respect to the price attribute (after introduction of a high-price decoy), so that brands A and B are not so many standard deviations apart on this dimension, and the fact that the prior mean is shifted farther in the case of the price attribute, making it particularly unlikely that brand B will be judged extreme on this dimension, contribute to this conclusion.

In panel (b) of Figure 6, we can similarly observe the consequences for the probability of judging one brand to be at least somewhat superior on a given attribute. Here movement farther from the $\mu = \mu_0$ axis again always reduces the probability of recognition of the superior brand as superior, but the effect of increasing σ does not always have the same sign. Nonetheless, one observes that it is possible for the net effect of the two changes to lower the probability of recognition of the superior brand

[57] Here I assume that while brand C is inferior to B in quality, it is nonetheless located to the right of the midpoint of goods A and B on the quality dimension, so that the mean rises for this attribute when brand C is introduced.

as superior. Hence it can easily be the case that introduction of the decoy results in there being a smaller probability of recognition of the low-price good as superior in price than of recognizing the high-quality good as superior in quality.

In the above discussion, it has been assumed that equal processing capacity is allocated to the perception of the two attributes; but this may not be realistic. In their meta-analysis of studies of decoy effects, Heath and Chatterjee (1995) find a greater ability of decoys to increase the market share of high-quality/high-price brands (the case shown in Figure 8) than to increase the market share of low-price/low-quality brands; but no such difference would be predicted by the model just sketched, which is perfectly antisymmetric between brands A and B.[58] Intriguingly, Heath and Chatterjee also find in their own experiments that decoys increase the market share of a high-quality/high-price target brand more, in the case of a "traditional" subject population (MBA students at an urban, semi-private research-oriented university),[59] but that decoys increase the market share of a low-price/low-quality target more in the case of a "non-traditional" subject population (undergraduate students at a rural, teaching-oriented state university). One might expect the former subjects to pay more attention to differentiations in quality when choosing purchases, and less attention to price, while the latter subjects pay more attention to price and less to quality.[60] In this case, the studies suggest that decoys are more effective in reducing the attention paid to an attribute to which the subject is *already relatively insensitive* than they are in reducing the attention paid to an attribute with which the subject is highly concerned.

The theory proposed here predicts such an asymmetry, if we do not assume that

[58] The situation described is unchanged if one interchanges brands A and B and at the same time interchanges the "price" and "quality" attributes.

[59] Here "traditional" means typical of the subjects used in other marketing studies, such as those reviewed in their meta-analysis.

[60] For example, Heath and Chatterjee report that the "non-traditional" subjects are significantly more likely to agree with the statements "I don't care what others think about what I buy," "I buy less expensive brands," and "I believe store brands are as good as national brands," and significantly less likely to agree that "I purchase products for their status value." Curiously, they do not find any significant difference in the two populations' responses to statements that directly ask about the importance of "quality" as opposed to saving money. It seems likely that while the "non-traditional" population does pay more attention to cost, they prefer not to think of themselves as sacrificing quality, but instead challenge whether conventional markers of quality represent higher quality.

equal processing capacity is allocated to perception of the two attributes. Suppose, for simplicity, that the amount of capacity allocated to perception of quality is large enough that we can (as an approximation) assume that the quality of each brand is accurately observed, while there is (as above) only 0.5 bits of capacity allocated to the perception of price. (This could be consistent with the theory of attention allocation proposed in the previous section, if $\sigma_{quality} >> \sigma_{price}$. Note that here I am assuming that the subject's perceptual system is adapted to a quality distribution that is not calibrated on the basis of the range of qualities in the current choice set.) Then brand B will be preferred to brand A if and only if

$$\hat{y}^A - \hat{y}^B < x^B - x^A, \tag{3.2}$$

where y and x are used to refer to the vertical and horizontal axes in Figure 8 (*i.e.*, to the price and quality attributes respectively), and $\hat{y}^b$ is the estimate of brand b's value on the price attribute, given the subjective representation of that attribute for that brand.

If we further suppose that the true price difference between the two brands exceeds their true quality difference, by a factor

$$x^B - x^A \approx 0.6(y^A - y^B),$$

then there will be a range of values of σ_{price} around $\sigma_0 \equiv (y^A - y^B)/2$ such that for any prior in that range, (3.2) will hold as long as A is not perceived to be strongly superior to B on the price dimension, but will fail if it is. If A is perceived to be superior, but not strongly superior to B, $\hat{y}_A - \hat{y}_B = 0.84\sigma_{price}$, but $0.84\sigma_0 = 0.42(y^A - y^B) < x^B - x^A$, so that (3.2) holds as long as σ is not too much greater than σ_0. If instead A is perceived to be strongly superior to B, $\hat{y}_A - \hat{y}^B$ is twice as large, and then since $1.68\sigma_0 = 0.84(y^A - y^B) > x^B - x^A$, (3.2) will not hold, unless σ_{price} is substantially less than σ_0.

In this case, the probability that brand A will be preferred to brand B is equal to $Pr(A >> B)$, the quantity plotted in panel (a) of Figure 6. If we suppose that perceptions of price are adapted to a prior that is calibrated to the mean and standard deviation of the prices of the brands in the current choice set, then the prior when only brands A and B are offered will correspond to point I in the figure. The model then predicts that brand A should be chosen nearly 45 percent of the time. However, the introduction of the asymmetrically dominated decoy C should lower μ_{price}, while it may or may not raise σ_{price}. It can be observed from Figure 6(a) that the reduction of μ should lower $Pr(A >> B)$, and hence lower the probability that brand A will be preferred to brand B. If there is no reduction in σ_{price}, or σ_{price} actually increases, then one can say unambiguously that the probability of preferring A to B must fall. At the same time, $Pr(A >> B)$ will still fall in the case of a sufficiently modest decline in σ_{price}. Hence if the degree to which C is more expensive than B is great enough, the model clearly predicts a reduction in the probability of preferring brand A to brand B.[61]

Of course, Figure 6 is computed under the assumption that the processing capacity C allocated to the particular attribute remains unchanged despite variations in μ and σ. In the previous section, we have discussed the fact that an increase in σ_a for some attribute a should increase C_a, in order to maintain a common value of θ across attributes. If the reallocation of processing capacity across different perceptual tasks occurs more slowly than the adaptation of the coding scheme to make more efficient use of the existing processing capacity for a given attribute, then the analysis above suffices for the short-run effect of the introduction of the decoy brand. If instead we suppose that the reallocation of processing capacity occurs equally quickly, then an increase in σ_{price} should increase C_{price} (as discussed in the previous section), and this should tend to counteract the reasons for a reduction in the perceived price advantage of brand A discussed above. However, even in the latter case, it remains possible for the introduction of the decoy to lower $Pr(A >> B)$. For example, in the case that brand C is not enough more expensive than brand B to actually raise σ_{price}, there would be no reason for C_{price} to increase, and the effects are as discussed above.

Thus the asymmetric model explains in a fairly simple way why the introduction of a decoy can shift purchases from a low-cost/low-quality brand to a high-quality/high-cost brand. The same model predicts that the introduction of a decoy that is asymmetrically dominated by the low-cost brand should not be equally effective at shifting purchases to that brand. If $C_{quality}$ is large, one cannot expect changes in the prior distribution for the quality attribute to make any material difference for the perceived

[61] Heath and Chatterjee also report a "range effect," according to which an increase in the range of prices in the choice set (by introducing a brand C that is more expensive even than B) results in a greater reduction in the market share of A than in the case that C is only dominated by B on the quality dimension. The desirability of increasing the range of prices, so as to ensure an increase in σ_{price}, can be explained by the analysis presented here.

difference in quality between the two brands; the effects of the prior on perceptions are only significant when C is small. Hence the introduction of an unusually low-quality decoy cannot be expected to reduce the salience of the quality differential between A and B. The decoy would only be effective at increasing the market share of brand A if it were able to increase the probability that A is perceived to be strongly superior on the price attribute (*i.e.*, the attribute on which it is superior). But Figure 6(a) shows that adding an additional inexpensive option, and raising μ_{price} above μ_0, will only *lower* the probability that A is perceived as strongly superior. The fact that the introduction of a new brand that is not as inexpensive as B necessarily lowers σ_{price} should be helpful, but will not necessarily outweigh the effect of the increase in μ_{price}. Furthermore, to the extent that a reduction in σ_{price} results in a reduction in the processing capacity allocated to the perception of price, the introduction of the decoy is even more counter-productive. Hence it should not be surprising if decoys are less effective in this case.

Of course, the roles of "price" and "quality" in the analysis above are reversed in the case of subjects for whom $C_{price} >> C_{quality}$. Hence the theory also provides a potential explanation for the finding by Heath and Chatterjee that in the case of their "non-traditional" subject pool, decoys are more effective in shifting purchases from a high-quality/high-price brand toward a low-price/low-quality brand than in the opposite direction.

Bordalo (2010) and Bordalo *et al.* (2011) similarly propose to explain decoy effects in terms of a shift in the relative salience of the brands' differences along different dimensions. In their theory, each choice option x^j has a subjective valuation $\sum_a \lambda_a^j x_a^j$, where the weight λ_a^j placed on attribute a in evaluating option j depends on the "salience ranking" of that attribute for that option. The salience of attribute a for option j depends on the degree of contrast between the value x_a^j of this attribute for option j and the distribution of values of attribute a across the entire choice set;[62] the weights $\{lambda_a^j\}$ are then determined by an ordinal ranking of the saliences of the various attributes of option j. Note that in this theory, unlike those of Koszegi and

[62] The precise mathematical definition of salience is slightly different between Bordalo (2010) and Bordalo *et al.* (2011). In the first paper, it is a sum of pairwise contrasts between option j and each of the other options in the choice set, whereas in the second it is simply a measure of the contrast between x_a^j and the mean value of attribute a over the elements of the choice set. These two definitions do not lead to equivalent rankings because "contrast" is not a linear function of the two quantities being compared.

Szeidl (2011) or Gabaix (2011) discussed above, the weights $\{lambda_a^j\}$ are specific to each option j, rather than being the same for all options, because the relative salience of the various attributes can differ across options.

The Bordalo *et al.* definition of "salience" as an increasing function of contrast implies that the sensitivity of the subjective valuation of option j to variation in the value of a given attribute a depends on the degree of contrast between the value x_a^j and a distribution of values for that attribute. The theory proposed here is similar, insofar as the probability of a given subjective valuation of an option along a particular dimension is predicted to be a function of where the value x_a^j falls within a prior probability distribution for attribute a. There are, however, some important differences between the present theory and theirs. In their theory, subjective valuation is a deterministic function of the objective characteristics of the options in the choice set. This yields simpler and sharper predictions, but fails to account for the stochasticity of observed choice. In their theory, salience is also a function purely of the attributes of the options in the choice set, rather than of "prior distributions" for the various attributes that may derive from other experience. This makes the theory's predictions in a given application much less ambiguous; but it means, for example, that their theory offers no explanation for the asymmetry between the use of a decoy to increase demand for a high-quality/high-price good and the use of a decoy to increase demand for a low-price/low-quality good. I believe that it will instead have to be accepted that "history and experience" matter for choice, as McFadden proposes, despite the difficulties that this implies for clean hypothesis testing.

3.4 Reference-Dependent Valuations

One of the most celebrated findings of Kahneman and Tversky (1979) is that when choosing among gambles, people appear to evaluate *gains and losses* from individual choices rather than the different *final situations* that they may reach after a sequence of events. This notion is, of course, quite contrary to a basic postulate of the standard theory of expected utility maximization. Kahneman (2003, 2011) calls this postulate "Bernoulli's error." "Bernoulli's model is flawed," he writes, "because it is *reference-independent:* it assumes that the value that is assigned to a given state of wealth does not vary with the decision maker's initial state of wealth" (Kahneman, 2003, p. 460).

For example, Kahneman and Tversky present different groups of experimental subjects with the following two choices:

Problem 1 *In addition to whatever you own, you have been given 1000. You are now asked to choose between (a) winning an additional 500 with certainty, or (b) a gamble with a 50 percent chance of winning 1000 and a 50 percent chance of winning nothing.*

Problem 2 *In addition to whatever you own, you have been given 2000. You are now asked to choose between (a) losing 500 with certainty, and (b) a gamble with a 50 percent chance of losing 1000 and a 50 percent chance of losing nothing.*

They report that substantial majorities of their subjects choose the sure thing in Problem 1, and the gamble in Problem 2. Yet in each of the two problems, the choice is between *identical* probability distributions over possible final wealth states: option (a) yields one's initial wealth plus 1500 with certainty, while option (b) yields one's initial wealth plus 1000 with a 50 percent probability and one's initial wealth plus 2000 the other 50 percent of the time. Thus the evaluation of these options is evidently not merely a function of the probabilities assigned to different final wealth states. Kahneman and Tversky propose instead that in each problem, the different possible final wealths are evaluated relative to a *reference point* corresponding to the wealth possessed prior to the decision; it is the fact that the reference point is higher by 1000 in Problem 2 that results in a different evaluation of the relative attractiveness of the two lotteries.

The theory of inattentive valuation proposed here provides an explanation for such findings, that remains a variant of rational choice theory, and that still include the standard (von Neumann-Morgenstern) theory of choice over lotteries as a limiting case (the case in which the processing capacity allocated to the evaluation of one's options is large enough). Suppose that, as in Bordalo *et al.* (2010), we treat each of the possible outcomes for a lottery as separate attributes that must be evaluated. In the above example, we may suppose that each of the four options (two in each problem) has two attributes: the final wealth achieved in each of two states, that occur with equal probability, and so receive equal weight in the DM's utility function. (In the case of a sure thing, we assign the option the same value of both attributes, as the outcome is the same in both states.) Choice between two options will then depend on the relative size of the subjective estimate $\sum_{s=1}^{2} \hat{x}_s$ for the two options, where for each option, $\hat{x}_s$ is the subjective estimate of the value of final wealth in state s under that option. The estimate $\hat{x}_s$ for each state will be a function of the subjective representation r_s of the value of final wealth in that state.

According to the theory presented above, the conditional distribution $p(r_s|x_s)$ over subjective representations in the case of a given objective final wealth x_s is adapted to a particular distribution of possible final wealths associated with choice situation of that kind. The distribution to which the DM's perceptions of value are adapted may well be different in situations like Problem 1 than in situations like Problem 2; hence the subjective perceptions of options (a) and (b) may be different in the two problems, even though the probability distributions over final wealth in the two options are the same in both problems. This is the way in which valuations are predicted to be *reference-dependent* in the present theory.

For example, suppose that the DM recognizes the class of choice situations in which "you have been given 1000, and now are asked to choose between lotteries" as different from the class of situations in which "you have been given 2000, and now are asked to choose between lotteries," and so perceives the options presented using a perceptual system that has been optimally adapted to different distributions of potential outcomes in the two cases. If the DM has no reason to expect the types of gains and losses that may be offered by the lotteries to depend on her wealth at the time that the choice is presented, then the prior distribution over possible levels of final wealth should indeed be different between the two classes of situations: the entire probability distribution should be shifted up by 1000 in the case of the second class. If the distribution of possible net gains from a lottery payoff in either state of the world is assumed to be a Gaussian distribution, then for either state of the world, in both classes of choice situations the prior distribution over possible final wealths should be a Gaussian distribution $N(\mu, \sigma^2)$, and the standard deviation σ of this distribution should be the same for both classes of situations, but the mean μ should be higher by 1000 for the second class. Since the standard deviation σ is the same for both classes of situations, then — supposing that the ex ante probability of encountering choice situations in the two classes is also expected to be equal — it makes sense for the same processing capacity C to be allocated to perceptions of the value of outcomes in each of the two cases. Then if the optimal information structure for situations in class 1 involves conditional probabilities $\{p^1(r_s|x_s)\}$, for some set of possible subjective representations $\{r_s\}$ and values of x_s on the real line, then the optimal information structure for situations in class 2 will involve conditional probabilities $\{p^2(r_s|x_s)\}$ defined over the same domain, where

$$p^2(r_s|x_s) = p^1(r_s|x_s - 1000)$$

for each subjective perception r_s and each possible final wealth level x_s.

Reversals of the preference ordering of the kind reported by Kahneman and Tversky can easily be explained by reference-dependence of this sort in the way that subjective perceptions of value are coded. For example, suppose that in each of the two classes of choice situations, the prior distribution over possible net gains in state s is expected to be $N(0, 1000)$, and the processing capacity allocated to the perception of the value of outcomes in each class of situations is one-half a binary digit. Then the optimal information structure in each case is of the kind shown in Figure 4, where one unit on the horizontal axis in the figure should now be interpreted as an increase in wealth by 1000. The mean normalized subjective value (MNSV) assigned to each lottery in each of the two choice situations is then shown in Figure 9. Here the horizontal axis x indicates the amount by which the DM's final wealth exceeds initial wealth, and the vertical axis plots $\hat{z}$, the DM's estimate of the normalized value (*i.e.*, $(\hat{-}\mu)/\sigma$, where $\hat{x}$ is the estimate of x). Each of the two sigmoid curves plots the function $\mathrm{E}[z|x]$ for one of the two classes of choice situations: the black curve for class 1 (the prior with $\mu = 1000$) and the grey curve for class 2 (the prior with $\mu = 2000$).[63]Note that since z is a linear transformation of x (with the same linear transformation applied for all possible outcomes in a given class of choice situations), whichever of a pair of lotteries has the higher MNSV $\hat{z}$ will also have the higher MSV, and should on average be preferred.

In the case of option (a), which results in $x = 1500$ with certainty, the MNSV is given by the black dot above $x = 1500$ on the curve corresponding to the given choice situation. In the case of option (b), the MNSV for one state will be given by the black dot above $x = 1000$ on the appropriate curve, while the MNSV for the other state will be given by the black dot above $x = 2000$ on that same curve. The overall MNSV for option (b), averaging the MNSVs for the two equiprobable states, will then be given by the white dot above $x = 1500$, the midpoint of the dashed line

[63] Each of these curves reproduces the function $\phi(z)$ plotted in Figure 7 for the case $C = 0.5$ bits, with a suitable linear transformation of the horizontal axis.

connecting the two black dots representing the MNSVs for the individual states. The figure clearly shows that in this numerical example, option (a) should be preferred on average to option (b) in Problem 1 (the black dot is higher than the white dot), while option (b) should be preferred on average to option (a) in Problem 2 (the white dot is higher than the black dot, in this case). Hence the experimental results of Kahneman and Tversky are quite consistent with this model of valuation.

While the theory proposed here exhibits reference-dependence, there is not really a "reference point" as proposed by Kahneman and Tversky. Perceptions are relative, in the proposed theory, but they are relative to a *probability distribution* of values to which the DM's perceptual system has adapted. In the example above, the only kind of change in the prior distribution that is considered is a shift in the mean, with the distribution of values relative to the mean remaining unchanged; in such a case, it is possible to say that what is perceived is *x relative to the prior mean,* and so one can think of the mean as a "reference point." But in fact, the theory of the coding of perceptions of value proposed here is not based in any way on direct comparison of the actual value with the prior mean or any other single reference value.[64] In the interpretation proposed in section 1.3, this is also true of adaptation in visual perception — the phenomenon to which Kahneman and Tversky appeal in arguing for the psychological realism of assuming there to be a "reference point."

The theory proposed here also makes it clear that, in general, there is no reason to associate the reference point with the *status quo,* either at the time that a decision is made or some earlier time. The coding of perceptions of value is predicted to depend on the prior distribution of potential values to which the DM's perceptual system has adapted. This prior is presumably based on previous experience, but it need not be determined by the DM's current situation or situation in the recent past — it might instead reflect a *range* of situations encountered in the past. Moreover, the

[64] In the theory of reference-dependent choice of Koszegi and Rabin (2006), the reference point is also replaced by an entire probability distribution of reference values, but the reason for this is not very closely related to the one developed here. In the theory of Koszegi and Rabin, the only thing that matters for choice, apart from "consumption utility" that is assumed to be observed with perfect precision, is a set of pairwise comparisons between each possible outcome and each of the individual reference values in a certain distribution, determined by the distribution of outcomes that may actually occur. Each of the contributions to "gain-loss utility" from one of these comparisons depends only on the outcome and the individual reference value in question; none of these individual components of the evaluation depend on the entire distribution.

range of past situations that should determine the prior are not those that have been experienced most *recently,* but rather those that represent situations most *similar* to what the DM understands his current situation to be. In fact, the prior represents an *expectation* about the possible outcomes in situations like the current one. Changes in the status quo are relevant only because one's current situation is often the most relevant basis for a prediction of one's future; when other bases for prediction exist, these are what should be relevant.[65]

Insofar as it predicts that outcomes are evaluated not in absolute terms, but relative to the DM's expectations, this theory is related to the theory of reference-dependent choice proposed by Koszegi and Rabin (2006); and a fair amount of evidence suggests that expectations are indeed important for reference-dependent behavior (*e.g.,* Crawford and Meng, 2011). There are, however, notable differences between this theory and that of Koszegi and Rabin. Koszegi and Rabin propose that the relevant expectation should be what they DM *ordinarily receives* in a situation like the current one, given the choice that she characteristically makes — what *could* have been received in options that are not chosen has no effect on the reference point. In the theory proposed here, instead, the DM's perceptual system is adapted to allow more accurate evaluation of the kind of options that one *expects to have occasion to evaluate* in a situation like the current one — hence all of the options that one expects to evaluate (and needs to be able to see without too much distortion) are relevant, not simply the ones that are chosen.

More generally, Koszegi and Rabin's theory is really about the *reaction* to obtaining an outcome different from what was expected, and the way in which choice should be affected by correct anticipation of that aspect of the way that one will be affected (through this predictable reaction) by the outcomes that can result from one's various choice options — under the assumption that outcomes are perfectly perceived when they occur, and also perfectly perceived in anticipation when considering which option to choose. (All surprises that occur represent intrinsic uncertainty about which state of the world will occur, not any misperceptions of what one will receive in a particular state.) Thus it is not a theory of distorted *perceptions* at all.

[65] Kahneman and Tversky themselves argue that the reference point may well be determined by expectations rather than correspondingly simply to the status quo (1979, pp. 286-288). The present theory proposes a more specific model of the way in which expectations are relevant, though still without fully specifying the time at which the relevant expectations are formed.

In the present theory, instead, differences between outcomes and expectations matter because of the difficulty that one has *recognizing* things that one was not prepared for.

The present theory can explain not only the reference-dependence of valuations, but also another key feature of the behavior documented by Kahneman and Tversky (1979): the coexistence of apparent *risk-aversion* in the domain of gains with apparently *risk-seeking* behavior in the domain of losses. Thus in the case of Problem 1 above (a choice between certain and uncertain *gains*), the typical respondent preferred the sure thing, even though both options imply the same mean final wealth; whereas in Problem 2 (a choice between certain and uncertain *losses*), the typical respondent preferred to gamble. In fact, both types of behavior are difficult to square with the standard theory that postulates maximization of the expected utility of final wealth; while the postulate of risk aversion (understood essentially in the way proposed by Bernoulli) is a staple of textbook accounts of choice under uncertainty, the *degree* of risk aversion with respect to final wealth that would have to be assumed to explain people's observed aversion to randomness of gains in individual gambles of modest size is extraordinary, and quite inconsistent with other observed choices, as shown by Rabin (2000). But the puzzle is even sharper when one recognizes that the hypothesis of risk aversion (a strictly concave von Neumann-Morgenstern utility function) that might be invoked to explain choice in Problem 1 is certainly contradicted by observed choice in Problem 2.

Kahneman and Tversky explain both types of behavior by hypothesizing that choice depends on an evaluation of potential gains and losses from the individual decision, and that such prospects are valued in accordance with a weighted average of a certain nonlinear function of the net gain in each possible outcome, where the weights depend on the probability of occurrence of the different outcomes,[66] and the nonlinear "value function" is concave for gains but convex for losses. The explanation provided by the present theory has a similar formal structure. The MSV assigned to each of the lotteries in a given choice situation will be an average of the MSVs of the final outcomes in each of the possible states, and the MSV for each state will

[66] Another key feature of prospect theory is the proposal that the relative weight on each outcome is a nonlinear function of the true probability. I do not propose any interpretation of this nonlinearity here; in cases like Problems 1 and 2 above, this issue does not arise, as the probability of each outcome is the same.

be a nonlinear function of the *normalized* objective final wealth in that state, which is a linear transformation of the net gain in wealth as a result of the choice. The nonlinear function used to convert objective values into mean subjective values is a linear transformation of the function $\phi(z)$ plotted in Figure 7 and again in Figure 9; and as we have seen, it is concave for $z > 0$ but convex for $z < 0$.

The present theory, however, derives this shape of the transformation function from deeper theoretical assumptions, rather than directly assuming it on the basis of observed behavior. The proposed theoretical derivation is closely related to the explanation for the shape of the "value function" proposed by Kahneman and Tversky (1979), who attribute both the concavity in the case of gains and the convexity in the case of losses to diminishing sensitivity to progressively larger changes of either sign, which they regard as a further instance of a general property of perception, that "the psychological response is a concave function of the magnitude of physical change" (p. 278). The present theory offers a deeper explanation, in terms of the efficient use of limited information-processing capacity, for this general phenomenon, which is then also applied to judgments of value, as Kahneman and Tversky propose. It is worth noting that the present theory predicts the existence of diminishing sensitivity to *both* gains and losses, without having to assume that gains and losses are coded *separately.* Instead, we have considered the optimal subjective representation of a single real variable that may be either larger or smaller than the prior mean, and found that there should be diminishing sensitivity to larger departures from the prior mean in either direction.

The theory presented here does not explain all aspects of the value function postulated by Kahneman and Tversky (1979). In the examples computed above, the transformation function is antisymmetric ($\phi(-z) = -\phi(z)$), whereas Kahneman and Tversky propose as one of the key properties of their value function that it is "steeper for losses than for gains." The antisymmetry is not, however, a general prediction of the type of theory proposed here; it obtains in the case analyzed in section 2 only because of the assumed symmetry of the loss function with respect to over- and underestimates of the true value and the assumed symmetry of the prior distribution around its mean. An extension of the theory to cases with other loss functions and/or other types of prior distributions will be an important topic for further investigation; it would be interesting to know if the kind of asymmetry proposed by Kahneman and Tversky can be explained through such a generalization.

Kahneman and Tversky (1979) also propose that there should be a kink in the value function at the point of zero gain or loss, so that the negative value assigned to a loss of a given size is a multiple greater than 1 of the negative of the value assigned to a gain of the same size, even when the size of the gain is made arbitrarily small. This feature of prospect theory has received particular emphasis in the subsequent literature, notably because of the explanation that it provides for "endowment effects."[67] The theory proposed here provides no explanation for a kink at the value of an attribute corresponding to the prior mean (or any other conception of the "reference point"); indeed, as noted above, there is really no "reference point" in this theory.

It is possible that a more adequate theory would hypothesize separate coding of gains and losses, as a consequence of which different valuations of small gains and small losses could easily be possible. (This would mean assuming additional constraints on the class of possible subjective representations than those assumed in the relatively parsimonious theory explored here.) On the other hand, there are other possible explanations for endowment effects than the hypothesis of a kink in the mapping of objective values into subjective perceptions of value. It may be that people perceive more accurately the attributes of objects already in their possession than ones that they do not possess (but can see); or it may be that they make inferences about the value of objects from the fact that someone else would offer an exchange, and not simply from what they can observe about the objects' characteristics. Hence it may be possible to understand endowment effects, within a more general theory, without changing the account given here of choice between lotteries.

4 Conclusions

I have proposed a theory of errors in the valuation of options confronting economic decisionmakers. In this theory, valuations are optimal, in the sense of minimizing the mean of squared evaluation error, subject to an information-theoretic constraint on the processing capacity required to observe the attributes of the options with a particular degree of precision. I have shown that the resulting theory implies an inevitable form of reference-dependence of valuations, insofar as the mapping from

[67] See, for example, Kahneman (2011), chap. 27.

objective characteristics to subjective valuations that will be optimal depends on the prior distribution of potential valuations to which the DM's perceptual system has adapted.

I have argued that a theory of this kind predicts a number of types of departures from the implications of full-information rational choice theory that have been observed to occur in human decisions. These include stochastic variation in the choice that is made from among a fixed set of options, focusing effects, decoy effects, valuation of risky options on the basis of the distribution of gains and losses from the individual choice rather than implications for final wealth after many random events, and the coexistence of risk-aversion with respect to gains with risk-seeking with respect to losses. Other modifications of standard decision theory have been proposed, of course, in order to account for each of these anomalies. The theory offered here, however, offers the prospect of unified explanation of all of them on the basis of a single departure from standard assumptions. For that reason, I believe that the implications of this type of theory deserve to be analyzed in further detail.

References

[1] Adrian, Edgar D.A., *The Basis of Sensation: The Action of the Sense Organs,* London: Christophers, 1928.

[2] Attneave, Fred, "Some Informational Aspects of Visual Perception," *Psychological Review* 61: 183-193 (1954).

[3] Barlow, Horace B., "Possible Principles Underlying the Transformation of Sensory Messages", in W.A. Rosenblith, ed., *Sensory Communication,* Cambridge, MA: MIT Press, 1961.

[4] Bordalo, Pedro, "Choice-Set Effects and Salience Theory," unpublished, Harvard University, December 2010.

[5] Bordalo, Pedro, Nicola Gennaioli, and Andrei Shleifer, "Salience Theory of Choice Under Risk," NBER Working Paper no. 16387, September 2010.

[6] Bordalo, Pedro, Nicola Gennaioli, and Andrei Shleifer, "Salience and Consumer Choice," unpublished, Harvard University, October 2011.

[7] Brenner, Naama, William Bialek, and Rob de Ruyter van Steveninck, "Adaptive Rescaling Maximizes Information Transmission," *Neuron* 26: 695-702 (2000).

[8] Britten, Kenneth H., Michael N. Shadlen, William T. Newsome, and J. Anthony Movshon, "The Analysis of Visual Motion: A Comparison of Neuronal and Psychophysical Performance," *Journal of Neuroscience* 12: 4745-4765 (1992).

[9] Caballero, Ricardo J., and Eduardo M.R.A. Engel, "Microeconomic Adjustment Hazards and Aggregate Dynamics," *Quarterly Journal of Economics* 108: 359-383 (1993).

[10] Caballero, Ricardo J., and Eduardo M.R.A. Engel, "Explaining Investment Dynamics in U.S. Manufacturing: A Generalized *(S,s)* Approach," *Econometrica* 67: 783-826 (1999).

[11] Caballero, Ricardo J., Eduardo M.R.A. Engel and John C. Haltiwanger, "Aggregate Emloyment Dynamics: Building from Microeconomic Evidence," *American Economic Review* 87: 115- 137 (1997).

[12] Cover, Thomas M., and Joy A. Thomas, *Elements of Information Theory,* New York: Wiley-Interscience, 2d ed., 2006.

[13] Cox, D.R., and D.V. Hinkley, *Theoretical Statistics,* London: Chapman and Hall, 1974.

[14] Crawford, Vincent P., and Juanjuan Meng, "New York City Cabdrivers' Labor Supply Revisited: Reference-Dependent Preferences with Rational-Expectations Targets for Hours and Income," *American Economic Review* 101: 1912-1932 (2011).

[15] Dayan, Peter, and L.F. Abbott, *Theoretical Neuroscience: Computational and Mathematical Modeling of Neural Systems,* Cambridge, MA: MIT Press, 2001.

[16] Dotsey, Michael, Robert G. King, and Alexander L. Wolman, "State Dependent Pricing and the General Equilibrium Dynamics of Money and Output," *Quarterly Journal of Economics* 114: 655-690 (1999).

[17] Fairhall, Adrienne, "Spike Coding," in K. Doya, S. Ishii, A. Pouget, and R.P.N. Rao, eds., *Bayesian Brain: Probabilistic Approaches to Neural Coding,* Cambridge, MA: MIT Press, 2007.

[18] Frisby, John P., and James V. Stone, *Seeing: The Computational Approach to Biological Vision,* 2d ed., Cambridge, MA: MIT Press, 2010.

[19] Gabaix, Xavier, "A Sparsity-Based Model of Bounded Rationality," unpublished, New York University, October 2010.

[20] Gabbiani, Fabrizio, and Steven J. Cox, *Mathematics for Neuroscientists,* Amsterdam: Academic Press, 2010.

[21] Glimcher, Paul W., *Foundations of Neuroeconomic Analysis,* Oxford: Oxford University Press, 2011.

[22] Goeree, Jacob K., Charles A. Holt, and Thomas R. Palfrey, "Quantal Response Equilibrium," in S.N. Durlauf and L.E. Blume, eds., *New Palgrave Dictionary of Economics,* 2d ed., Palgrave Macmillan, 2008.

[23] Green, David M., and John A. Swets, *Signal Detection Theory and Psychophysics,* New York: Wiley, 1966.

[24] Gul, Faruk, Wolfgang Pesendorfer, and Tomasz Strzalecki, "Behavioral Competitive Equilibrium and Extreme Prices," unpublished, Princeton University, August 2011.

[25] Heath, Timothy B., and Subimal Chatterjee, "Asymmetric Decoy Effects on Lower-Quality versus Higher-Quality Brands: Meta-Analytic and Experimental Evidence," *Journal of Consumer Research* 22: 268-284 (1995).

[26] Huber, Joel, John W. Payne, and Christopher Puto, "Adding Asymmetrically Dominated Alternatives: Violations of Regularity and the Similarity Hypothesis," *Journal of Consumer Research* 9: 90-98 (1982).

[27] Kahneman, Daniel, *Attention and Effort,* Englewood Cliffs, NJ: Prentice-Hall, 1973.

[28] Kahneman, Daniel, "Reference Points, Anchors, Norms, and Mixed Feelings," *Organizational Behavior and Human Decision Processes* 51: 296-312 (1992).

[29] Kahneman, Daniel, "Maps of Bounded Rationality: A Perspective on Intuitive Judgment and Choice," in T. Frangsmyr, ed., *Les Prix Nobel 2002,* Stockholm: Almquist & Wiksell International, 2003.

[30] Kahneman, Daniel, Thinking, Fast and Slow, New York: Farrar, Straus and Giroux (2011).

[31] Kahneman, Daniel, Alan B. Krueger, David Schkade, Norbert Schwartz, and Arthur A. Stone, "Would You Be Happier if You Were Richer? A Focusing Illusion," *Science* 312: 1908-1910 (2006).

[32] Kahneman, Daniel, and Amos Tversky, "Prospect Theory: An Analysis of Decision Under Risk," *Econometrica* 47: 263-291 (1979).

[33] Kandel, Eric R., James H. Schwartz, and Thomas M. Jessell, *Principles of Neural Science,* 4th ed., New York: McGraw-Hill, 2000.

[34] Kets, Willemien, "The Minority Game: An Economics Perspective," CentER Discussion Paper no. 2007-53, Tilburg University, June 2007.

[35] Koszegi, Botond, and Matthew Rabin, "A Model of Reference-Dependent Preferences," *Quarterly Journal of Economics* 121: 1133-1165 (2006).

[36] Koszegi, Botond, and Adam Szeidl, "A Model of Focusing and Economic Choice," unpublished, U.C. Berkeley, July 2011.

[37] Laughlin, Simon, "A Simple Coding Procedure Enhances a Neuron's Information Capacity," *Zeitschrift fur Naturforschung* 36: 910-912 (1981).

[38] Laughlin, Simon B., and Roger C. Hardie, "Common Strategies for Light Adaptation in the Peripheral Visual Systems of Fly and Dragonfly," *Journal of Comparative Physiology A* 128: 319-340 (1978).

[39] Lipman, Barton, "Information Processing and Bounded Rationality: A Survey," *Canadian Journal of Economics* 28: 42-67 (1995).

[40] Luce, R. Duncan, *Individual Choice Behavior: A Theoretical Analysis,* New York: Wiley, 1959.

[41] Mackowiak, Bartosz, and Mirko Wiederholt, "Optimal Sticky Prices under Rational Inattention," *American Economic Review* 99: 769-803 (2009).

[42] Matejka, Filip, and Alisdair McKay, "Rational Inattention to Discrete Choices: A New Foundation for the Multinomial Logit Model," unpublished, Boston University, February 2011.

[43] McFadden, Daniel, "Conditional Logit Analysis of Qualitative Choice Behavior," in P. Zarembka, ed., *Frontiers in Econometrics,* New York: Academic Press, 1974.

[44] McFadden, Daniel, "Rationality for Economists?" *Journal of Risk and Uncertainty* 19: 73-105 (1999).

[45] McFadden, Daniel L., "Economic Choices," *American Economic Review* 91: 351-378 (2001).

[46] McKelvey, Richard D., and Thomas R. Palfrey, "Quantal Response Equilibria for Normal-Form Games," *Games and Economic Behavior* 10: 6-38 (1995).

[47] Midrigan, Virgiliu, “Menu Costs, Multi-Product Firms, and Aggregate Fluctuations,” unpublished, New York University, January 2010.

[48] Miller, George A., “The Magical Number Seven, Plus or Minus Two: Some Limits on Our Capacity for Processing Information,” *Psychological Review* 63: 81-97 (1956).

[49] Morwitz, Vicki G., Eric A. Greenleaf, and Eric J. Johnson, “Divide and Prosper: Consumers’ Reactions to Partitioned Prices,” *Journal of Marketing Research* 35: 453-463 (1998).

[50] Neyman, Abraham, “Bounded Rationality Justifies Cooperation in the Finitely Repeated Prisoners’ Dilemma Game,” *Economic Letters* 19: 227-229 (1985).

[51] Rabin, Matthew, “Risk Aversion and Expected-Utility Theory: A Calibration Theorem,” *Econometrica* 68: 1281-1292 (2000).

[52] Rieke, Fred, David Warland, Rob de Ruyter van Steveninck, and William Bialek, *Spikes: Exploring the Neural Code,* Cambridge, MA: MIT Press, 1997.

[53] Rubinstein, Ariel, “Finite Automata Play the Repeated Prisoners’ Dilemma,” *Journal of Economic Theory* 39: 83-96 (1986).

[54] Sayood, Khalid, *Introduction to Data Compression,* 3d ed., Amsterdam: Morgan Kaufmann, 2005.

[55] Schkade, David A., and Daniel Kahneman, “Does Living in California Make People Happy? A Focusing Illusion in Judgments of Life Satisfaction,” *Psychological Science* 9: 340-346 (1998).

[56] Shannon, Claude E., “A Mathematical Theory of Communication,” *Bell System Technical Journal* 27: 379-423 and 623-656 (1948).

[57] Sims, Christopher A., “Stickiness,” *Carnegie-Rochester Conference Series on Public Policy* 49: 317-356 (1998).

[58] Sims, Christopher A., “Implications of Rational Inattention,” *Journal of Monetary Economics* 50: 665-690 (2003).

[59] Sims, Christopher A., "Rational Inattention and Monetary Economics," in B.M. Friedman and M. Woodford, eds., *Handbook of Monetary Economics,* vol. 3A, Amsterdam: Elsevier, 2011.

[60] Sperling, George, and Barbara Ann Dosher, "Strategy and Optimization in Human Information Processing," in K.R. Boff, L. Kaufman, and J.P. Thomas, eds., *Handbook of Perception and Human Performance,* New York: Wiley, 1986.

[61] Thurstone, Louis L., "A Law of Comparative Judgment," *Psychological Review* 34: 273-286 (1927).

[62] Thurstone, Louis L., *The Measurement of Values,* Chicago: University of Chicago Press, 1959.

[63] Tversky, Amos, and Daniel Kahneman, "Judgment Under Uncertainty: Heuristics and Biases," *Science* 185: 1124-1131 (1974).

[64] Weber, Elke U., "Perception Matters: Psychophysics for Economists," in I. Brocas and J.D. Carrillo, eds., *The Psychology of Economic Decisions, Volume 2: Reasons and Choices,* Oxford: Oxford University Press, 2004.

[65] Woodford, Michael, "Inattentiveness as a Source of Randomized Discrete Adjustment," unpublished, Columbia University, April 2008.

[66] Woodford, Michael, "Information-Constrained State-Dependent Pricing," *Journal of Monetary Economics* 56(S): 100-124 (2009).

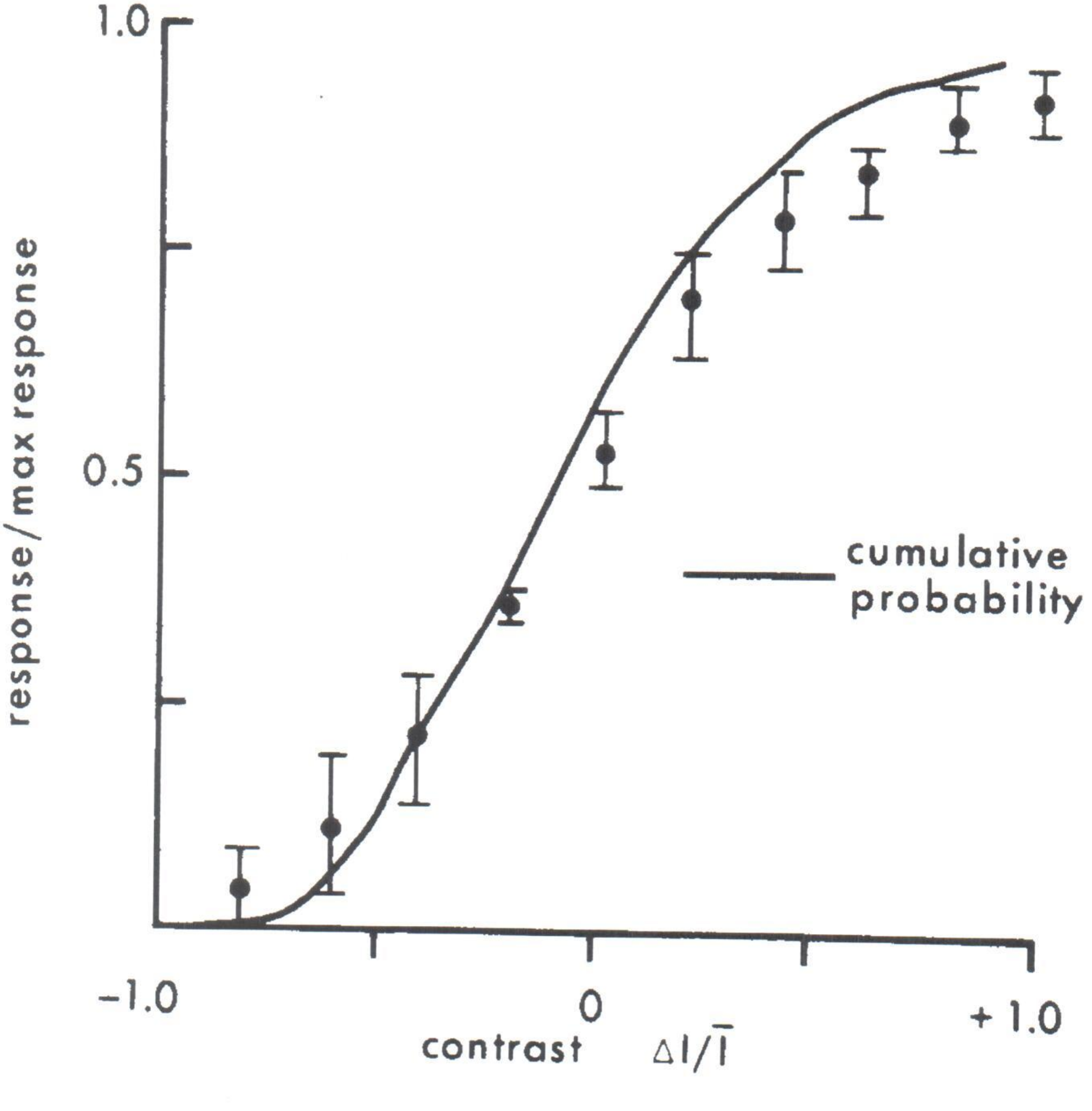

Figure 3. From Laughlin (1981).

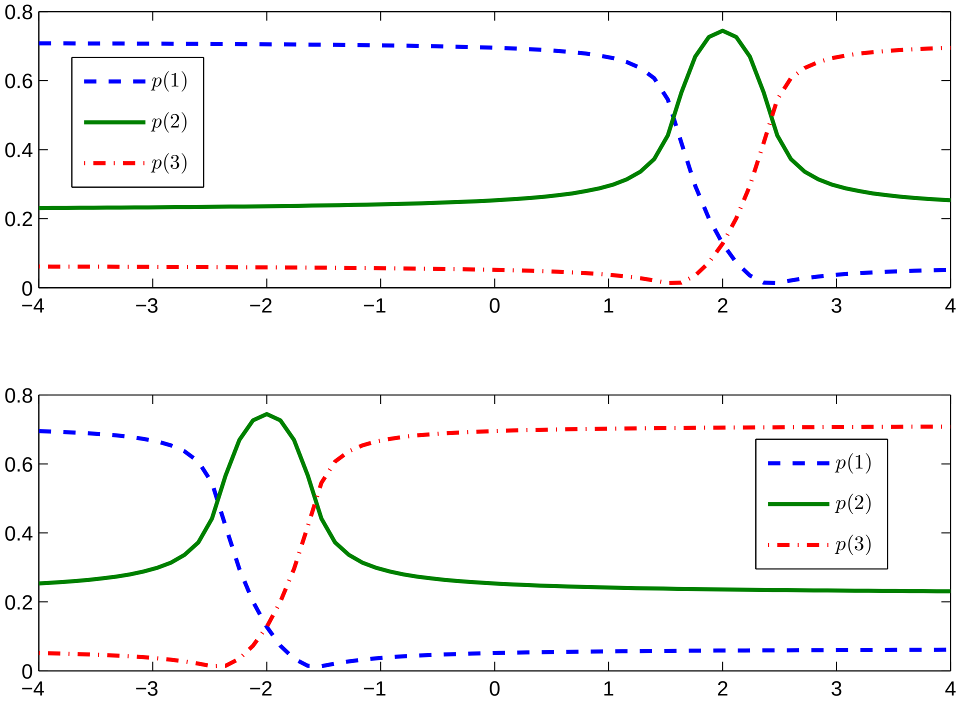

Figure 4: Optimal information structures for a capacity limit C equal to one-half a binary digit, when the prior distribution is $N(\mu, 1)$. Plots show the probability of each of three possible subjective representations, conditional on the true state. Panel (a): $\mu = -2$. Panel (b): $\mu = +2$.

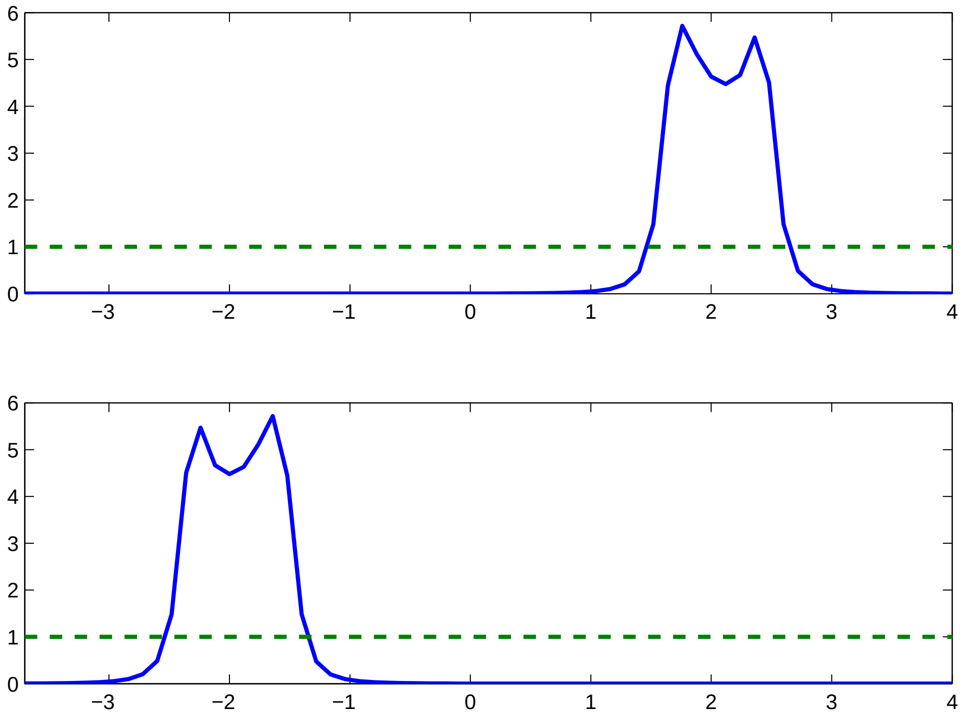

Figure 5: Fisher information $I^{Fisher}(x)$ measuring the discriminability of each objective state x from nearby states under optimal information structures. Solid line corresponds to the optimal structure subject to a limit on the capacity C, dashed line to the optimal structure subject to a limit on mutual information. The two panels correspond to the same two prior distributions as in Figure 4.

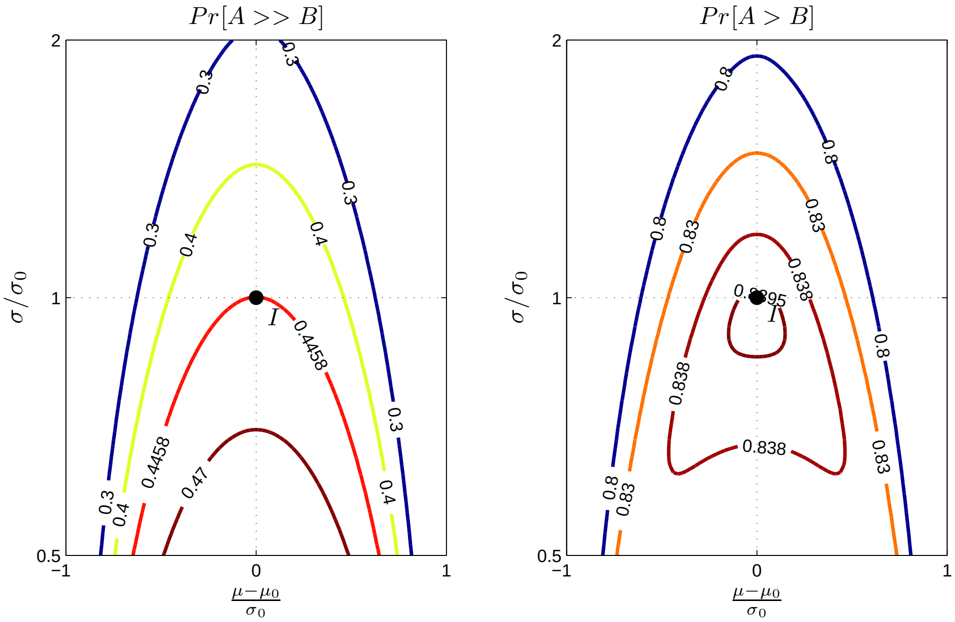

Figure 6: The probability that the superior option (with respect to a particular attribute) will be perceived as superior, if $x^A = \mu_0 + \sigma_0$, $x^B = \mu_0 - \sigma_0$, and the processing capacity allocated to perception of the attribute is 0.5 bits. Panel (a) shows the probability that A is perceived as strongly superior, panel (b) the probability that it is perceived as at least somewhat superior. The point I indicates the baseline prior (μ_0, σ_0).

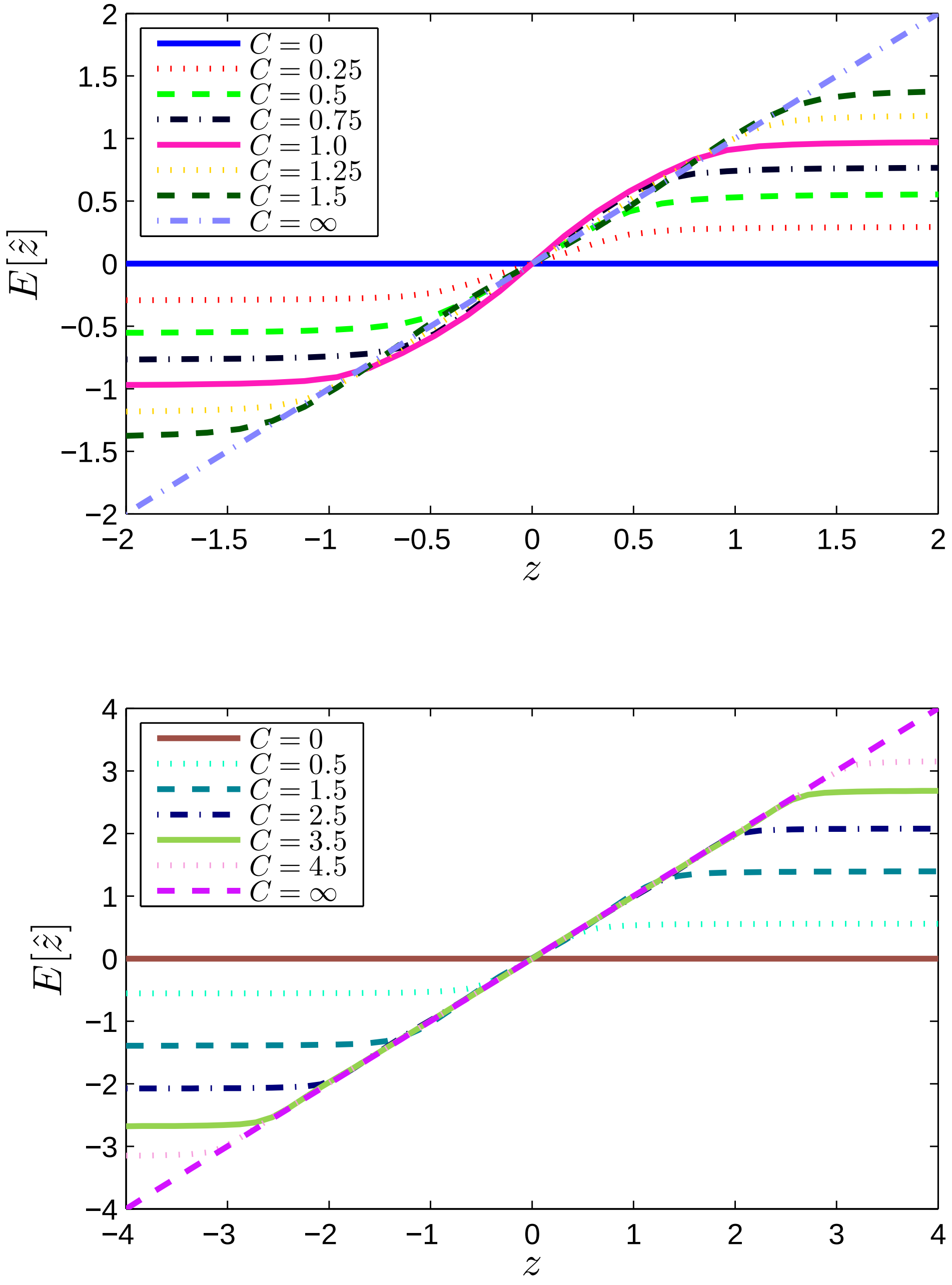

Figure 7: Mean normalized subjective value as a function of actual normalized value z of a particular attribute, under different values of the capacity limit C. Panel (a) gives more details of the effects of tight capacity limits for z of modest size. Panel (b) shows the effects of higher capacity limits in the case of more extreme z.

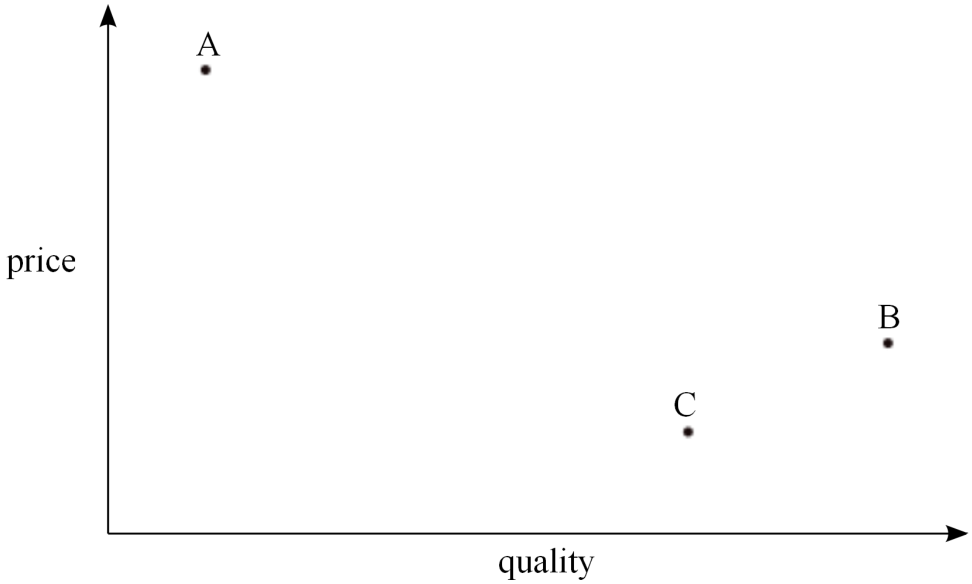

Figure 8: Location of three brands in attribute space. Introduction of the “decoy” brand C increases market share of “target” brand B at the expense of competitor A.

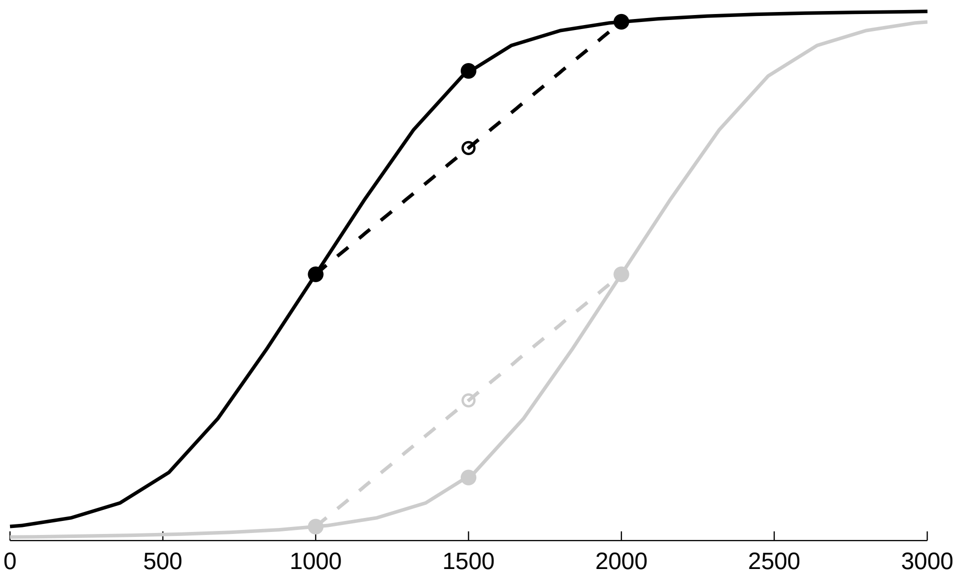

Figure 9: Mean normalized subjective value (vertical axis) for each of the possible lotteries in two choice problems studied by Kahneman and Tversky (1979). The horizontal axis shows final wealth in excess of initial wealth, in each of the two possible states under a given lottery. Black dots plot MNSV of the outcome in an individual state; white dots plot the MNSV for a gamble in which the outcomes in the two possible states are the two black dots at the endpoints of the dashed line. Black line: plot of E$[z|x]$ when the DM is initially given 1000. Black line: plot of E$[z|x]$ when the DM is initially given 1000. Grey line: plot of E$[z|x]$ when the DM is initially given 2000.