Title

Toward a Synthesis of Cognitive Biases: How Noisy Information Processing Can Bias Human Decision Making

Journal

Psychological Bulletin, 138(2)

ISSN

0033-2909

Author

Hilbert, Martin

Publication Date

2012-03-01

DOI

10.1037/a0025940

Author's version of article published in

Psychological Bulletin

Online First Version, Nov 28, 2011. doi: 10.1037/a0025940

Toward a Synthesis of Cognitive Biases: How Noisy Information Processing Can Bias Human Decision Making

Martin Hilbert

Annenberg School of Communication, University of Southern California (USC), in collaboration with USC Viterbi School of Engineering; and USC Department of Psychology

Email: mhilbert [at] usc [dot] edu

Acknowledgment: I am highly indebted to Gerhard Kramer, from USC's Department of Electrical Engineering (now TU Munich), and would like to thank him, as well as Ashok Patel from the same Department, for their patience with my questions and for the excitement with which they introduced me to the important subtleties of the intriguing worlds of information- and probability theory.

Abstract

A single coherent framework is proposed to synthesize longstanding research on eight seemingly unrelated cognitive decision-making biases. During the past six decades, hundreds of empirical studies have resulted in a variety of rules of thumb that specify how humans systematically deviate from what is normatively expected from their decisions. Several complementary generative mechanisms have been proposed to explain those cognitive biases. Here it is suggested that (at least) eight of these empirically detected decision-making biases can be produced by simply assuming noisy deviations in the memory-based information processes that convert objective evidence (observations) into subjective estimates (decisions). An integrative framework is presented to show how similar noise-based mechanisms can lead to conservatism, the Bayesian likelihood bias, illusory correlations, biased self-other placement, subadditivity, exaggerated expectation, the confidence bias, and the hard-easy effect. Analytical tools from information theory are used to explore the nature and limitations that characterize such information processes for binary and multiary decision-making exercises. The ensuing synthesis offers formal mathematical definitions of the biases and their underlying generative mechanism, which permits a consolidated analysis of how they are related. This synthesis contributes to the larger goal of carving a coherent picture out of the myriad of seemingly unrelated biases and their potential psychological generative mechanisms. Limitations and research questions are discussed.

Table of Contents

When human judges make decisions, they essentially make a choice among several alternatives (Edwards, 1954). It turns out that these choices are "predictably irrational" (Ariely, 2008). We always end up with the same kinds of deviations from what is normatively predicted by classical probability and utility theory to be the optimal outcome of those choices. The consistency of such systematic biases can be useful for predicting individual behavior and can also have disastrous large-scale consequences for society as a whole. If the mistakes in our judgments were random, the deviations would cancel each other out. For example, in a specific situation, some investors would overestimate and others underestimate risk. The overall result would be a self-regulating social system, indistinguishable from the one proposed by the efficient market hypothesis with its rational actors. Contrary to such views, however, our judgments are systematically biased toward one side or the other, and in specific situations, the large majority of investors will either over- or underestimate risk, not both. The worldwide economic crisis of 2008 delivered hard evidence for such dynamics (Ariely, 2008) and left many previous defenders of the rational and efficient market hypothesis in "shocked disbelief" (Greenspan, 2008; p. 16).

Biases, Their Models, and Outline

For psychologists, human irrationality is old news. Six decades of psychological research on human judgment and decision-making have produced an impressive list of "heuristics and biases" (Tversky & Kahneman, 1974). A bias usually takes the following form: when confronted with evidence of type X, a judge will consistently chose alternative B instead of the expected alternative A. Because we are very consistent with our biases, rules that describe such biases have large predictive power (Baron, 2007; Hastie & Dawes, 2001; Rubenstein, 1998;; Wilkinson, 2007). A popular text book lists 53 such biases (Baron, 2008; Table 2.1).

Yet despite the predictive power of the rules that describe these biases, many are still hesitant to take these findings as a solid foundation for larger theories. One reason for this hesitation is that the list of biases is a quite loose grab bag of empirical regularities that still lacks the foundation of a thorough theory itself. This state of affairs has contributed to what another popular textbook describes as "quite conflicting beliefs regarding fundamental aspects" of human decision-making (Wilkinson, 2008, p.12). The identification of "if X, then B" statements is an important first task in describing human behavior, but it does not explain the origin of these deviations and how they are interrelated. In short, we have a solid description of many parts but do not yet see the big picture. A coherent and solid theory of human decision-making would require such understanding. In this article I attempt to contribute to this goal by offering a conceptual foundation for the integration of several biases within one common frame of reference.

Complex and Simple Generative Mechanisms for Cognitive Biases

In the search for generative mechanisms of our cognitive biases, an obvious first suggestion is that their consistencies occur because we all share the same information processing system: the human mind. Because the mind's information processing capacity is biologically limited (for example, we possess neither infinite nor photographic memory), we end up with "bounded rationality" (Simon, 1955, 1956). Additionally, we also seem to employ short-cuts in our information processing that aim at reducing cognitive effort, known as "heuristics" (Goldstein, & Gigerenzer, 2002; Kahneman, Slovic, & Tversky, 1982; Shah & Oppenheimer, 2008). These simple but often effective approximations make us use a representative case instead of the specific one (representativeness). They also make us work whatever first comes to mind (availability); and based on our first thoughts, it turns out that the subsequent mental search process is limited (adjustment and anchoring). Other potential generative mechanisms for our cognitive biases are emotional and moral motivations (Loewenstein, Weber, Hsee, Welch, & Ned, 2001; Pfister & Böhm, 2008), as well as social influences (Wang, Simons, & Bredart, 2001; Yechiam, Druyan, & Ert, 2008). Last but not least, it has been proposed that noisy information processing leads to bias in our decision-making. Loosely speaking, we understand "noise" as "distorted mixing of information flows" (a more concrete definition will follow later). These are to be expected in human judgment. Our mind is the result of biological evolution, which does not strive for perfection or even theoretical optimization, but simply for a competitive degree of "fitness" in a specific environment. From this perspective, it should not be surprising that the design of the information processing system we employ when making decisions is imperfect and that a certain degree of distortion takes place arising simply from the sloppy design of the system.

In this article, I exclusively focus on this last kind of generative mechanism for our cognitive biases: noise. I identify four distinct mental processes and assert that eight empirically detected decision-making biases are the inevitable result if we suppose a certain and justifiable kind of distortion in these processes. Thus, the present argument is about sufficient rather than necessary conditions. The major theoretical point is that simple properties of a noisy information processing system are sufficient to produce several biases. This argument does not exclude the possible contribution of more complex factors—such as heuristics, higher cognitive functions, emotions, motivations or social influences—which may also be sufficient but not necessary to understand the phenomena of interest. Yet because the supposition of noisy information processing is a simple and elegant way to account for the chosen biases, the conclusions of this article will argue that Occam's razor favors naturally occurring noise as the most likely explanation.

A Synthesis of Models

The approach taken in this article follows the logic of existing computer models that simulate several cognitive biases. I do not, however, present a review of all existing computer models, propose a new model, or offer a competitive test of different models. Rather, my purpose is to present an analytical synthesis of several existing models through a systematic literature review. The presented framework provides a unifying theoretical framework for a considerable number of existing noise-based models of cognitive biases. I provide mathematical proofs that many of these models are merely a special case of the broader information theoretic logic outlined in this article.

My approach is inspired by memory models suggesting that judgments and decisions are produced by storing and subsequently retrieving objective evidence in memory and that biases are the result of distortions in this mental process. These models have their roots in the late 1970s and early 1980s (e.g. McClelland & Rumelhart, 1985; Medin & Schaffer, 1978; Whittlesea 1983; see also Hintzman's well-known MINERVA model, 1984, 1988). In these models, computers are essentially fed with some input and then researchers search for a specific kind of distortion of this input (noise) that results in a retrieved output resembling the irrational behavior of human judges. Some models, for example, study the effects of incomplete storage in and retrieval from memory (e.g., Fiedler, 1996; Linville, Fischer, & Salovey, 1989; Metcalfe, 1990;). These models are much in line with the previously mentioned argument that our biological information processing capacities are limited.

I also focus on another kind of noise in the process of storage in and retrieval from memory: confusion and mix-up of evidence. In so-called "stochastic models," the input is distorted according to a probability distribution that represents the error, which leads to a mixed-up output consisting of input+error. In 1994, Erev, Wallsten, and Budescu showed that a specific kind of distortion of some evidence (for example, log-odds plus normally distributed error) can lead to judgments that simultaneously reveal two seemingly unrelated biases: what they call conservatism and overconfidence. At the same time, Wallsten and González-Vallejo (1994) set up a computer-aided "stochastic model of judgment and response," which disturbs input values with symmetric and single peaked noise. Encouraged by these successes, various researchers proposed a series of so-called "random error" or "stochastic models" (see special issue of *Journal of Behavioral Decision Making,* Vol.10,3; Budescu, Erev, Wallsten, &Yates, 1997a). All of them follow the same logic but vary in the kind of noise applied (for example, Budescu, Erev, & Wallsten, 1997b, use a binomial error distribution) and other kinds of technical fine-tuning of matching and parameters setting (see also Budescu, Wallsten, Au, 1997c; Juslin, Olsson, & Bjorkman, 1997;;Merkle, 2009). In 1999, Dougherty, Gettys, and Ogden showed that the logic of Hintzman's MINERVA model (1984, 1988) can be used to replicate several distinct violations of rationality, including conservatism, overconfidence, and the hard-easy effect (called MINERVA-DM = decision-making). This model was subsequently used to artificially

replicate other biases (Bearden & Wallsten, 2004; Dougherty, 2001). Appendix B reviews the basic modus operandi of MINERVA-DM (using the information-theoretic channel logic that we will work with in this article, see also Appendix A).

These models suggest that several of the empirically detected human decision-making biases can be understood as distortions in the process of storage in and retrieval from memory. The problem with these kinds of computer simulations, however, is that they often seem quite arbitrary: the parameters that define the distortion are naturally fine-tuned to favor the desired outcome. The results show only that a specific configuration of a computer program (with one out of many possible parameter settings) can replicate the empirical regularity of biases X and Y, and another specification can replicate the patterns of bias Z. We know little about the boundary properties of these settings or the margins within which they work and do not work. Nor do we know if it is inevitable that the same parameter settings must inevitably produce different biases,or if a specific kind of noise is merely a special case. Worse, we know even less about how these parameter settings are tied to psychological processes and what the generative mechanisms might be that lead to different kinds of noise. What are the limits within which a specific kind of noise works? Is there a class of error distributions that can simultaneously explain a range of seemingly unrelated biases? How are the biases related? Is it psychologically justifiable to suppose this or that kind of distortion? These are some of the questions that previous work leaves open, and that the present synthesis proposes to answer.

Outline of and Contributions to Selected Biases

Table 1 provides guidance throughout the article. Many of its ingredients will become clearer as the different arguments are elaborated. In general, I postulate that a similar kind of distortion can appear in four different mental processes involved in storage in and retrieval from memory. The left column in Table 1 lists these four kinds of distortions: noise between input evidence and output estimate (referred to as $\hat{E}|\underline{E}$), noise between output estimate and input evidence (called $\underline{E}|\hat{E}$), noise between memory and output estimate ($\hat{E}|M$), and finally, noise between output estimate and memory ($M|\hat{E}$). The second column (from the left) in Table 1 lists the eight cognitive biases I will synthesize in this study. As shown in the Table, five of the biases can be attributed to the same cause: a specific kind of noise between the objective input evidence and the subjective output estimate. The other three biases originate from the remaining three noise-based generative mechanisms. The fourth column in Table 1 lists selected studies (both historical and more recent) that have empirically detected these irrational regularities . The list of studies is not exhaustive; it could easily be expanded to a list that contains hundreds of controlled experiments. I elaborate the rest of the specifications in Table 1 throughout the article. The formulas in the third column show the mathematical formalizations of the distinct biases, based on the definitions of the synthesizing framework. These allow unambiguous formalizations of the co-dependencies among and relations across those seemingly

unrelated biases. The properties in the left column specify the kind of noise sufficient to produce the identified kinds of biases.

Table 1: Summary of the biases included in this study and related empirical studies, including mathematical formalizations based on the noisy memory channel framework

Generative mechanism: Noise properties that can produce biases	**Bias:** Empirically observed deviation from normative expectation	**Mathematical formalization of the bias** based on the noisy memory channel	**Selected examples of empirical research** that detected the bias
Noise ($\hat{E}$|$\underline{E}$) between $\underline{E}$vidence and $\hat{E}$stimate: -For binary decision-making tasks: Properties B and N_i. - For equidistant decision-making tasks: Properties S and N, or S and U. -All other kinds of decision-making tasks: Properties D and N.	**Conservatism:** Based on the observed evidence, estimates are not extreme enough	$0 \le [r \times \sigma_{\hat{E}}] \le \sigma_{\underline{E}}$ $0 \le [r \times \sigma_{P(\hat{E})}] \le \sigma_{P(\underline{E})}$	Kaufman, et.al. (1949); Attneave (1953); Fischhoff, et.al. (1977); MacGregor, et.al (1988); Fiedler (1991).
	Bayesian likelihood: estimates of conditional probabilities are conservative	$0 \le [r \times \sigma_{P(\hat{E}\|c)}] \le \sigma_{P(\underline{E}\|c)}$ with c being some overall conditioning event for the task	Phillips and Edwards (1966); Phillips, et.al. (1966); Edwards (1968); DuCharme (1970); Messick and Campos (1972).
	Illusory correlation of minority stereotyping: estimates on a two-dimensional distribution become correlated	$0 \le [r \times \sigma_{\hat{E}\|c}] \le \sigma_{\underline{E}\|c}$ $0 \le [r \times \sigma_{P(\hat{E}\|c)}] \le \sigma_{P(\underline{E}\|c)}$ With c being a cross-tabulated event	Hamilton and Gifford (1976); Hamilton, Dugan, and Trolier (1985); Pryor (1986); Fiedler (1991); Smith, (1991).
	Placement: estimates about myself are better than estimates about others	$0 \le slope_{\underline{E}\ others} \le$ $\le slope_{\underline{E}\ own} \le 1$	Cooper, et.al. (1988); Kruger and Dunning, (1999); Kruger (1999); Moore and Cain (2007); Moore and Healy (2008).
	Subadditivity: estimate of a likelihood is less than the sum of its (more than two) mutually exclusive components	$p(\hat{e}_i) \le \Sigma\, p(\hat{e}_d)$ with d being a decomposition of event i.	Tversky and Koehler (1994); Redelmeier, et.al. (1995); Fox and Levav (2000); Bearden and Wallsten (2004).
Noise ($\underline{E}$|$\hat{E}$) between $\hat{E}$stimate and $\underline{E}$vidence: -Binary: Properties B and N_i. - Equidistant: Properties S and N, or S and U. -All others: Properties D and N.	**Exaggerated expectation:** Based on the estimates, real-world evidence turns out to be less extreme than our expectations (conditionally inverse of the conservatism bias)	$\sigma_{P(\hat{E})} \ge [r \times \sigma_{P(\underline{E})}]$	Waagenar and Keren (1985); Erev, Wallsten and Budescu (1994).

Noise (Ê|M) between Memory and Êstimate: -Binary: Properties B and N_i. - Equidistant: Properties S and N, or S and U. -All others: Properties D and N.	**Confidence:** Based on a specific level of confidence, the confidence in judgments is too extreme	$0 \leq [r \times \sigma_{P(\hat{E})}] \leq \sigma_{P(M)}$	Adams and Adams, (1960); Lichtenstein and Fischhoff (1977); Lichtenstein, Fischhoff, and Phillips (1982); McClelland and Bolger (1994); Keren (1997); Fischer and Budescu (2005).
Noise (M|Ê) between Êstimate and Memory: -Binary: Properties B and N_i. - Equidistant: Properties S and N, or S and U. -All others: Properties D and N.	**Hard-easy**: Based on a specific level of task difficulty, the confidence in judgments is too conservative (conditionally inverse of the confidence bias)	$\sigma_{P(\hat{E})} \geq [r \times \sigma_{P(M)}]$	Lichtenstein and Fischhoff (1977); Keren (1988); Suantak, et.al. (1996); Juslin, et.al. (2000); Merkle (2009).

Source: Author.

Based on Table 1, it can be seen that the contributions of this article are three-fold:

Formalization of eight biases: This integrative framework allows us to unambiguously define each of the biases in mathematical terms, enabling clear distinctions between them. Until now, literature has often given different names to the same biases and the same names to different ones (Shah & Oppenheimer, 2008). Seminal papers in the field, like Erev et al. (1994), Dougherty et al., (1999), or Moore and Healy (2008), refer to different findings when using terms like conservatism, underconfidence, and overconfidence. One the one hand, this is nothing new in science, where conventional wisdom holds that "a scientist would rather use someone else's toothbrush than another scientist's nomenclature" (Gell-Mann, 1995a, p. 18). On the other hand, verbal formulations are often simply too slippery to clearly define the subject matter of interest. We should credit Galileo's insight (1623): "the book [of nature] cannot be understood unless one first learns to comprehend the language and read the letters in which it is composed. It is written in the language of mathematics […] without these, one wanders about in a dark labyrinth." (p. 238) I present clear-cut and unambiguous mathematical definitions of each bias, based on one common conceptual framework. I also give names to these biases, but the names are smoke and mirrors: It is the mathematical definition according to a common conceptual framework that defines the phenomenon, not an arbitrarily chosen verbal description.

Formalization of four noise-based generative mechanisms: This framework allows us to provide formalizations of the distinctive generative mechanisms that can produce these biases. The Appendices to this article present mathematical proofs showing that the presented kinds of noise must produce the corresponding biases. I also relate the identified kinds of noise to intuitively plausible psychological processes and propose some justifications for the identified kinds of mental noise.

Formalization of the relation between these biases: Given the use of a common conceptual framework and the mathematical formalization of biases and their generative mechanisms, I am able to formalize the relationships between and interdependencies among different biases and their potential causes, allowing the identification of limits to and trade-offs among these phenomena. The hope is that a better understanding of the relations between seemingly unrelated biases will not only clarify their existence by providing a solid theoretical foundation but also eventually enhance the search for coherent prescriptive strategies to confront, moderate, or remedy these psychological irrationalities.

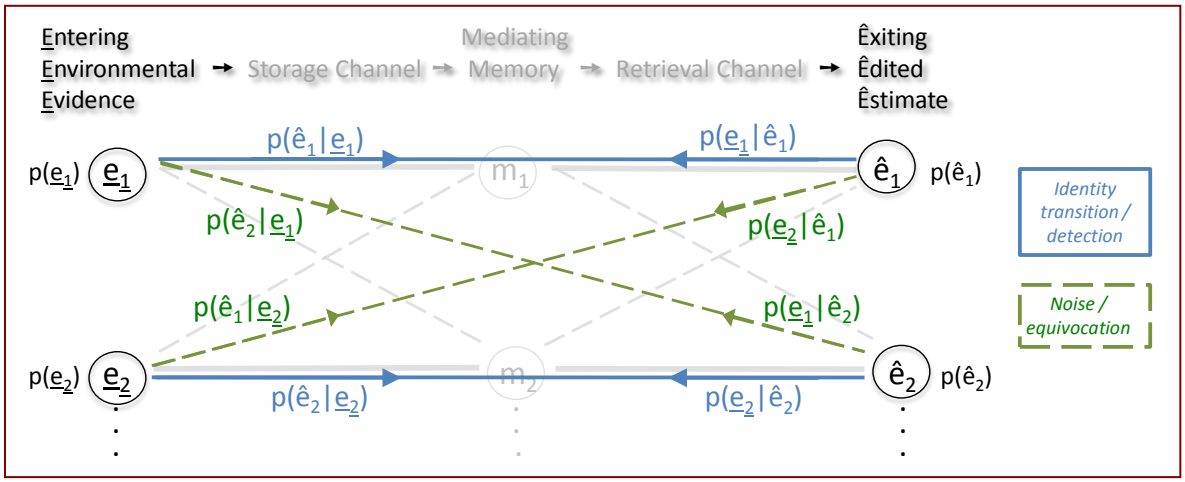
The theoretical background of this synthesizing framework is derived from the conceptual tools of information theory[1] (see the introductory texts Cover & Thomas, 2006, and Massey, 1998). The application of information theory to psychological decision-making goes back at least to Miller (1956), one of the most cited articles in the history of psychological research. However, information theoretic approaches to psychology have more recently been quietly abandoned and even declared to be an enterprise in vain (Duncan, 2003). While information theory has developed a large variety of ideas and concepts, I will focus on one very specific contribution: the rigorous analysis of the mechanisms involved when information is processed through a noisy channel (going back to Shannon, 1948; see also Chapter 4 in Massey, 1998; and Chapters 7 and 9 in Cover & Thomas, 2006). This set of tools will help us to explore key characteristics and boundaries of stochastic transformations from objective input evidence to subjective output estimate.

The Noisy Memory Channel

I term this conceptual synthesis the "noisy memory channel" (Figures 1 and 2). This kind of schematization of an information process might seem a little unconventional. For those readers who would like to warm up before getting more formal, I provide an introductory analogy in Appendix A. As intimidating as these figures might appear, this way of representing information processes turns out to be extremely useful when reasoning analytically about the properties of such processes.

The channel represents a probabilistic (stochastic) transform of one random variable into another. We use capital letters, like E, to range over all possible values e_1, e_2, … e_i of the random variable. Each of these realizations has a different probability of appearance: $p(e_i)$. The random variable "$\underline{E}$" represents the objective input evidence encountered by a decision-maker. This objective input evidence is transformed by the noisy memory channel into the subjective output estimate (represented by the variable $\hat{E}$, or "E hat") (Figure 1). More specifically, this transformation is intermediated by a transitional storage of the input evidence in some kind of memory (M) (Figure 2).

Figure 1: The overall noisy memory channel from evidence to estimate (binary choice)

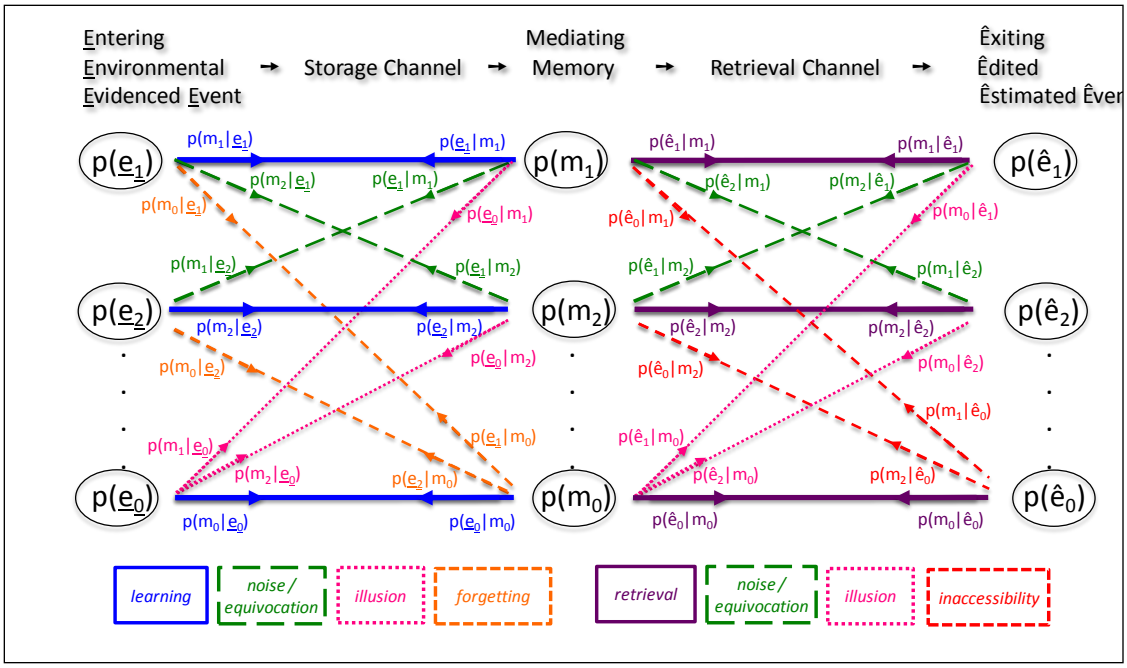


Source: Author.

Most computer simulations treat the storage and retrieval subchannels as one and the same channel and focus on the noisy transformation of objective evidence $\underline{E}$ into some distorted subjective estimate $\hat{E}$ (e.g., Budescu et al., 1997b; Erev, et al. 1994; Juslin et al., 1997; Merkle, 2009; Wallsten & González-Vallejo, 1994;). Figure 1 represents such a channel for a binary decision-making exercise—i.e., a choice between two options. The input for the memory channel usually originates from the external environment, as evidence or experienced events. Thus the different values of $\underline{E}$ can be any observable cognitive chunk, such as colors, words, ideas, concepts, events, or numbers, among many others. In a decision-making task, however, the number of values $\underline{e}_1$, $\underline{e}_2$, $\underline{e}_3$, etc. over which $\underline{E}$ ranges depends on the number of choices between which the judge must choose.

The transformation of $\underline{E}$ into the subjective estimate $\hat{E}$ depends on the conditional probabilities $P(\hat{E}|\underline{E})$: given evidence $\underline{e}_1$, what is the probability of obtaining estimate $\hat{e}_1$? In the best-case scenario, evidence $\underline{e}_1$ would be flawlessly converted into estimate $\hat{e}_1$, and $\underline{e}_2$ into estimate $\hat{e}_2$, etc. This is the case when $p(\hat{e}_1|\underline{e}_1) = 1$, or $p(\hat{e}_2|\underline{e}_2) = 1$, etc. We will call these kinds of horizontal transformations $p(\hat{e}_i|\underline{e}_i)$ "identity transitions" (or more precisely "mutual information transitions", see Massey, 1998, Ch.1; Cover and Thomas, 2006, Ch.2). Suppose, however, that as shown in Figure 1, there are several crossover possibilities, e.g. $p(\hat{e}_2|\underline{e}_1) \neq 0$ and/or $p(\underline{e}_1|\hat{e}_2) \neq 0$. These kinds of crossover transitions, $p(\hat{e}_x|\underline{e}_i)$ with $x \neq i$, are the kind of "noise" we are interested in in this article. This background provides us with a graphical and clear definition of noise as distortion and mixing.

Figure 2: The noisy memory channel consisting of storage and retrieval subchannels



Source: Author.

The storage of input evidence in memory creates another random variable, which we denote with "M." "Remembering," "learning," "perceiving," or simply "sensing" are processes that convert objective evidence E into memories M. We then retrieve our estimate from M to obtain our estimate Ê. Probabilistically, the overall channel from evidence to estimate, P(Ê|E)(Figure 1), is the product of the two transition matrices of the storage, P(M|E), and retrieval channel, P(Ê|M) (see Figure 2).[2]

The noisy memory channel is of analytical interest because it helps us to understand what can go wrong where, for example, given the probabilities of each realization of E, and of each transition probability P(Ê|E), we are able to calculate the probability of each value of Ê.[3] We can also analyze the channel from the point of view of our estimates, with conditional probabilities P(E|Ê) (sometimes referred to in information theory as "equivocation"). This perspective looks at the noisy memory channel from the left to the right and asks about the probability of the objective evidence E, given our estimate Ê. The "identity transition", $p(e_i|\hat{e}_i)$, detects the "hit rate" of the estimate, or how often a particular estimate is correct. To keep things simple, however, we will often refer to both P(Ê|E) and P(E|Ê) as noise. Of course, the attentive reader has already realized that both kinds of noise are related by Bayes' theorem: P(E|Ê) = [P(Ê|E) x P(E)]/P(Ê). In words: the probability of a wrong estimate is directly related to the noise of the memory channel (by Bayes' theorem), and vice versa.

In our application of this model to the first six biases (conservatism, Bayesian likelihood, illusionary correlation, placement, subadditivity and exaggerated expectation, see Table 1), it will not be necessary to worry about the precise route some evidence input $\underline{e}_i$ takes through memory M to be converted into an estimate $\hat{E}$. We will be concerned only with the transformation between evidence $\underline{E}$ and estimate $\hat{E}$ (Figure 1). To understand the other biases (the confidence bias and the hard-easy effect), memory becomes important (Figure 2). In reality, some kind of memory always plays a role in every judgment and decision-making process. Yet M might represent different kinds of memories, such as sensory-, working-, short-term, or long-term memory, episodic, semantic, etc. Nevertheless, because "information is physical" (Landauer, 1991), ((pg #))without any intermediate internal representation in some kind of storage (however short and unstable), information could not be processed.

Figure 2 also presents the psychologically special cases of illusions, inaccessibility, and forgetting, which are represented by $\underline{e}_0$, m_0 and $\hat{e}_0$.[4] Unfortunately, the empirical evidence analyzed in this article does not allow us to go deeply into the issues raised by forgetting and illusions. Nevertheless, they are part of the model as a whole. Furthermore, the model does not require that the number of values of the input variable $\underline{e}_1, \underline{e}_2, \ldots \underline{e}_y$ be the same as the number of values of the output variable $\hat{e}_1, \hat{e}_2, \ldots \hat{e}_z$ (i.e., it need not be the case that y = z, but for our purposes it is reasonable to assume that y = z).

In principle, channels can have millions of different input and output variables, each with different input-, crossover-, and deletion probability distributions. The transition probabilities and characteristics of these channels can quickly get incredibly complex and laborious to analyze (for the most common channels see Cover & Thomas, 2006, Chap.7; Massey, 1998, Chap.4). Appendix B, for example, shows that the MINERVA-DM model (Dougherty, et al. 1999; Hintzman, 1984, 1988;) is merely one special realization of the boarder logic presented in Figure 2. The goal is to find those properties that describe the human memory channel in agreement with empirically detected decision-making biases.

Noise in the Overall Memory Channel

Noise (Ê|E) between Evidence and Êstimate: Five Biases from one Generative Mechanism

I start my analysis of biases with those that can be generated by noise between the evidence and estimate (Ê|E) (see Figure 1). I show that some basic properties of this channel can generate five distinctively recognized empirical findings. All of them can be explained with what is known as "regression," "averaging," or "conservatism."

Formalizing the conservatism bias.

"Conservatism" refers to the experimental finding that people tend to underestimate high values and high likelihoods/probabilities/frequencies and overestimate low ones. As even this rough statement of conservatism indicates, I must distinguish between two kinds of input for our channel: the transformation of the values of the random variable (E, which can be nominal, ordinal, interval, or ratio) or the transformation of the respective likelihoods of occurrence (p(E)) (compare with Figures 1 and 2). I distinguish these two cases and make clear which one is under consideration, but in essence, conservatism works for any kind of measurable values and for likelihoods/probabilities/frequencies.

In a classic study, Kaufman, Lord, Reese, and Volkmann (1949) found that people tend to overestimate the number of dots that were flashed on a screen in a random pattern when only few dots are shown (between 5-10 dots) and underestimate the number of dots when many dots are shown (15-210 dots). In this case, the bias concerns the estimation of absolute values on an interval scale (number of dots, represented by the values of $\underline{e}_1$ and $\underline{e}_2$). Another example is that people overestimate the number of practicing physicians in Lane Country (subjective estimate: 456; versus objective evidence: 350), and underestimate the number of cigarettes consumed in the U.S. (subjective estimate: 1.5 billion; versus objective evidence: 604 billion) (MacGregor, Lichtenstein, & Slovic, 1988).

One limitation of the case of non-normalized numbers on an interval scale is that it is often difficult to say which values/numbers are "high" and which are "low". Is the number of cigarettes consumed in the U.S. high or low? Compared to China it seems low, but compared to Lichtenstein is seems high. Without a normalization scale it is often tricky to detect the conservatism bias for absolute values.

This insight provides a rationale for why the most straightforward and reliable studies on the conservatism bias are based on so-called likelihood, probability, or frequency estimates of some input, represented by the values of $p(\underline{e}_1)$ and $p(\underline{e}_2)$ (e.g., Greene, 1988; Hasher & Zacks, 1984; Hintzman, 1969; Howell, 1973; Zuroff, 1989; Fiedler, 1991). In these kinds of tasks, subjects do not estimate the value of $\underline{e}_i$, but its probability $p(\underline{e}_i)$ (compare Figures 1 and 2). Because probabilities are normalized between [0-1], it is straightforward to define "high" (close to 1) and "low" (close to 0) when estimating

likelihoods/probabilities/frequencies. For example, we overestimate the probability of rare causes of death and underestimate frequent ones (Fischhoff, Slovic, & Lichtenstein. 1977); we are conservative in estimating the likelihood of being male or female after being presented with the height of a person (DuCharme, 1970); and we are conservative in our estimates of how often specific letters appear in newspaper articles (Attneave, 1953). For these kinds of exercises the input random variable is not required to be an interval or ratio variable; it might also be categorical (such as causes of death, man or woman, colors, letters of an alphabet, etc).

Empirical studies distinguish between two different methods of obtaining $P(\hat{E})$: likelihood and probability estimations (e.g., what is the likelihood or probability of an event expressed as a percentage?) and frequency estimates (e.g., how often does an event occur?). Probabilities and frequencies are essentially related by the law of the large numbers (when numbers are large: "frequencies approximate probabilities"). Kahneman and Tversky (1982) referred to these modes of judgment as singular and distributional, respectively, and argued that frequency estimates usually provide more accurate results than estimates based on singular belief assessments (see also Tversky and Koehler, 1994). Of course, it is often not possible to count frequencies (what is the probability that the world will end tomorrow?). Notwithstanding these differences, for reasons of simplicity we will treat them both equally and refer to them as likelihoods in this article.

Traditionally, the input and output of a decision-making task are represented on a two-dimensional x/y-plane, such as in Figure 3a. Empirical studies of the conservatism bias have detected that subjective estimates ($\hat{e}_i$ for absolute values, or $p(\hat{e}_i)$ for likelihood estimates) are, on average, closer to their mean than objective evidences, e_i or $p(e_i)$, which are "spread out" toward their extremes. In other words, estimates are "conservative" with respect to the evidence, because the estimates' extremes are less accentuated and closer to their means.

Let us formalize this bias in mathematical terms. There are several measures that can be applied to quantify the notion of how distant measures are from their mean. The measure most widely used in psychology is variance, or its square root, the standard deviation. In these terms, conservatism holds that the standard deviation of the output estimate (typically depicted on the y-axis) is smaller than the standard deviation of the input evidence (on the x-axis): $\sigma_{\hat{E}} \leq \sigma_{\underline{E}}$.

As long as there is a positive correlation between evidence and estimates, the slope of the regression line between $\underline{E}$ and $\hat{E}$, based on the x-axis $\underline{E}$, is between 0 and 1 (see Figure 3a). This characteristic finding can be seen from the following transformations:

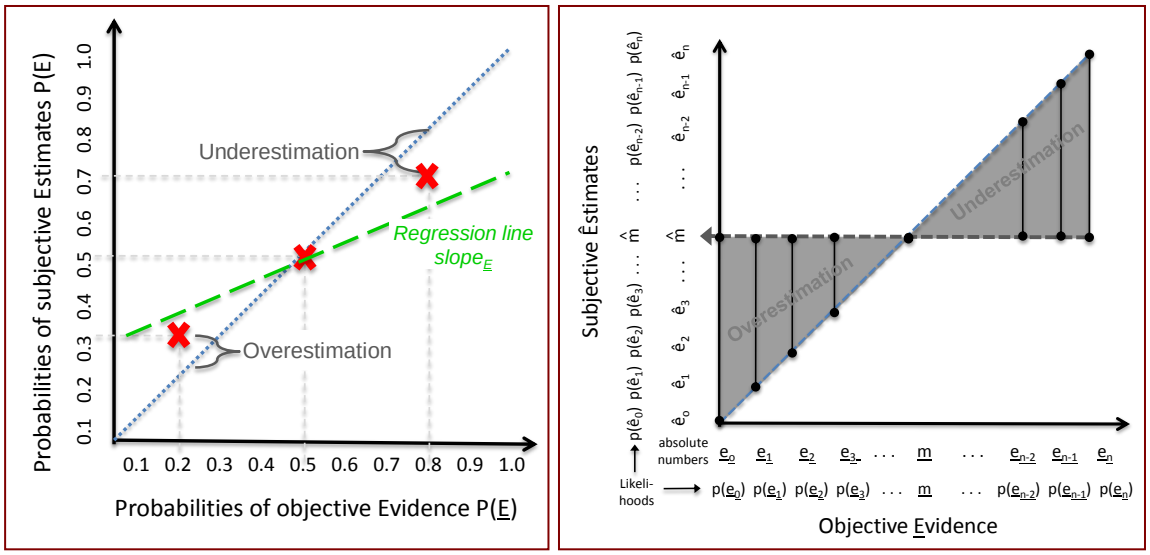
$$0 \leq \text{slope}_{\underline{E}} \leq 1 \qquad \text{(I)}$$

$$0 \leq [r^{x} \sigma_{\hat{E}}] / \sigma_{\underline{E}} \leq 1$$

$$0 \leq [r^{x} \sigma_{\hat{E}}] \leq \sigma_{\underline{E}} \qquad \text{(II)}$$

$$0 \leq \text{cov}(\underline{E},\hat{E}) \leq \text{Var}(\underline{E}) \qquad \text{(III)}$$

Figure 3: (a) Traditional representation of likelihood conservatism in a binary decision-making task; (b) Result of a memory channel that fulfills Properties N and S for (Ê|E) (bounded-noise conservatism channel); or Properties S and U for (Ê|E) (unbounded-noise conservatism channel).



Source: Author.

Equations (I), (II), and (III) are different reformulations of what will be our official definition of conservatism (as presented in Table 1). As long as the regression coefficient r is positive (r > 0), the inequality of equation (II) shows that $\sigma_{\hat{E}} \leq \sigma_{E}$ will always lead to $0 \leq \text{slope}_{E} \leq 1$ (because r is always ≤ 1), in agreement with the inequality of equation (I).

We are left with several alternative cases for which conservatism has been detected, including estimates of non-normalized numerical value inputs and estimates of likelihoods, as well as for selections among only two different choices (binary decision-making tasks, compare the set up in Figure 1) or among a multitude of choices (multiary decision-making tasks, compare Figure 2). Instead of going through all of the resulting four cases I will focus the present analysis on the case of binary likelihood estimates for the Bayesian likelihood bias—which provides conservatism in case of conditioned probabilities—and the case of pure conservatism among non-normalized numerical value estimates in a multiary decision-making task. Traditionally, these biases are treated separately in the literature, and I will follow this custom. Still, these biases are actually quite very similar.

Regressive Bayesian Likelihood: The Case of Binary Likelihood Estimates

Although rather simple, binary decisions are the most important case of human decision-making. The vast majority of empirical decision-making research has focused on binary choices (e.g., yes or no, right or wrong, man or woman, more or less likely, this or that, more or less, continue or stop, increase or decrease, back or forth, believe it or not,

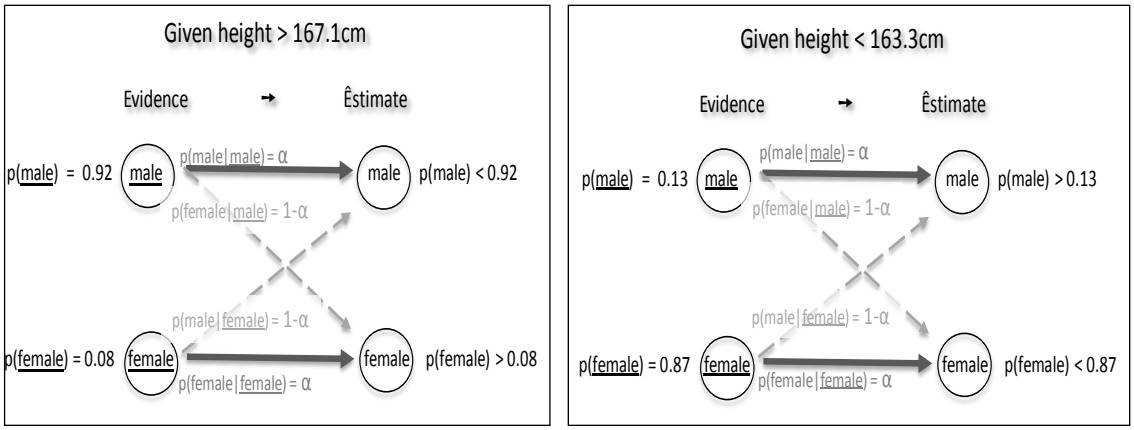
etc.). Many judgments naturally come down to binary choices, because all kinds of questions can be formulated as requests for an estimate about the event in question and its complement (event versus not-event).

The "Bayesian likelihood" or "Bayesian odds" bias refers to the case of estimating conditional probabilities (DuCharme, 1970; Edwards, 1968). This is conceptually identical to straightforward likelihood estimates discussed in the previous section; the sole difference is that the estimates are conditioned on some concretely defined event. The underlying logic with regard to conservatism remains the same. In controlled experiments, people are typically presented with some condition c_i and based on this condition asked to estimate a probability, $p(?|c_i)$. The results show that human estimates are conservative when compared with objective conditional probability, as calculated by definition of conditional probability, $[p(?|c_i) = p(?,c_i)/ p(c_i)]$, or alternatively through Bayes' theorem, $[p(?|c_i) = p(c_i|?) \times p(?) / p(c_i)]$. This is also the reason for the somewhat unfortunate name of the effect. In this regard it is interesting to point out that original studies of the phenomenon (Edwards, 1968; Phillips & Edwards, 1966;) simply referred to the bias as "conservatism in human information processing", which is more straightforward as, strictly speaking, Bayes' theorem is not necessary in this case.

In a classic study, DuCharme (1970) presented judges with the height of a person and asked them to estimate the person's gender : that is, to infer p(gender|height). We know that height and gender are related: men tend to be taller (back in the 1960s the average height of men was 173cm or 68') and women smaller (160cm or 63'). In agreement with the conservative Bayesian likelihood effect, judges tended to underestimate the number of tall men and overestimate the number of tall women. Dougherty et.al. (1999) have shown that this finding can be replicated with the MINERVA-DM application. The MINERVA-DM multi-trace memory model (based on Hintzman, 1988) essentially executes the logic of a so-called "binary symmetric channel" (BSC) in the retrieval channel (see Appendix B). In a binary symmetric channel, both identity transitions, and both noise transitions are equal, p(male|male) = p(female|female) = α; and p(female|male) = p(male|female) = 1 – α (see Figure 4; for more see Cover & Thomas, 2006, Ch.7; and Massey, 1998, Ch.4).

Figure 4 models DuCharme's (1970) findings with a binary symmetric channel between the two possibilities. In Figure 4a, the judgment is conditioned on people taller than 167.1cm. As argued by Dougherty et.al. (1999), this can be understood as limiting search in memory to a specific subgroup of memory traces. Within this identified subgroup of tall people, we know that there are many more men than women. According to DuCharme's objective evidence, 92% of the people taller than 167.1cm are men and only 8% are women. The channel transforms both likelihoods into values that lie "somewhere in between" 0.08 and 0.92. The estimates are "regressive" or "conservative". The same logic, in reverse, accounts for people with height smaller than 163.3cm (Figure 6b).

Figure 4: Bayesian likelihood bias: Conservatism of conditional probabilities modeled with a binary channel. (a) conditioned on people taller than 167.1cm; (b) conditioned one people smaller than 163.3cm.



Source: Author, based on DuCharme (1970)

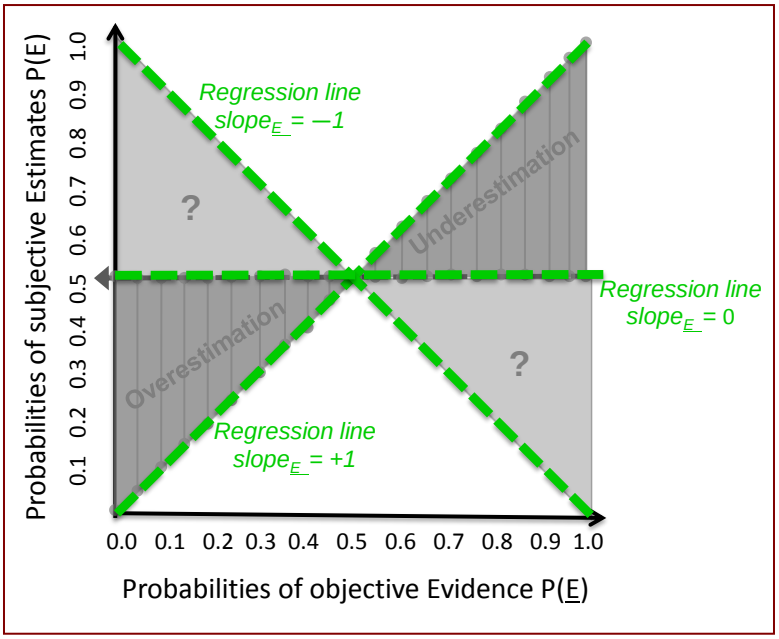
For our specific purposes, it does not matter if we represent the judge's memories as a multi-trace "episodic memory" (Baddeley et. al., 2009, Ch. 5), in which each event has its own memory trace (out of 100 memories of people taller than 167.1cm, the participant remembers that 92 are men), or as a form of "semantic memory" (Baddeley et.al., 2009, Ch.6), in which there are only two events with different weights (0.92 for men, and 0.08 for women) and the participant learned each weight (see also Fiedler, 1996).

Channel Properties for Binary Decision-making Tasks

Does the binary symmetric channel always produce conservatism? Does it matter which kinds of values we choose for the identity and noise transitions? To answer these questions, we compare the traditional representation of conservatism on the x-y plane (Figure 3) with the channel logic (Figure 4). If the identity transitions of Figure 4 would carry all the weight, p(male|male) = p(female|female) = α = 1, the resulting estimates would be located along the dashed diagonal 45° line in Figure 3a (which implies that $slope_{E}$=1). Deviations would exclusively depend on the variance and the sample size of the objective evidence and the sampling from memory, while deviations would cancel each other out over large numbers (see also Fiedler, 1996). If the identity transition and the crossover noise transition had the same weight, so that α = 1 – α = 0.5, input evidence and output estimate would be independent, and the resulting regression line would have $slope_{E}$ = 0. In this case, the specific input value of the evidence E does not tell us anything about the specific output value of the estimate Ê. In information theoretic terms we would say that there is no "mutual information" between evidence and estimate: the channel does not transmit information and is at its highest state of uncertainty (maximum entropy).[5] If all weight were placed on the crossover noise transitions, p(female|male) = p(male|female) = 1 – α = 1, a given input would be perfectly transformed into its opposite: we would consider that all

males are females and vice versa. This transformation would suggest that the human mind is a system that perfectly confuses things. In terms of the traditional representation of Figure 3, this would imply $slope_{\underline{E}} = -1$ (see Figure 5).

Figure 5: Possible outcomes of the binary symmetric channel



Source: Author.

This work shows that the binary symmetric channel will produce estimates that lie inside all (light and dark) grey areas in Figure 5 simply by being symmetric. Typical empirical findings do not show estimates inside the light grey areas but rather in the darker ones. We have defined conservatism with equation (I) as $0 \leq slope_{\underline{E}|c} \leq 1$ (corresponding to the dark areas in Figure 5), or the conditioned version of equation (II), $0 \leq [r \times \sigma_{P(\hat{\underline{E}}|c)}] \leq \sigma_{P(\underline{E}|c)}$, with c being some overall conditioning event for the task (compare our formal definition of the Bayesian likelihood bias in Table 1). Thus, there is a positive correlation between objective evidence and subjective estimate, not a negative one: the smaller of both inputs stays smaller, and the larger one stays larger. This condition is satisfied only if the identity transition is larger than the crossover noise transition, with $0.5 \leq \alpha \leq 1$: on average we are more right than wrong.

From a psychological perspective, it seems reasonable and intuitive to assume that the identity transition is larger than the noise transition. If this were not the case, we would on average classify men as women and vice versa. Our predictions and expectations would persistently be in contradiction with reality, and we could not make much sense of the world around us. This does not seem to be the case. We can therefore model the conservatism bias for binary decision-making tasks with a binary symmetric channel in

which the identity transition is larger than the crossover noise. We thus end up with the following two properties of our binary channel:

Property N_i: (*More right than wrong*). The identity transition probability is larger than or equal to each of the noise transition probabilities: $p(\hat{e}_i|\underline{e}_i) \geq p(\hat{e}_x|\underline{e}_i)$, for all $x \neq i$.

Property B: (*Binary symmetric*). The channel is a binary symmetric channel: $p(\hat{e}_1|\underline{e}_1) = p(\hat{e}_2|\underline{e}_2)$.

Linking Property N_i to psychological processes implies that evolution has provided us with an information processing design in which the correct connection weighs more than any wrong connection: the way we are "wired" makes sense. We might systematically fail in the case of specific tasks (such as trick questions and illusions), and patients with cognitive dysfunctions might persistently confuse inputs, but these seem to be the exceptions rather than the norm. In general, evolution has endowed us with an impressively accurate and remarkably well tuned information processing system.

Property B states that both input evidences are affected by noise with the same intensity. This property also seems plausible psychologically, especially if one considers that in well-defined binary decision-making tasks there is actually only one event, with the "other event" serving as its complement (not-the-event, or "everything else"). Property B supposes that both are distorted equally. Property B is also backed up empirically. For example, DuCharme (1970) reports no bias for estimates about similarly sized men and women (both between 167.1-163.3cm, for which 51% are men and 49% are women). This result (50%-50% in, and 50%-50% out) is characteristic for a binary symmetric channel (Cover and Thomas, 2006: Ch. 7; Massey, 1998: Ch4), which confirms our choice of model.

Conservatism: The Case of Equidistant Interval Estimates

The first example focused on a binary decision-making exercise that estimated the likelihood of an event, $p(e_i)$. Empirical studies show that we are also conservative in our estimates of the value of a non-normalized numerical variable itself, e_i, even if we have more than two choices. In this case we will refer to an exercise that, unlike the Bayesian likelihood bias, does not depend on previous conditioning. So instead of asking "given A, what is B?" we simply ask "what is B?". As will be seen, the same conservative logic applies even without existence of a conditioning variable.

In an experiment with a very reliable sample size, Hockley (1984) analyzed 47,120 judgments on the repetitions of listed words (Experiment 1, p.230). The presented words were repeated up to three times. The incoming probability distribution was close to uniform (which makes the math easier, but is not required in our setup). The subjects tended to overestimate the number of repetitions for low numbers (the mean estimation for 1st repetitions was 1.03) and underestimate the number of repetitions for high numbers (the mean estimation for 3rd repetition was 2.72). Figure 6a shows the finding in its

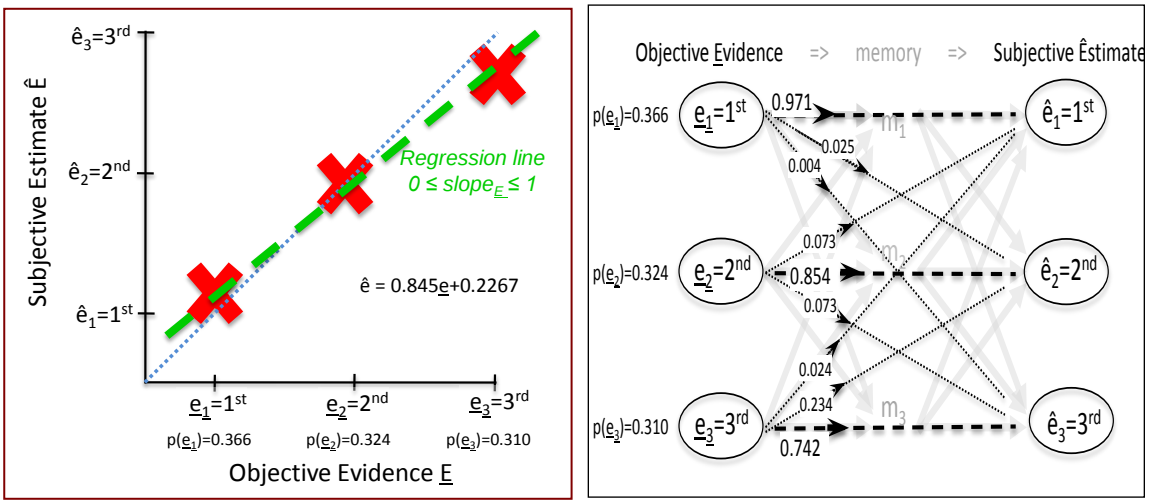
traditional presentation, which reveals conservatism in agreement with our definition (equation I).

Hintzman (1988) simulated Hockley's findings with the computer program MINERVA2. Given that MINERVA basically follows the logic of a noisy memory channel (see Appendix B), we should be able to analyze it with our information-theoretic tools. Figure 6b shows the same result as Figure 6a, but this time in the representational form of the noisy memory channel.[6] The Figure cites the values of Hockley's Table 1 (1984: p.230), which shows the noise P(Ê|E) of the channel. Subjects tended to include "false friends" in their estimates. When a word was presented for the 3rd time ($\underline{e}_3$), people identified it as a third repetition only 74.2% of the times: 23.4% of the time they estimated it to be a 2nd repetition, and 2.4% of the times they thought the word appeared for the 1st time. On the basis of these results, we can calculate the expected value of the estimate Ê given a 3rd repetition $\underline{e}_3$:

Expected value of the estimate Ê, given a 3rd repetition $\underline{e}_3$ =

$$= [0.024^{x}1 + 0.234^{x}2 + 0.742^{x}3] = 2.72 = [p(\hat{e}_1|\underline{e}_3)^{x}\,\hat{e}_1 + p(\hat{e}_2|\underline{e}_3)^{x}\,\hat{e}_2 + p(\hat{e}_3|\underline{e}_3)^{x}\,\hat{e}_3]$$

Figure 6: Conservatism: (a) traditional representation; and (b) memory channel representation



Source: Author, based on Hockley (1984).

We can now use all existing theorems and rules of probability to analyze additional properties of the channel.[3] For example, using the total probability theorem we can calculate the probability of each estimate, P(Ê) (how probable is it that subjects estimated a 1st repetition). Through Bayes' theorem we can calculate the equivocations, P(E|Ê) (given a specific estimate, what is the probability that a specific objective input evidence was presented?).[7]

In this sense, conservatism can be understood as the fact that "false friends" sneak into our estimates, due to noise in the memory channel, which confirms the notion of noise as mixing. Hockley (1984: p237) performed the same experiment with up to 5 word repetitions (sample size: 45,760), which led to a similar conservative result (see Table 3).

Channel Properties for Equidistant Decision-making Tasks

Hockley's exercise is not binary but ternary. Therefore, our property B does not apply. We could demand that the ternary channel be symmetric, i.e. that all three evidences are affected by the same kind of noise. However, this is not in agreement with Hockely's (1984) findings. For example, whereas 1st and 2nd as well as 3rd and 2nd repetitions are equally distant from each other (1Δ apart), it is much more probable to confuse a 2nd and a 3rd repetition, with $p(\hat{e}_2|\underline{e}_3)$= 23.4 %, than a 2nd and a 1st repetition, $p(\hat{e}_2|\underline{e}_1)$= 2.5 % (see Figure 6b). This result suggests, plausibly enough, that it is easier for people to recognize when a word appears for the 1st time than when it appears for a 2nd or 3rd time. The technical conclusion is that different inputs seem to be affected by noise of different intensities. We therefore relax our requirements and basically demand that judges "confuse equally different things, equally likely":

Property S: (*identity-symmetric noise*). Noise is symmetric around the identity transition for all defined values $e_{i\pm j}$: $p(\hat{e}_{i-j}|\underline{e}_i) = p(\hat{e}_{i+j}|\underline{e}_i)$.

This property is in agreement with Hockley's empirical findings. Being presented with a word for the 2nd time, it is as probable that people erroneously think that the word appears for the 1st or a 3rd time: $p(\hat{e}_1|\underline{e}_2)= p(\hat{e}_3|\underline{e}_2)$= 7.3 %. It is important to note that not all empirical findings reconfirm this property (e.g., Hockley's five word repetition exercise deviates slightly; compare also with Table 3). Nevertheless, I will stick to the simplified assumption of identity-symmetric noise for modeling purposes.

For this kind of decision-making exercise that draws from a equidistant interval scale (such as 1, 2, 3, etc. all 1Δ apart), it turns out that satisfying Property S is enough to fulfill the right side of the double inequality in our definition (I) for conservatism (for a formal proof, see Appendixes C and E). For ternary input, satisfying Property S is also enough to fulfill the left side of definition (I). The left side of our conservatism definition (equation I), however, is not assured for decision-making exercises on an equidistant scale with more than three inputs. These distributions have to be quite particular, but they do exist.[8] I therefore return to the psychologically pleasing Property N_i and extend it to the multiary case. I continue to demand that our estimates are on average "more right than wrong" and additionally postulate that we are more likely to confuse "more similar things" than we are "less similar ones":

Property N: (*single-peaked unimodal noise*). The transition probabilities get smaller the larger the distance between a noisy estimate and the identity estimate: $p(\hat{e}_x|\underline{e}_i) \geq p(\hat{e}_z|\underline{e}_i)$, for all $|x-i| \leq |z-i|$.

Because the case x = i defines the weight of the identity transition (which by the definition of Property N is the largest transition), Property N subsumes Property N_i, which thus becomes redundant.[9]

Hockely's (1984) empirical findings also confirm Property N. For example, he detected that identity transitions are most likely and that it is more probable to confuse a 1st and a 2nd repetition (which are more similar, $p(\hat{e}_2|\underline{e}_1)$= 2.5 %), than a 1st with a 3rd (which are more distinct, $p(\hat{e}_3|\underline{e}_1)$= 0.4 %) (see Figure 6b; similar for his quinary five word repetition exercise, see Table 3).

Property N states that we are more likely to confuse something with "something similar" than with something "less similar" (more or less similarity between x and z in the formulation of Property N). This raises the question of what defines "similar" and "dissimilar". This question is trickier to answer for nominal categorical or ordinal variables (such as the estimation of a color, etc.). It is quite straightforward to the presented case of word repetitions, since 1, 2, 3, 4 etc. consist of an equidistant interval scale. 3 is more different from 1 than 2. The proof in Appendix C shows that the combination of Properties S and N produces conservatism for all decision-making exercises that focus on evidence taken from an equidistant interval scale (more formally, for which: $\hat{e}_i = \hat{e}_o + i\Delta$, for some constant Δ with i = {0, 1, 2, 3 … n}). The objects are equidistant because each of them is an equal distance Δ apart from its next valid neighbor (e.g. 1, 2, 3…; or 10 %, 20 %, 30 %...).[10] The importance of the requirement for an equidistant scale might not be clear at this point, but the details of the proof in Appendix C shows that this is an important requirement to assure that Property N produces conservatism.

With this in mind, we now know that Properties S and N define a channel that must produce conservatism for estimates drawn from an equidistant interval scale. The "turning point" between over- and underestimation is defined half way between both extremes, the midrange point $m = [\underline{e}_0+\underline{e}_n]/2$ (see Figure 3b). For values with equal weights, the midrange point is equal to the mean when all values on the equidistant interval scale are applicable: $\sum_{i=0} e_i /[n+1] = m = [\underline{e}_0+\underline{e}_n]/2$.[11] For example, for Hockely's (1984) ternary exercise: [1+2+3]/[2+1] = 1+([3-1]/2) = 2 (compare Figure 6a).

This argument shows that we can use the same underlying logic to explain the existence of two biases traditionally treated separately in the literature: conservatism and Bayesian likelihood. It turns out that we can use the same logic to explain additional biases.

Illusory Correlation of Minority Stereotyping

Illusory correlation is a bias in which one's judgments are based on a relation one expects to see even when no such relationship exists. One socially very relevant and delicate case of illusionary correlation is that decision-makers form false associations between people with rare (typically negative) behaviors and membership in statistical minority groups (Hamilton, Dugan, & Trolier, 1985;Hamilton & Gifford, 1976; Jones et al., 1977;

Mullen & Johnson, 1990 Pryor; 1986; Spears, van der Pligt, & Eiser, 1985,1986;for its role in stereotyping and discrimination see Bar-Tal, Graumann, Kruglanski, &Stroebe, 1989). The typical empirical setup (based on Hamilton & Gifford, 1976) assumes a 2 by 2 matrix of subjects, distinguishing between majority and minority groups and between positive and negative behavior (Table 2). Studies indicate that the smallest group is overestimated and that the largest group is underestimated. Because the minority is part of the smaller group (by definition), the result is in an exaggerated impression of the behavior of the minority, which students of this phenomena have interpreted as stereotyping, discrimination, or a illusionary correlation between group membership and behavior.

Most studies conclude that this bias is produced by the availability heuristic (Tversky & Kahneman, 1973; 1974), claiming that decision-makers are good at recognizing distinctive minority behavior. However, both Fiedler (1991) and Smith (1991) have shown that there is a simpler explanation for this bias: statistical base rates alone may lead to discrimination against minorities. Smith (1991) simulated this bias with the help of MINERVA (Hintzman, 1984; 1988), which does not include any mechanism of availability that fosters the distinctiveness of any group (see Appendix B). We will follow this idea and show that simple noise can produce the bias without the need for a more sophisticated availability heuristic.

Table 2 shows the traditional set up, with the numbers representing the size of the groups. The underlined numbers show the objective evidence $\underline{e_i}$, the arrow in brackets the tendency of the judgment [underestimation ▼; or overestimation ▲], and the subsequent number the obtained subjective estimate $\hat{e}_i$. Note that the distribution of evidence is independent among the 2 by 2 matrix (relationship 18/9 = 8/4; and 18/8 = 9/4)—there exists no relation between group membership and behavior. Hamilton and Gifford (1976) did two experiments. In Experiment 1 they assumed a prevalence of positive traits for both the majority and minority (in agreement with Kanouse *&* Hansen, 1972). In Experiment 2 they assumed a prevalence of negative behavior. In both experiments, they asked two questions, one about group membership, given the trait (read the left two columns of Table 3 from left to right: given the trait as input, which of the two outputs: majority or minority?), and one about traits, given group membership (read the right two columns of Table 3 from top down: given the group's size as input, which of the two outputs: positive or negative trait?).

As shown by the triangle arrows in Table 2, the direction of the question (i.e., the conditioning on group membership or trait) seems to influence the bias. In both Experiments 1 and 2, when asked about group membership, subjects underestimated the majority and overestimated the minority. When asked about traits, subjects underestimated the larger group and overestimated the smaller group. It is striking that these findings appear in both Experiment 1 and 2. In other words, it does not seem to matter whether the traits are positive or negative. Rather than being an indication of discrimination and

stereotyping, one can interpret this finding purely in terms of group size and base rates (compare with Fiedler, 1991; Smith, 1991).

Table 2: Empirical findings of illusory correlation of minority stereotyping: evidence $\underline{e_i}$, [tendency: underestimation ▼; or overestimation ▲], and estimate $\hat{e}_i$.

(Experiment 1) **More positive than negative traits**					
	Majority or minority?			Positive or negative?	
	Majority	Minority		**Majority**	*Minority*
Positive	**18 [▼] 17.5**	**9 [▲] 9.5**	Positive	**18 [▼] 17.1**	*9 [▼] 7.3*
Negative	*8 [▼] 5.8*	*4 [▲] 6.2*	Negative	**8 [▲] 8.9**	*4 [▲] 5.7*
(Experiment 2) **More negative than positive traits**					
	Majority or minority?			Positive or negative?	
	Majority	Minority		**Majority**	*Minority*
Positive	**8 [▼] 5.9**	**4 [▲] 6.1**	Positive	**8 [▲] 8.2**	*4 [▲] 6.6*
Negative	*16 [▼] 15.7*	*8 [▲] 8.3*	Negative	**16 [▼] 15.8**	*8 [▼] 5.4*

Source: Hamilton and Gifford (1976).

This bias can be therefore be explained with the same noise-based generative mechanism as the conservative Bayesian likelihood bias. The conditioning variable in this case is the cross-tabulating second group of attributes. Formally, $0 \leq [r \times \sigma_{\hat{E}|c}] \leq \sigma_{\underline{E}|c}$, and $0 \leq [r \times \sigma_{P(\hat{E}|c)}] \leq \sigma_{P(\underline{E}|c)}$ for likelihood estimates, or $0 \leq [r \times \sigma_{P(\hat{E}|c)}] \leq \sigma_{P(\underline{E}|c)}$ for non-normalized numerical estimates, with c being a cross-tabulated event (see Table 1). In the case of a binary exercise (choice between two classes) we have shown that Properties B and N_i produce this result. In a decision-making exercise that would estimate the likelihood (or frequency) of more than two choices (e.g., positive, neutral, negative, and horrible behavior), we have shown that the related Properties S and N define the kind of noise that produces this empirical finding.

Self-other Placement

Another seemingly unrelated bias can be explained by assuming the very same mechanism of conservative noise ($\hat{E}|\underline{E}$) between evidence $\underline{E}$ and estimate $\hat{E}$. The placement-effect is often simply called "overplacement," sometimes known as the "better-than-average" and "worse-than-average" effect. It is based on the empirical finding that we tend to believe ourselves to be better than others at tasks at which we rate ourselves above average (Kruger and Dunning, 1999), and tend to believe ourselves to be worse than others

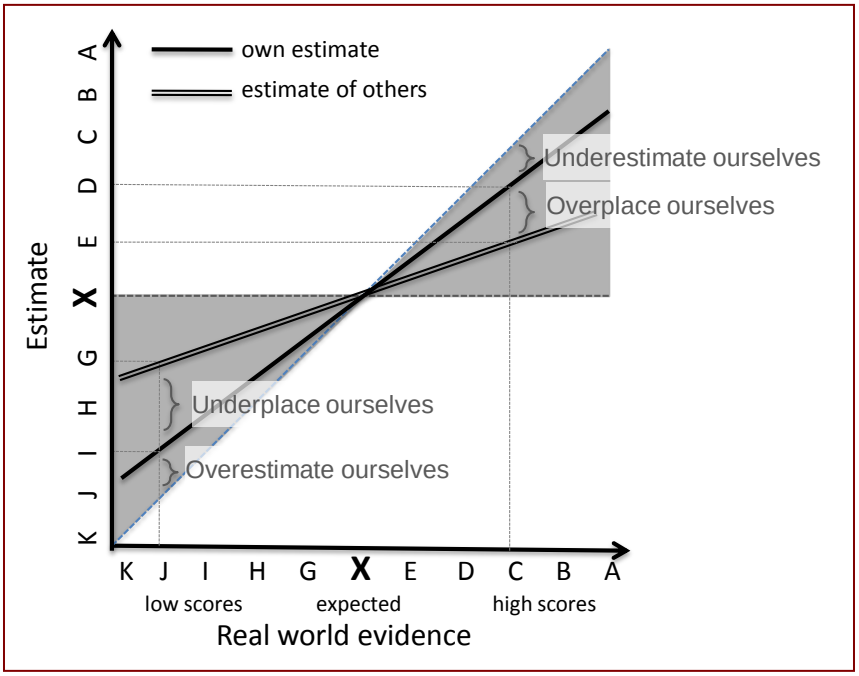
at tasks at which we rate ourselves below average (Kruger, 1999). What should this bias have to do with the previously analyzed logic of conservatism?

One potential explanation is based on the assumption that we have better information about ourselves than about others (Sande, Goethals, & Radloff, 1988). In a convincing presentation, Moore and Cain (2007) and Moore and Healy (2008) have shown that this effect can be traced back to the fact that

> "people often have imperfect information about their own performance, abilities, or chance of success. However, they have even worse information about [the performance] of others. As a result, people's estimates of themselves are regressive [conservative], and their estimates of others are even more regressive [conservative]. Consequently, when performance is high, people will underestimate their own performances, underestimate others even more so, and thus believe that they are better than others. When performance is low, people will overestimate themselves, overestimate others even more so, and thus believe that they are worse than others" (Moore & Healy, 2008: p.503)

Figure 7 illustrates the placement effect in a way similar to the traditional representation of conservatism. The example presents 11 equidistant test scores from the worst score K to the best score A (for example score K=0, J=1, I=2, … B=9, A=10). Therefore, the example is based on an 11-ary decision-making exercise that results in conservatism and can be modeled with the help of a noisy memory channel that satisfies Properties S and N.

Figure 7: Self-Other-Placement



Source: based on Moore and Healy (2008: p.504)

With no other evidence, and with total uncertainty, the guess with the greatest chance of success would be to estimate that everybody receives a midrange score of X (see also the principle of maximum entropy, Jaynes, 1957a, 1957b). Upon taking the test, we receive some evidence about our performance (we have a "feeling about how it went"); thus, our uncertainty is no longer at the maximum. This evidence tells us that we are below, above, or near the score that we had initially expected. If our estimate of this score were perfect, our estimation would be placed somewhere on the diagonal 45° line; it would be in agreement with reality. But our estimates are noisy. When reasoning about our own performance, we mix the evidence of our score with the initially expected score X, the only two pieces of evidence we have. The result is a binary choice between the prior and the updated evidence, which results in conservatism (solid black line in Figure 7).

When we estimate the scores of others at the same time, we use the same two sources. In this case, however, the evidence of our personal score is much less influential. The initially expected score X continues to weigh more heavily in our estimates of the performance of others. According to Moore and Healy (2008), if we perform better than expected (above X), we feel that the test was easier than we expected, and will readjust our estimates about others' performance upward. The argument goes the other way around if we perform below our initially expected score. Given, however, that our new evidence is based exclusively on our own performance, it influences the readjustment of our estimates of own performance more than the readjustment of the estimates of others' performances. This implies that our estimates of the scores of others are even more conservative (more influenced by the previous expectation) than our estimates of our own performace (more influenced by the new evidence received after giving the test), and the resulting line has a stronger inclination toward the horizontal line (double solid line in Figure 7). Formally, $0 \leq \text{slope}_{\underline{E}\ \text{others}} \leq \text{slope}_{\underline{E}\ \text{own}} \leq 1$ (see Table 1). Reformulating Moore's and Healy's (2008) logic, we can say that our estimates about our own performance are less noisy (more accurate) than our estimates about the performance of others (more formally, our estimates of others' performance have higher "entropy", see Cover & Thomas, 2006, Ch. 2Massey, 1998, Ch. 1). This finding is intuitively pleasing, as it simply restates that we know more about ourselves than about others. As illustrated by Figure 7, this leads to the well-known over/underplacement effect.

Subadditivity

The conservative noisy memory channel can provide a possible explanation for yet a fifth seemingly unrelated bias. The "subadditivity effect" refers to the empirical finding that an estimate of a likelihood is normally smaller than the sum of the estimates of each of its (more than two) mutually exclusive components (Fox & Levav, 2000;Tversky & Koehler,

1994;). Fiedler (1991, his Demonstration 3) shows that something similar holds for non-normalized absolute value frequency estimates: the summed estimates of two component estimates are higher than the compound frequency. Bearden and Wallsten (2004) have shown that the subadditivity effect can be computationally simulated with MINERVA-DM (see Appendix B) and Fiedler, Unkelbach, and Freytag (2009) have provided an additional and seemingly independent interpretation of subadditivity produced by an imperfect (noisy) transformation from evidence to estimate.

Formally, I describe the subadditivity bias as $p(\hat{e}_i) \leq \Sigma\ p(\hat{e}_d)$, with d being a decomposition of event i (see Table 1). For example, Redelmeier, Koehler, Liberman, & Tversky (1995) asked physicians to provide probabilities for the likelihood of four exclusive and exhaustive survival chances of patients: dies during hospitalization, dies within a year after release, lives between 1-10 years, lives more than 10 years. Subjects were confronted with each of those alternatives, making four separate judgments. The sum of the likelihood estimates should add up to 100 %, but was instead equal to 164 % (Redelmeier et al., 1995). The decomposed estimate is therefore said to be subadditive with 100/164 = 0.61. In a similar exercise, Witteman, Renooij, and Koele (2007) detected that the subadditivity effect increases with the level of unpacking: the decomposition into three alternatives led to a sum of 120 % on average, while the decomposition into six alternatives led to a sum of 180 % on average (see also Fiedler et al., 2009). Additionally, Fiedler et al. showed that the degree of subadditivity depends on the extremity of the input evidence.

Following the logic outlined in our discussion of the Bayesian likelihood bias (see Figure 4), the judge makes a series of binary choices on one single event: the scenario in question and its complement, which encompasses all other possibilities. For binary choices, the turning point between under- and overestimation is expected to be at $\sum_{i=1} p(\underline{e}_i)/n = 1/n = ½$. Dividing the general event into multiple smaller parts, must leave more choices below this mark than above. In other words: most of the decomposed subevents can be expected to have a probability less than 50 %. Following the logic of Figure 3, all estimates with likelihood values p(event) < 0.5 will result in overestimates. The sum of a series of overestimated components is of course larger than the total.

This model can also explain the effect of the level of unpacking on subadditivity, because the overestimation of a very small input will be larger when mixed with very high input, given the same level of noise. Property N (*single-peaked unimodal noise*) tells us that increasing similarity between events (moving them "closer together" on a one-dimensional scale), while keeping noise constant, should increase the subadditivity effect, which is in agreement with Tversky and Koehler (1994, see also the argument in Bearden & Wallsten, 2004). Our model also predicts that increasing noise should increase the effect, as more of the large complement is mixed into the estimation of the small subevent. This formulation is in agreement with the empirical findings by Bearden and Wallsten (2004) and Fiedler and Armbruster (1994).

The noisy memory channel model also affords us a possible explanation for another interesting finding that Tversky and Koehler (1994) outlined in their related support theory: "judged probabilities are complementary in the binary case and subadditive in the general case." (p. 547) . Soliciting estimations for both events of a binary choice (e.g., survives or dies), a symmetric binary channel will overestimate the lower value, p(event) < 0.5 but will also underestimate the higher value, p(event) > 0.5. The sum of an equally overestimated and underestimated value stays equal, which explains why no subadditivity effect is detected in the binary case (see also Redelmeier et.al., 1995). This conclusion suggests that conservative noise between the objective evidence and the subjective estimate is sufficient to produce this finding, without the explicit need for any additional heuristic. In the words of Fiedler et al. (2009, p. 383): "Although other factors may contribute to subadditivity, their influence needs to exceed the baseline expected from the [conservative] regression model alone".

Noise ($\underline{E}|\hat{E}$) between Êstimate and Evidence: Exaggerated Expectation

Until now I have analyzed the overall channel between evidence $\underline{E}$ and estimate $\hat{E}$ conditioned on the evidence ($\hat{E}|\underline{E}$). Bayes' theorem allows us to also look at the channel from the side of the estimates, ($\underline{E}|\hat{E}$). In a much-cited article, Erev et.al. (1994) observed that this way of conditioning results in a bias similar to conservatism, but looked at the other way around: not from the evidence to the estimate, but from the estimate to the evidence. The bias can be detected with the same data set as the conservatism bias. The key to understanding the difference between them lies in the conditioning variable. Erev et.al. simulated the introduced errors with normally distributed noise and were able to replicate both phenomena simultaneously.

In mathematical terms, the bias implies that the variance of the evidence ($\underline{E}$) is smaller than the variance of the estimate ($\hat{E}$). Turning equation (II) from above around, this implies that $\sigma_{P(\hat{E})} \geq [r \times \sigma_{P(\underline{E})}]$. In order to be able to clearly distinguish this bias from others (and because terms like "confidence" and "conservatism" are often used very loosely in the literature), we will refer to this bias as the "exaggerated expectation" bias, simply to give it a name (compare with Table 1). It is the same as the conservatism bias when the memory channel is looked at not from the side of evidence to the estimate (noise), but from estimate to evidence (equivocation).

The exaggerated expectation effect says that given a high subjective estimate, the mean objective value is not high enough, and given a low subjective estimate, the mean objective value is higher than the estimate. This means, for example, that in empirical findings that state that of all times in which we expect high-frequency events, $p(\hat{e}_i) > 0.5$, events are on average less frequent than we would expect, ($E[P(\underline{E})|p(\hat{e}_i)] < p(\hat{e}_i)$—and vice

versa for low expectations. It also means that given the expectation of a low return/grade/score, on average we receive higher returns/grades/scores than expected. The opposite is true for high expectations.

Exaggerated expectation occurs when the equivocation side of the noisy memory channel, $P(\underline{E}|\hat{E})$, is governed by properties similar to those governing its noise-side, $P(\hat{E}|\underline{E})$. Conditioned on $\hat{e}_i$, instead of on $\underline{e}_i$, Property N_i (*more right than wrong*) would claim that the identity detection $p(\underline{e}_i|\hat{e}_i)$ is larger than any of the equivocations $p(\underline{e}_i|\hat{e}_x)$, for all $x \neq i$. In other words, a specific estimate tells us more about the reality than about any other wrong evidence. The same conditional inverse can be applied for Properties S and N and the respective proof in Appendix C. This results from the fact that the regression line based on the estimate $\hat{E}$ (the y-axis of the traditional form of representation) has a slope of: $1 \leq \text{slope}_{\hat{E}}$. Remember that each correlation has two regression lines: one conditioned on the a-axis (which is the most commonly used, such as in the confidence-bias), and another one conditioned on the y-axis (see Freedman, et.al. 2007: p.174; Furby, 1973). Thus, the exaggerated expectation bias is clearly distinct from conservatism or the confidence bias, because it supposes noise from the estimate to the evidence, not the other way around.

Noise in the Retrieval Sub-channel

Until now I have worked with the overall noisy memory channel (see Figure 2), which treats the noisy storage and retrieval sub-channels as one single process. I now open up this black box (see Figure 3) to investigate a group of biases based explicitly on the judge's memory and the retrieval sub-channels.

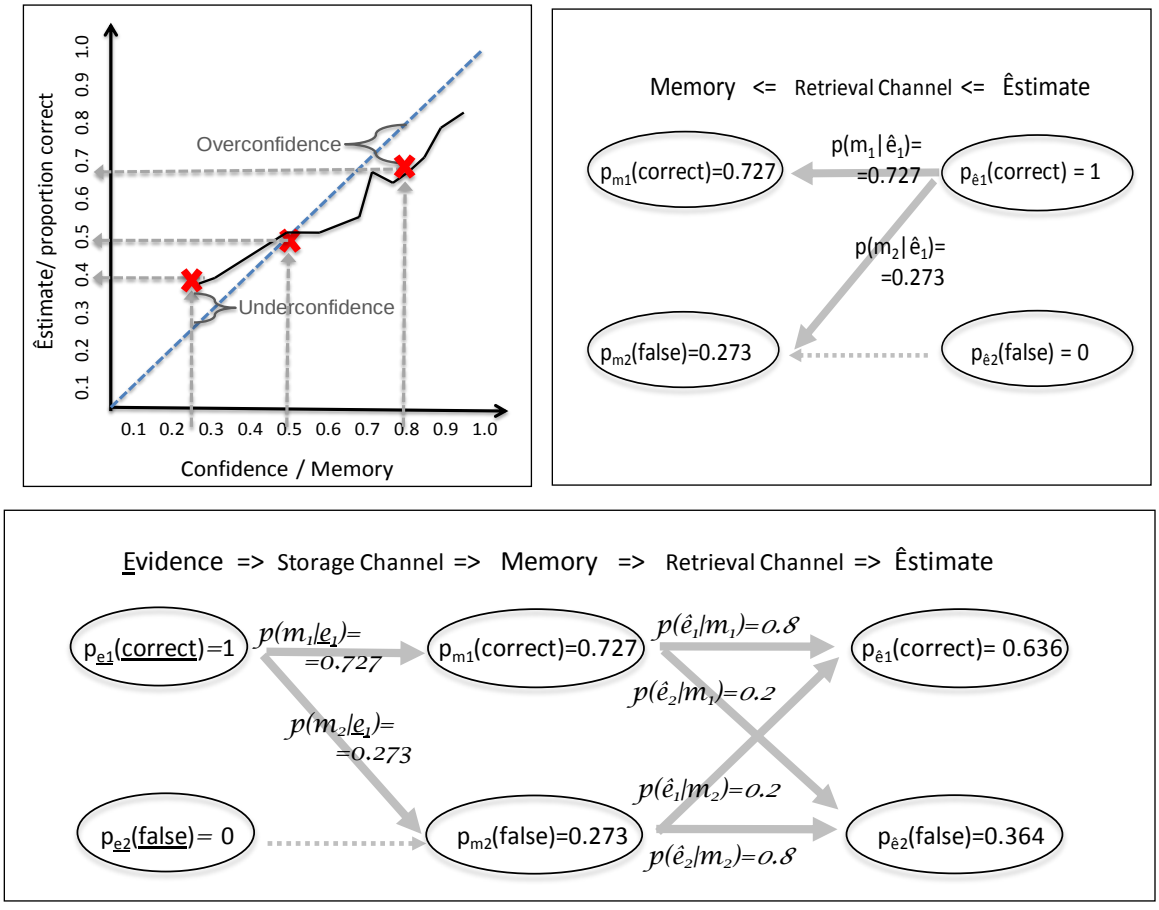
Noise ($\hat{E}|M$) between Memory and Êstimate: the Confidence Bias

The confidence bias originates in the internal uncertainty of the judge, rather than environmental uncertainty regarding the objective evidence (like conservatism or the exaggerated expectation bias). It refers to subjective uncertainty about the objective facts (see Wagenaar & Keren, 1985). More specifically, the confidence bias is the experimentally confirmed fact that we tend to be overconfident in our judgments when we are fairly certain about something, and underconfident when we have a high level of subjective uncertainty (for discussions see Keren, 1997; Liberman & Tversky, 1993; McClelland and Bolger). For example, in cases where people are fairly certain that their answer is correct (let's say 80% certainty), it usually turns out that fewer than 80% of all answers that they judge to be correct are indeed correct (perhaps only 70%). In contrast, in cases where judges are only 25% sure about the correctness of the answer, more than 25% (perhaps 40%) of those answers are actually correct. (See the traditional representation in Figure 8a, based on Adams & Adams, 1960.)

Fitting empirical findings to the noisy retrieval sub-channel.

The traditional set-up of confidence tasks involves asking general knowledge questions, such as, "Absinthe is (a) a liqueur or (b) a precious stone?"[12] The judges are asked to give two responses. First, they are asked to choose one of the two alternatives as their best guess at the correct response. Second, they are asked to indicate their confidence in their estimate. This confidence rating is usually solicited on a scale from 0%-100%. (Unfortunately, it is often also solicited on a rather distorting scale from 50%-100% for binary decision-making tasks, which leads to distorted results.[13]) Figure 8a shows the empirical findings from a classic study of the confidence bias (Adams & Adams, 1960). Here, people were asked how confident they were that the spelling of a word was correct, after reading and writing the potentially misspelled word. Perfect judgment would imply that our confidence judgments would line up on the diagonal 45° line with the proportion correct of our estimates (the "hit-rate").

Figure 8: Confidence bias, (a) traditional representation; (b) looking into memory; (c) memory channel representation of a general knowledge task.



Source: Author, based on (a) Adams and Adams (1960); (b, c) Lichtenstein and Fischhoff (1977).

How can we model the confidence bias with the tools of our noisy memory channel? I enter the memory channel from the side of our estimates, $P(\hat{E})$. The equivocation $P(M|\hat{e}_i)$ represents the level of confidence the judge has when "looking into his or her own memory." The judge asks, "given that my estimate is $p(\hat{e}_i)$, what is the probability that the original evidence is $p(\underline{e}_i)$? How confident am I about my judgment?" During a decision task our memory separates us from the original evidence (that is, $P(\underline{E})$ and $P(\hat{E})$ are conditionally independent given P(M), meaning that the memory channel is a Markov chain), thus, a judge has to base his or her confidence estimate on $P(M|\hat{e}_i)$: "given that my estimate is $p(\hat{e}_i)$, what do I find in my memory P(M) to support this claim?" In short, the equivocation $P(M|\hat{e}_i)$ represents the confidence that a judge has in his or her own judgment. When reporting confidence, the judge reports on the distribution of what is found in memory.

After collecting over 9000 responses to binary general knowledge questions, Lichtenstein and Fischhoff (1977) reported in their much-citied study that respondents had, on average, 72.7% confidence in their judgments. Figure 8b represents this finding in the form of the retrieval channel. As a judge, one first chooses what she or he believes to be the correct option. The participant then "looks into her/his own memory" for an estimate of the support her/his memory provides for the choice $p(M|\hat{e}_i)$. The average person found that the content of his or her memory supported this estimate to 72.7%, with 27.3% of memories supporting other estimates. Again, for our purposes, it does not matter whether these probabilities are the result of "episodic multi-trace memory" (see Baddeley et. al., 2009, Ch.5), a "semantic memory" (see Baddeley et.al., 2009, Ch.6), or a mix of both. We can imagine that 727 out of 1000 relevant episodic memory traces end up with this result, or we can assume that the average person learned or "believes in" 7 (or so) reasons in favor, and roughly three reasons that speak against this.

Having an idea about the content of the judge's memory has two implications. First, we can estimate the storage channel. Because we are concerned only with general knowledge questions, we can assume that there is no uncertainty involved in the original evidence $P(\underline{E})$. Absinthe is a liqueur, no doubt about it. Many questions, of course, do contain an innate uncertainty: questions about the future, for example (will it rain tomorrow?), or questions about uncertain aspects of the past (did the island of Atlantis really exist?). Furthermore, even with general knowledge questions without uncertainty, we do not know if our subject has been confronted with evidence suggesting that absinthe is a stone (if, for example, somebody lied to our judge or provided manipulated evidence). In this article, we do not consider the problematic of wrong input evidence, such as lies and deceptions, because false input does not make us irrational. We would still act completely rational, based on false premises. Therefore, we consider only general knowledge question that are not manipulated by misinformation and do not contain innate uncertainty. Given this assumption, we know that the difference between $P(\underline{E})$ and P(M) must be attributed exclusively to noise in the storage channel, as shown in the graph of Figure 8c.

In addition to P(E) and P(M), we also have empirical evidence for the distribution of our estimate, P(Ê). Lichtenstein and Fischhoff (1977) found that only 63.6% of the answers by subjects in their study were correct. Comparing 72.7% with 63.6% gives us the typical result of overconfidence. Finally, with P(M) and P(Ê) we can approximate the noise in the retrieval channel, P(Ê|M). Assuming Property B (*binary symmetric*) for binary decision-making exercises, we suppose that a judge is equally likely to confuse memories about absinthe being a liqueur with absinthe being a stone. The weights of 0.2 crossover probability and 0.8 identity transitions happen to fit this empirical finding, following the total probability theorem with a binary symmetric channel: $0.636 = [0.727*p(\hat{e}_1|m_1)] + [0.273*p(\hat{e}_1|m_2)]$, with $p(\hat{e}_1|m_2) = [1 - p(\hat{e}_1|m_1)]$ (because binary symmetric), and solving for $p(\hat{e}_1|m_1)=0.8$. The resulting picture of the entire memory channel is given in Figure 8c.

Formalizing the confidence bias.

Similar to our formal definitions of conservatism, I can also provide a precise mathematical definition of the confidence bias. On a traditional x/y-axis plane I can depict the level of confidence on the x-axis and the proportion correct (hit-rate) on the y-axis (see Figure 8a). Empirical evidence typically shows that the slope of the regression line based on the x-axis is between 0 and 1. As I have shown, the level of confidence can be understood as the distribution of evidence in the judge's memory, P(M). This gives us the following formal definition of conservatism:

$0 \leq \text{slope}_{P(M)}$ of regression between P(M) and P(Ê), based on $P(M) \leq 1$.

As before, this inequality can be reformulated. The $\text{slope}_{P(M)}$ depends on the regression coefficient r, and the standard deviations of P(M) ($\sigma_{P(M)}$) and P(Ê) ($\sigma_{P(\hat{E})}$):

$$0 \leq \text{slope}_{P(M)} \leq 1 \qquad \text{(IV)}$$

$$0 \leq [r \times \sigma_{P(\hat{E})}] / \sigma_{P(M)} \leq 1$$

$$0 \leq [r \times \sigma_{P(\hat{E})}] \leq \sigma_{P(M)} \qquad \text{(V) (compare with Table 1)}$$

Inequality (IV) illustrates how estimates turn out to be conservative in comparison to confidence levels. It is surely satisfied when the correlation between the estimate P(Ê) and the evidence P(M) is positive ($r > 0$), and the standard deviation of the subjective estimates P(Ê) is smaller (on average closer to its mean) than the standard deviation of the evidence in memory P(M), which is more "spread out" toward the extremes (further from its mean). In other words: estimates are "conservative" with respect to confidence levels derived from memory because such estimates' extremes are less accentuated (closer to their means).

The presented characteristic of the retrieval channel follows the same logic observed for conservatism in the overall memory channel from P(E) to P(Ê), discussed previously. When we detect high probabilities in our memories, it seems to be the case that noise mixes memories of lower likelihoods into our estimates. Likewise, when faced with a

question about which we have little evidence in memory, memories of higher likelihoods seem to sneak in. We apply the same Properties N_i, B, S and N to the retrieval channel, P(M) to P(Ê). Following the same logic as before, but replacing Ê|$\underline{E}$ with Ê|M, we can show that these properties lead to the confidence bias. Consider that in the typical case of the confidence bias, the standard exercise focuses on a very simple binary likelihood decision-making task: how confident are you that the estimate is correct? Traditionally, empirical confidence bias studies therefore merely test for binary, not multiary decision-making tasks, and given the nature of the question (how confident are you?), they focus on assessing likelihoods (not absolute values).

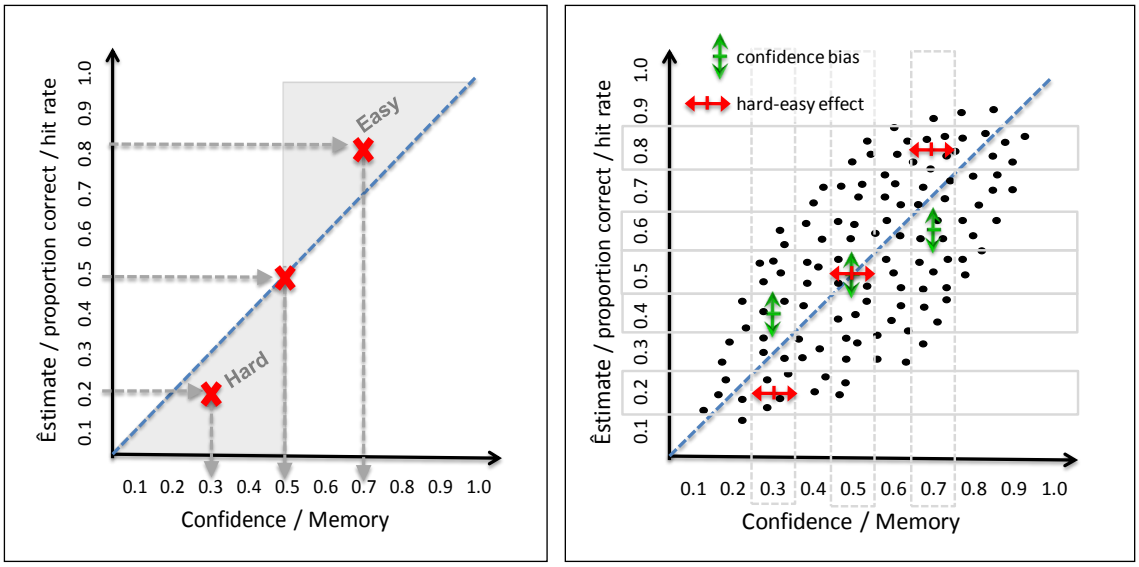
Noise (M|Ê) between Êstimate and Memory: the Hard-Easy Bias

There is another bias that can be explained by modeling noise in the retrieval sub-channel. Known as the hard-easy effect (e.g. Lichtenstein and Fischhoff, 1977; Lichtenstein, Fischhoff, and Phillips, 1982; Keren, 1988; Suantak, Bolger and Ferrell, 1996), it looks at the retrieval sub-channel the other way around, from estimate Ê to memory M. It states that based on a specific level of task difficulty, our confidence in judgments is too conservative. It is therefore the conditional inverse of the confidence bias. The hard easy bias has also been simulated by computer programs (Juslin, et.al., 2000; Merkle, 2009), and often leads to some confusion. Unlike the confidence bias, it is based not on the expected hit-rate values, conditioned on a specific confidence level, Expected value (EV) of [P(Ê)|p(m_i)], but transformed by Bayes' theorem, on the expected value of the confidence, conditioned on a specific hit-rate, EV[P(M)|p($\hat{e}_i$)]. By the definition of what is hard and what it easy, hard tasks are defined as having low average hit-rates, while easy tasks have high ones. In general knowledge tasks without innate uncertainty (with p($\underline{e}_{correct}$)=1), the distribution of the estimate, p(Ê), gives us the overall difficulty of the task.

Traditionally, different exercises are executed for tasks with different levels of difficulty (hard and easy). Figure 9a schematizes the typical finding of the hard-easy bias. In contrast to the usual convention, Figure 9a leaves the x- and y-axis at the same order as we had in Figure 8a. Since the hard-easy bias looks at the respective average level of confidence for each task, given the task difficulty, this order requires us to read the graph from the left y-axis to the bottom x-axis, in contrast to the traditional reading of the bottom x-axis to the left y-axis (see arrows in Figure 9a). It turns out that for tasks with a low average hit-rate (hard: 0.2 hit rate in Figure 9a) the average level of confidence is higher than it should be (0.3 confidence in Figure 9a), and that for tasks with a high average hit-rate (easy: 0.8 hit rate), the average confidence is lower than it should be (0.7 confidence), or, more formally: EV[P(M)|p($\hat{e}_{0.2}$)] = 0.3; EV[P(M)|p($\hat{e}_{0.5}$)] = 0.5; EV[P(M)|p($\hat{e}_{0.8}$0] = 0.7. In contrast to the previous exercise, where we conditioned Figure 8a on the level of confidence ("x-axis"), we now condition on the estimate "y-axis" and ask: given a certain level of hit-rate, what is the expected value of confidence in memory, EV[P(M)|p($\hat{e}_i$)]?

Students of this phenomena often determine the relation between the average hit-rate and the average confidence by subtracting the expected value of the hit-rate from the expected value of confidence, EV[P(M)]–EV[P(Ê)]. The result is then ranked according to the level of difficulty. This is just a more compressed way of achieving the same result.

Figure 9: Hard-easy effect; (a) conditioning on proportions correct (y-axis); (b): confidence bias (conditioned on x-axis) and hard-easy effect (conditioned on y-axis)



Source: Author.

ollecting these averages for many diverse tasks (hard and easy), we can run a regression line through these averages**Error! Bookmark not defined.**, getting a regression with slope$_{P(\hat{E})}$. The hard-easy effect claims that this regression line will typically be between 0 and 1 (see Juslin, Winman and Olsson, 2000; Merkle, 2009)[14]. We therefore define the hard-easy effect with:

$$\text{slope}_{P(\hat{E})} \geq 1 \qquad \text{(VI)}$$

$$\sigma_{P(\hat{E})} / [r \times \sigma_{P(M)}] \geq 1$$

$$\sigma_{P(\hat{E})} \geq [r \times \sigma_{P(M)}] \qquad \text{(VII) (compare with Table 1)}$$

This definition suggests that the equivocation side of the retrieval channel, $p(M|\hat{e}_i)$, seems to be governed by the same properties as the noise side of the channel, $p(\hat{E}|m_i)$. Following the reversing logic suggested by Erev, et al. (1994), we simply have to exchange the conditioned and the conditioning variable. Applied to this end, Property S (*identity-symmetric noise*) states that we are equally likely to confuse equally different memories. Property N (*single-peaked unimodal noise*) states that we can on average trust our judgment based on our memories: the largest part of our estimate comes from the correct memory. It also states that equivocation in our retrieval process becomes smaller the more dissimilar

the equivocated memory is from the correct memory. From a psychological perspective, these properties are intuitively pleasing.

Discussion of Several Biases within one Theoretical Framework

One of the main benefits of defining several biases within one common conceptual framework is the ability to show how they are related. For example, we are now in a position to explain both the confidence bias and the hard easy effect with one single theoretical framework. The key to understanding the difference between them lies in the conditioning variable (Erev, et.al., 1994). The micro-data that can give rise to both effects are schematized in Figure 9b, which presents what introductory statistics textbooks call a "football-shaped-cloud" (see Freedman, Pisani and Purves, 2007). Combining our definition of the confidence bias (equation V) and our definition of the hard-easy effect (equation VII) gives us the following results:

$$0 \le [r^{x}\, \sigma_{P(\hat{E})}] \le \sigma_{P(M)} \qquad \text{(V)}$$

$$\Rightarrow 1/r \ge \sigma_{P(\hat{E})}/\sigma_{P(M)} \ge 0 \qquad \text{(Va)}$$

$$\sigma_{P(\hat{E})} \ge [r^{x}\, \sigma_{P(M)}] \qquad \text{(VII)}$$

$$\Rightarrow \quad \sigma_{P(\hat{E})}/\sigma_{P(M)} \ge r \qquad \text{(VIIa)}$$

Combining (Va) and (VIIa):

$$1/r \ge \sigma_{P(\hat{E})}/\sigma_{P(M)} \ge r \ge 0 \qquad \text{(VIIIa)}$$

or: $$[Var(P(M)) / cov(P(M),P(\hat{E}))] \ge 1 \ge [cov(P(M),P(\hat{E})) / Var(P(\hat{E}))/] \ge 0 \qquad \text{(VIIIb)}$$

Equations (VIIIa) and (VIIb) describe a retrieval channel that leads to both the confidence bias and the hard-easy effect. The equations clearly show how the effects are related and define the co-dependence and trade-off that exist between the biases. The limitations between M and $\hat{E}$ arise as a result of applying Properties B and N for binary cases, or S and N for exercises from an equidistant interval scale. We see that the confidence bias and the hard easy effect are two sides of the same coin: the first one can be modeled with the noise of the retrieval channel ($\hat{E}|M$), and the second one can be understood as a consequence of the properties of the equivocation of the retrieval channel ($M|\hat{E}$). Both conditioned variables are related by Bayes' theorem: $P(\hat{E}|M) = {}^{[P(M|\hat{E})^{*}\, P(\hat{E})]}/_{P(M)}$.

Following the logic outlined in equations (V) to (VIII), but replacing M with $\underline{E}$ and applying Properties B and N, or S and N, to the overall channel between objective evidence $\underline{E}$ and estimate $\hat{E}$, it is straightforward to show that the overall channel is governed by the same limitations and trade-offs.

Possible Psychological Generative Mechanisms for Noise in Binary and Equidistant Decision-Making Tasks

From a psychological perspective, Properties B, N, and S do not seem too farfetched. I would argue that they seem reasonable and their psychological interpretations, briefly sketched above, are intuitive. But what are the psychological mechanisms that could possibly generate noise of such nature? In this section I discuss some possible candidates.

The Gaussian Channel

Normal noise (following the "bell curve") is a popular choice in many of the above-mentioned "random error models" (e.g. Erev, et.al., 1994; Wallsten and González-Vallejo; 1994). But why should the kind of noise that interferes with human decision-making be normally distributed? Why not binomial noise (such as suggested by Budescu, et.al., 1997b)? Or maybe there is another form of interference into the process of human judgment (exponential, beta, gamma, Cauchy, Poisson, etc.)? Where, if the kind of noise that interferes with human decision-making is normally distributed, does the normally distributed noise come from?

There are many possible sources of interference in the process from objective evidence to memory, and from memory to estimates. Though it may appear as a weakness of the current argument, this multiplicity of different effects that may lead to the cumulative effect of "mixing things" (noise) in fact supports the proposal of normally distributed noise, since the central limit theorem states that the cumulative effect of a large number of independent random effects will be approximately normally distributed. Thus, the fact that there are so many potential sources for noise, makes it in fact quite reasonable to assume that noise is normally distributed. The resulting Gaussian channel already serves as a successful model for some of the most common communication technology channels, such as wired and wireless telephone channels and satellite links (which also suffer from interference from a large number of independent causes). It is one of the most common channels studied in information theory and its properties are well understood (see Cover and Thomas, 2006, Ch.9).

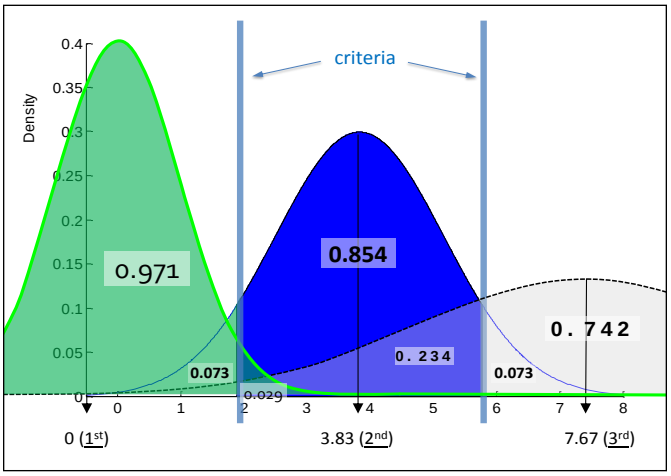
If normal noise of the same variance is applied, Properties B and N_i are satisfied. The Gaussian channel also satisfies Properties S and N, since the normal distribution is symmetric and single-peaked unimodal. It is worth noting that Properties S and N do not require each piece of evidence to be affected by normal noise with the same variance. Noise can affect different estimates differently, and the study of these differences can give us much insight into the inner workings of our irrationality.

Figure 10a models Hockely's (1984) ternary experiment with normal noise: [signal + normally distributed noise around the identity transition] (compare with Figure 6). The

reader might see a similarity with the information theoretic logic applied to psychological Signal Detection Theory (Goldstein, 2002, Appendix A; Heeger, 1997; Swets, 1964; see also Wallsten & González-Vallejo, 1994). The vertical lines represent the criteria by which a judge classifies input evidence wrongly. Figure 6a represents the fact that 3rd word repetitions are most affected by noise: the normal curve around $\underline{e}_3$ has the largest variance.

There are, however, limitations when fitting normal curves to empirical findings concerning decision-making tasks of more than two binary choices. With multiary choices it is not possible to fit all the curves perfectly to the empirical findings because each normal curve only has two degrees of freedom by which to adjust it to the data; its mean and variance. As the number of variables grows from ternary to quaternary exercises and beyond, this limit on the degrees of freedom makes it increasingly difficult to fit normal curves to all aspects of the empirical findings. The result is thus only a rough approximation, but it often works quite well for the replication of communication systems in electrical engineering. I discuss the fitting process further in Appendix D.

Figure 10: (a) replication of Hockley's (1984) ternary experiment with normal noise; (b) transition matrix P(Ê|E) for empirical finding of Hockley (1984) with indication of model deviations.



Evidence E	Estimate Ê			
	1	2	3	Expct. [Ê|e_i]
1	0.971	0.025 *+0.004*	0.004 *-0.004*	*≈1.029*
2	0.073	0.854	0.073	*2.00*
3	0.024	0.234	0.742	*2.72*

Source: Author, in reference to empirical data from Hockley (1984).

Figure 10b shows the fitting limitations in a transition matrix. In this solution I decided to limit this misfit to evidence 1 (1st repetition), but this is not binding. As the name indicates, such matrices focus on the transition probabilities P(Ê|E) and are often used to work with channels. These matrices represent the visual logic of channel graphs (like Figures 1, 2, 4, and 6b), but with the analytical benefit that the well-known tools of matrix algebra can be applied to analyze the given properties. Figure 10b corresponds to the traditional x/y-representation of Hockley's (1984) ternary experiment in Figure 6a, the channel representation in Figure 6b, and the normal noise approximation in figure 10a. All four kinds of representation show the same finding in different ways.

Other Candidate Mechanisms

Modeling noise with the normal distribution is quite popular and justifiable, not least because there is a justifiable generative mechanism: the central limit theorem. There might, however, be other mechanisms that generate noise satisfying Properties B, S, and N. It is beyond the scope of this article to dig much deeper into the search for further candidate mechanisms, but I will present one additional possibility.

This possibility is to interpret the noisy memory channel as an evolving network. The analysis of complex networks has become a rich and quite sophisticated area of research over recent years (see, for example, Albert & Barbasi, 2002; Neuman, 2003; 2010). For purposes of illustration, I use a very simple case. Suppose that the first presented evidence results in a correct identity transition and therefore creates a valid connection in the noisy memory channel. Suppose, too, that this connection is not stable and that over time the (initially correct) link starts to "wander" up and down the estimation scale Ê at random. A simple 50%-50% random walk up and down on a one-dimensional scale results in a distribution based on a binomial coefficient: P(k)= P(k)= $2^{-n}\binom{n}{(n+k)/2}$. Here, n is the number of steps taken, k the final position on the scale, and p(k) the probability of ending up at position k. Modeling noise according to this "random-walk" logic leads to an open-ended scale that satisfies Properties N and S. Thus, there are other potential generative mechanisms for the noise properties that we identified. Following the logic of complex networks, it is possible to come up with much more sophisticated models that could provide reasonable explanations why the noisy memory channel produces the selected biases.

Besides the logic of Gaussian noise and random walk, there might be many other possible generative mechanisms that can create the properties that generate the kind of noise we are looking for.

Additional Channel Properties

For reasons of completeness, in this section I now discuss additional channel properties that also produce the identified biases. The first one is an alternative to Property N; the second one loosens the quite strict requirement of symmetric noise (Property S) and therefore applies to a much broader category of decision-making exercises involving multiary choices whose values are not equidistant (for example, instead of choosing from values 1, 2, 3, which are all 1Δ apart; choose from values 1, 7, 99, with 1 and 7 being 6Δ apart, and 7 and 99 being 92Δ apart). The third section discusses approximate channel properties, which do guarantee to produce the biases, but provide more flexibility for psychological interpretations.

A Property for Unbounded Noise Distributions

Normal noise and open-ended random-walks are unbounded (from -∞ to +∞), which leads to the question of what to do with the "overshoot". The "overshoot" refers to estimates outside of the defined scope of the decision-making exercise (such as illusions)[4]. In line with many empirical studies, Hockley (1984), for example, simply "truncates" or "cuts off" the overshooting estimates (p.230). He deletes 354 answers that referred to a (non-existent) 4th repetition in his ternary word repetition exercise. The same is done in the random-error model of Juslin et.al. (1997). This, however, seems to amount to simply denying that the mind creates such kind of illusions.

Another alternative is to add the overshoot to the extremes, which seems psychologically reasonable: if it appears to a judge that some event occurred more times than allowed by the nature of the decision-making exercise, it is reasonable to expect that the judge would simply add this "overshoot" to the highest possible option. This leads to an additional property:

Property U: (*Adding overshoot of unbounded noise to extremes*). In cases where noise is modeled through an unbounded distribution, add the transition probabilities of the overshoots to the extremes, $\hat{e}_0$ and $\hat{e}_n$:

$p(\hat{e}_{low\text{-}extreme}|\underline{e}_i) = p(\hat{e}_0|\underline{e}_i) + p(\hat{e}_{values<0}|\underline{e}_i)$; and

$p(\hat{e}_{high\text{-}extreme}|\underline{e}_i) = p(\hat{e}_n|\underline{e}_i) + p(\hat{e}_{values>n}|\underline{e}_i)$.

Note that the definition of Property S (*identity-symmetric noise*) assumes an unbounded scale from -∞ to +∞, depending on the nature of j. Therefore, Properties S and U can be applied to the noisy memory channel simultaneously. As I show in Appendix E, it turns out that Property U renders Property N (*single-peaked unimodal noise*) redundant. Properties S and U will also always lead to conservative noise. If we follow the psychologically justifiable procedure of adding the overshoot of unbounded noise to the respective extremes, it turns out that the noise distribution does not need to be unimodal single-peaked. The noise distribution might as well be W or U-shaped, as long as it is symmetric and we suppose that the unbounded overshoot is added to the extremes. There is no harm if it is unimodal single-peaked, but Appendix E shows that this is not a necessary condition.

A Property for non-Equidistant Multiary Decision-Making Tasks

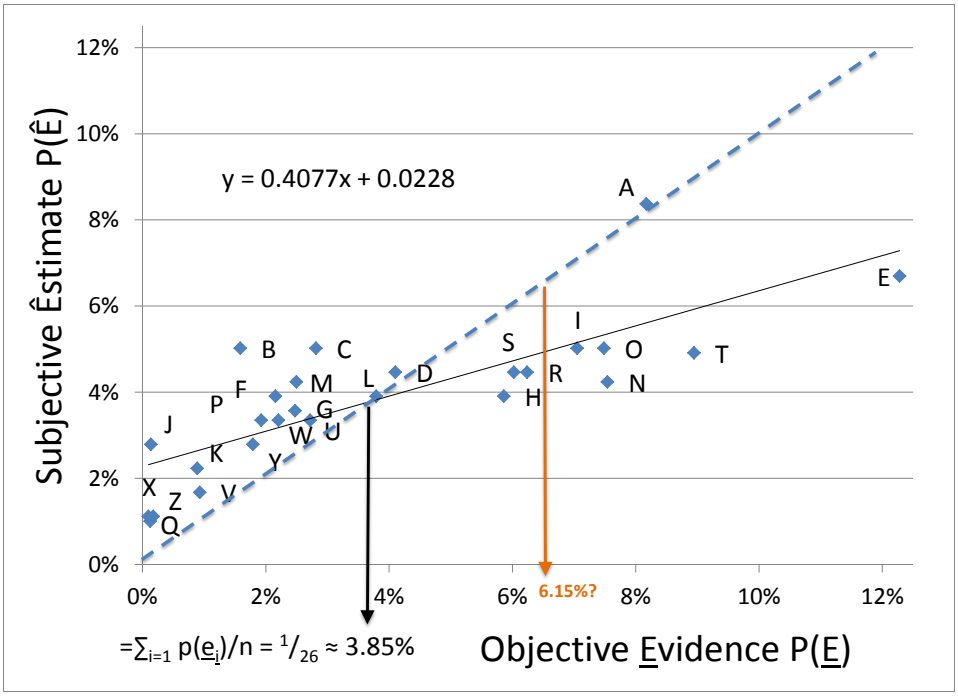
So far, I have exclusively discussed binary decision-making tasks and judgments that rely on an equidistant interval scale, which are the most common decision-making tasks. The overwhelming majority of existent empirical evidence refers to them, as do existing computer simulated "random-error models", which typically model choices from a binary decision space (e.g. Budescu, et.al., 1997c; Dougherty, et.al., 1999; Erev et.al., 1994; Juslin, et.al, 1997; Merkle, 2009; Wallsten and González-Vallejo, 1994). These kind of decision-

making exercises are also the basis for influential principles that combine different biases, such as prospect theory (Kahneman & Tversky, 1979, 1992). Presenting a synthetic framework that provides a possible explanation for these two kinds of decision-making exercises, as I have done, is a valuable finding by itself. It shows what these exercises have in common, within what parameter settings the models work, and the kinds of simple generative mechanisms that can account for them. Yet despite this success, it is important to point out that the identified properties fail when we try to use them to explain the same biases for multiary exercises on a non-equidistant skewed scale. This is also the reason why I worded the previous sections carefully when discussing the kind of exercises for which the presented explanations work flawlessly.

For example, in a pioneering and now classic study on "psychological probability as a function of experienced frequency", Attneave (1953) showed that we are conservative in our estimates of letter frequencies. There are 26 letters in the English alphabet (26-ary decision-making exercise) and their probability is not equidistant: the letter E appears 12.3 % of the time, the letter T 8.9 %, the letter A 8.2 %, whereas more than half of the letters appear less than 3 % of the time. There is no fixed value Δ that separates the likelihood of each letter from adjacent letters. Figure 11 depicts Attneave's findings in the traditional presentation style of conservatism. The x-axis represents objective evidence, $P(\underline{E})$, and the y-axis empirically detected subjective estimates, $P(\hat{E})$. The estimates are conservative: the slope of the regression line, based on $P(\underline{E})$, is between 0 and 1 (0.4077 in this case) and satisfies the definition of conservatism from equation (I). It shows that the "turning point" between over- and underestimation can be found around the mean value of the input, which is at $\sum_{i=1} p(\underline{e}_i)/n = {}^1/_n$. In this 26-ary alphabet it is at ${}^1/_{26} = 0.03846$. This fits the empirically determined turning point remarkably well and agrees with it until the fifth digit: $y = 0.4077x + 0.0228$, with $y = x = 0.03849$.

According to Property S (*identity-symmetric noise*), however, the turning point should not be at the mean, but at the mid-range point of $m = [\underline{e}_0+\underline{e}_n]/2 = [0+0.123]/2 = 0.0615$.[11] In other words, if one supposes a symmetric noise distribution (such as used by most computer simulations), there would be conservatism around the mid-range point of 6.15 % (i.e. the slope of the regression line would cross the diagonal at 6.15 %, see Figure 11), not around the empirically detected turning-point around 3.85 %. It turns out that in binary and multiary equidistant exercises this issue does not arisebecause the mid-range point and mean are equivalent.[11] In other words, the previously mentioned "random error models" with symmetric noise work well for binary or multiary equidistant decision-making exercises (e.g. Budescu, et.al., 1997; Erev, et.al., 1994; Juslin, et.al, 1997; Merkle, 2009), but if used to replicate the findings of Attneave's (1953) multiary non-equidistant exercise, Properties S and N (and even Properties S and U) would fail to replicate this empirical finding (Figure 11). Although the existing computer simulations unfortunately have not been tested for the replication of multiary non-equidistant exercises, our theoretical analysis shows this limitation very clearly. There must be other properties for our channel in order to provide a working model for these additional cases.

Figure 11: Conservative estimate of the likelihood of letters in a newspaper.



Source: based on Attneave, 1953.

Inspired by a well-known finding from information theory (Cover & Thomas, 2006, Ch.4, p. 88, Exercise 4.1)[15], I introduce an additional property:

Property D: (*Doubly stochastic mixing*). The transition matrix is doubly stochastic: $\Sigma_j p(\hat{e}_j|e_k) = 1$; and $\Sigma_k p(\hat{e}_j|e_k) = 1$.

A transition matrix is said to be *doubly stochastic* if all marginal rows and all columns of the conditional transition probabilities sum up to 1. For example, the binary symmetric channel (Property B) is doubly stochastic, because $p(\hat{e}_1|e_1)+p(\hat{e}_2|e_1) = 1$, and $p(\hat{e}_1|e_1)+p(\hat{e}_1|e_2) = 1$, whereas the our simulated Gaussian channel from Figure 10b is not doubly stochastic: $p(\hat{e}_1|e_1)+p(\hat{e}_2|e_1)+p(\hat{e}_2|e_1) = 0.971+0.025+0.004 = 1$, but the marginal probability of $p(\hat{e}_1|e_1)+p(\hat{e}_1|e_2)+p(\hat{e}_1|e_3) = 0.971+0.073+0.024 = 1.068 \neq 1$.

As shown in Appendix F, the combination of Properties D (*doubly stochastic mixing*) and N (*single-peaked unimodal noise*) also produces the conservatism bias for likelihood estimates on all kinds of random variables and for uniformly distributed input variable of absolute values, regardless of distance between the absolute values. Property S is not required in this case. Noise does not have to be symmetric around the identity transition. As a result, the turning-point is not fixed at the mid-range point.

How well does Property D fit empirical findings? Table 3 shows the before-mentioned empirical finding of Hockley's (1984) 5-ary repetition experiment (with up to five word repetitions, p.237) (sample size: 45,760). The result is conservative: 1st and 2nd

word repetitions are overestimated (1.064 and 2.222), while 4th and 5th repetitions are underestimated (3.883 and 4.410). The empirically detected transition matrix, however, is not perfectly doubly stochastic, because the columns do not sum up to 1. The 4th and 5th estimates, in particular, deviate from Property D (the 4th column sums up to 1.281, and the 5th to 0.695). We could improve this fit between Property D and Hockley's empirical findings by setting up a program (for example in MATLAB) which minimizes the distance between the empirical transition matrix and the modeled transition matrix, while being restricted by double stochasticity. Least squares can be used. But this would only improve the fit, not the fact that it does not it must not fit exactly. Furthermore, unlike Property S, I am not aware of any plausible psychological reason that might justify why information processing should be doubly-stochastic.

Table 3: Transition matrices of P(Ê|E): empirical finding of Hockley (1984) for 5-ary decision-making exercise

		Estimate Ê						
		1	2	3	4	5	∑ of rows	*Resulting mean estimate*
Evidence E	1	0.959	0.028	0.008	0.005	0.001	**1**	1.064
	2	0.028	0.774	0.149	0.041	0.007	**1**	2.222
	3	0.006	0.092	0.659	0.206	0.037	**1**	3.176
	4	0.003	0.02	0.214	0.617	0.146	**1**	3.883
	5	0.001	0.008	0.075	0.412	0.504	**1**	4.410
	∑ of columns	**0.997**	**0.922**	**1.105**	**1.281**	**0.695**	**5**	

Source: Hockley (1984), p. 237.

Still, on mathematical grounds doubly stochastic transition matrices (in combination with Property N) must produce conservative noise for all kinds of decision-making exercises. This is a very solid and comforting finding, because we now know that there is at least one well-defined constellation of properties in our framework that can explain the eight biases even for multiary exercises on a skewed scale. Currently, however, there is a lack of empirical and psychological backing for this alternative channel. Therefore, Property D should be viewed as a rough approximation and must be used with these caveats until there is a more thorough understanding of its implications. It might well turn out that further research shows that another noise distribution fits empirical findings much better than Property D.

Approximate Channel Properties

So far, I have been very rigorous with the exploration of channel properties. Much in the tradition of information theory, I have provided unambiguous mathematical proofs that the proposed channel properties must lead to the empirically detected results (see Appendices C, D, E, and F). However, psychology is not always as exact. Cognitive biases are stylized facts, not mathematical laws. We find them on average and as general tendencies, not strictly and absolutely always. In some studies, a particular bias might not even be detected or only show up partially or approximately. Therefore, it seems reasonable to also explore some approximate channel properties, which hold most of the time, but not always. Applying Property N (*single-peaked unimodal noise*) by itself to the overall memory channel and/or the retrieval sub-channel, the respective biases would result for many kinds of tasks and for many kinds of remaining channel constellations, but not for all. The lost precision of the prediction is recompensed by the elimination of the constraints of Properties S (*identity-symmetric noise*) or D (*doubly stochastic mixing*), which provides more flexibility for the psychological interpretation of the involved cognitive process.

Property N can be fortified with another property, which expands the psychologically pleasing Property N_i ("*more right than wrong*") to all kinds of binary- and multiary, equidistant- and non-equidistant tasks:

Property N_d: (*Identity transition Dominance*). The magnitude of each identity transition exceeds the sum of the magnitudes of all noise transitions: $p(\hat{e}_i|\underline{e}_i) \geq \Sigma\, p(\hat{e}_x|\underline{e}_i)$, for all $x \neq i$.

Property N_d describes so-called diagonally dominant transition matrices. These are nonsingular matrices that occur naturally in a wide variety of practical applications (Meyer, 2001). Combining Property N with Property N_d, the respective biases will result most of the time, but still not always. Nevertheless, the latter would require that the noise transitions had a very strong and peculiar tendency toward one side, or that the decision-making task was set up in a very unbalanced (non-equidistant) manner. By and large, however, the channel will produce the detected biases. The psychological interpretation of Properties N and N_d is straightforward: on average, our estimates are more right than (the combined) wrong (N_d), and we are more likely to confuse "similar things" than "less similar ones" (N). Both Hockley's ternary and quinary exercises (1984) fulfill Properties N and N_d perfectly (see Figures 6b and 10b, and Table 3).

Conclusions and Limitations

In this article I have shown that some intuitively pleasing properties of one single theoretical framework—following one single underlying logic (mixing noise)—are sufficient to provide a possible explanation of eight seemingly unrelated decision-making biases. I identified several channel properties and shown via mathematical proofs (in the Appendices) that those must inevitably give rise to the regularities detected in empirical studies on these biases. By doing so, I have also shown the limitations within which specific properties work and when they do not.

We are now in a position to return to and fully read Table 1, which lists the eight empirically detected biases analyzed in this article. The table also summarizes the mathematical formalizations that define the biases and the channel properties that can produce them. The benefit of using one single theoretical framework is that it enables us to formulate a set of unambiguous and formal definitions of those (often slippery) concepts that explain our irrationality (biases). This clarity makes them also more susceptible to rebuttal from future empirical experiments, which can aim at fine-tuning the our understanding of human judgment by testing clearly defined mathematical hypothesis. For example, defining the hard-easy bias as $\sigma_{P(\hat{E})} \geq [r \times \sigma_{P(M)}]$, one is now in a position to quantify the question of how strongly this condition is satisfied when testing judges with different levels of relevant expertise.

Of course, eight is only a small percentage of the vast and ever-growing list of cognitive biases. Whereas Baron (2008; Table 2.1) lists over 50 biases, the collective and open-access online encyclopedia Wikipedia enlists 35 biases for probability and belief, and 44 behavioral biases ("List of cognitive biases", 2010). Synthesizing eight of them into one framework and one single basic generative mechanism is a humble but promising start.

Resulting Research Questions

In the future, it will be interesting to explore whether there is a margin of flexibility in the application of Properties B, N, S, U, and D, or whether there are other properties that both fit the empirical findings and have satisfying psychological explanations. This search should include the exploration of approximate properties, properties that work most of the time, but not always (such as Properties N and N_d, by themselves or in combination). Carefully designed empirical tests can help to differentiate between competing properties.

It will also be worthwhile to explore in which exceptional cases the identified properties are not satisfied. For example, Properties N and N_d have to be adjusted to rates of (selective) forgetting or inaccessibility.[4] It might be that a certain percentage of each choice is forgotten (uniform forgetting rate) or that some selected choices are more or less affected by forgetting or inaccessibility. Illusion and forgetting might require adjustments to the presented model.

A further interesting question is whether it matters how the channel is used and how the channel changes while using it. For example, we know that there are differences between learned and experienced frequencies and that frequent retrieval of memories has a similar rehearsal effect to the repeated consumption of a fact (Baddeley et.al., 2009, Ch.9). Also, sample size matters when storing to and retrieving from memory, which can account for additional biases (see Fiedler, 1996).

Another line of research opens up when considering that it is not required that both sides of the channel consist of the same number of different values. In reality it is likely that the input evidence (what we receive) is much larger than the output estimates that we usually rely on. In other words, our mind reduces the complexity of a myriad of inputs by categorizing them in a reduced set of classified memory traces. We create prototypes and typical groups (e.g., Goldstein, 2005: p.276), and we codify and re-codify them into cognitive chunks with varying informational content (Miller, 1956). In information theory, this logic is well studied and known as coding theory. Block-codes are often used to group items highly dependent on each other. This leads to important efficiency gains and increases the reliability of the processed information. Dynamical systems and chaos theory refer to this as the level of "coarse-graining" with which the environment is perceived, and it plays a crucial role in the interplay between description and prediction of a dynamical system (Gell-Mann, 1995b; Strogatz, 2001). It should come as no surprise that evolution has figured out a similar way to make use of information compression or coarse-graining through sophisticated coding in order to improve the effectiveness of our decisions (i.e., how much level of detail do we need to get by?). The before-mentioned "representative heuristic" (Tversky &Kahneman, 1974) and framing (Tversky &Kahneman, 1981) aim at the same question.

Much in line with coding theory, a major line of research opens up when considering that the analysis in this article focuses on non-normalized numerical variables, or on their likelihood. This implies that we can meaningfully assign one-dimensional numerical values to the observed evidence or its likelihood. I used variance in our analysis, which as a measure of difference works fine for decision making tasks involving absolute numbers or discrete probabilities but does not work for categorical or nominal variables. There are, however, other measures of variation, such as distance and attribute measures. These ask about the cognitive similarity between concepts like "tables, chairs, and elephants", for which variance is meaningless as a measure of difference[16]. A valid theory of human decision-making must also work for decision-making exercises that process non-numerical cognitive chunks, and variance might not be the right measure of distinctiveness.

Outlook

Perhaps the largest line of future research involves justifying the psychological generative mechanisms for judgment and decision-making biases. Traditional computer

simulation models simply plug in and utilize several possible noise distributions and see what works for which bias. This method of trial-and-error has led researchers to suggest a wide range of input and noise distributions to simulate biases, including normally distributed errors, log-odds plus normal and binomial distributed errors, uniform distributions, U-shaped, W-shaped, and beta-distributions (i.e. Erev., et al., 1994; Juslin, et al., 1997; Budescu, et al. 1997b; Merkle, 2009). All of them fall under the larger umbrella of the properties that I elaborated in this article. In retrospect, is not clear why these distributions and not others were chosen for testing in computer simulations. There are many other distributions that satisfy Properties S (*identity-symmetric noise*) and N (*single-peaked unimodal noise*), e.g. gamma-, poisson-, laplace-, gauchy-, skellam-, erlang-distributions, etc. All of them could be used to replicate cognitive biases for binary and equidistant decision-making tasks, yet each of them is produced by a different specific generative mechanism. Can we find a cognitive justification in favor of one or the other generative mechanism for noise in the human mind? Is it possible to device ingenuous empirical tests to find evidence in favor of one or the other generative mechanism?

Thus the future challenge shifts from statistically describing and blindly replicating judgment and decision-making through trial-and-error methods, to deepening our understanding of the informational processes that create these biases. I have proposed two possible candidates: the central limit theorem and a random-walk. Both are justifiable and work for binary and equidistant tasks, but not for multiary exercises on a skewed scale. This is a major current limitation.

The analytical tools derived from information theory (going back to Shannon, 1948) have been useful in identifying and understanding the related noise processes. Thus, there is reason to believe that the painstaking theoretical groundwork elaborated by communication engineers and computer scientists over the last 60 years will provide the right language to better understand the underlying informational processes. Still, despite the importance of the new analytical tools that have been provided by communication and computer engineers (such as the systematic analysis of noisy channels used here), the psychological challenge is different from the engineering challenge. Engineers aim at perfecting information processes by embedding them into optimized technological solutions (see Cover &Thomas, 2006; Massey, 1998). The psychological challenge aims at modeling and understanding the particular nature of an imperfect information processing system that resulted not from intelligent engineering design but from millions of years of (often accidental) evolution. Nevertheless, that the system is not perfect but is instead systematically irrational does not prevent us from modeling and analyzing its modus operandi with the same analytical tools used by engineers and computer scientists. We just have to build these imperfections into the model. This would never occur to an engineer, who strives for optimization. Although the ends of engineers and psychologists are different, the means are not: both are based on the nature of information processes such as defined by information theory, and related fields such as coding theory and computer science.

As always with models, such efforts have to balance the tradeoff between clarity and complexity, aiming at a level of abstraction that enables the inclusion of as many empirical findings as possible, balanced with sufficient clarity and tractability to be illuminating: as simple as possible and as complex as necessary. In this sense, information theory might even turn out to be at the right level of abstraction to bridge the neurological basis of information processing (e.g. Berger, 2003; Berger & Levy, 2010; Borst & Theunissen, 1999) with observable psychological effects and therefore provide a long-sought theoretical language and set of analytical tools to bridge the gap between neuroscience and psychology.

So This is it?

When making a decision, is it that we retrieve what we have stored in memory, and because we systematically "mess up" during the storage and retrieval processes, our judgments turn out to be predictably irrational? Are there no higher cognitive functions involved in these eight cognitive biases? No emotions, motivations, or unfathomable feelings? No homunculus in our mind fooling us? Is it simply an almost mechanical flaw in the design of the system, reminiscent of a sloppily constructed information processing machine? It is as simple as that?

Occam's razor—a fundamental principle of science—would argue yes, this is it: "among the theories that are consistent with the observed phenomena, one should select the simplest theory" (see Li &Vitanyi, 2008, p.341).[17] We cannot deny the possibility that additional heuristics, emotions, or social influences might also play a role in the explanation of the discussed biases. But they are in sensu stricto not necessary to explain them. It might, of course, also be the case that the model will have to be expanded in order to increase its explanatory power to include other biases. In other words, additional explanations can explain parts of the biases that mixing noise cannot explain. Occam's razor is based on the assumption that the theory is "consistent with the observed phenomena", which supposes "all other things being equal". Models with larger explanatory power can be more complex than ones with less explanatory power and still fulfill Occam's razor.

Of course, the integration of additional biases might require additional methodological assumptions. Dougherty et al. (1999) have shown that it is possible to replicate the empirical regularities of additional cognitive biases through computer simulations like MINERVA, which essentially follow the channel logic presented here (see Appendix B). They replicate the empirical regularities of the hindsight bias (Fischhoff, 1975), the availability heuristic (Tversky & Kahneman, 1973), the representativeness heuristic (Kahneman &Tversky, 1973), the conjunction fallacy (Tversky & Kahneman, 1983), base-rate neglect (Bar-Hillel, 1980), the validity effect (Hasher, Goldstein, & Toppino 1977), and simulation (Kahneman & Tversky, 1982). It should be possible to determine the underlying mathematical properties of the noisy memory channel that lead to these results with help of our presented framework. Other heuristics, such as framing (Tversky &

Kahneman, 1981) or anchoring and adjustment (Kahneman et.al., 1982) might need more profound extension of the model, such as (sequentially dependent) stochastic search processes inside memory (such as those that can be modeled with simple Markov-chains with a certain kind of memory; e.g., Cover & Thomas, 2006, Ch. 4). These potential extensions notwithstanding, noisy storage and retrieval from memory is certainly part of the human decision-making process, given that any process of cognition (or even perception) always requires some kind of internal information representation (some kind of "memory"), and information processes are always noisy, at least to some degree.

In this sense, the properties proposed in this article should not be seen as final but as a starting point for future empirical testing and theoretical discussion. We are far from understanding the concrete properties of the human information processing system involved in decision-making. In the long term, the overall contribution of this article will certainly not consist in arguing in favor of one or another specific property but in showing the logic of systematically defining such properties. It might well turn out that noise in our minds has a different distribution than the ones proposed here, or that a larger class of channels exists, affording explanation of these and many other empirical findings, including the processing of non-numerical cognitive chunks.

Nevertheless, the overall logic of the approach presented here is sound: if human beings employ a physical information processing system that gives rise to particular empirically detectable input-output combinations (cognitive biases), it must be the inner design of the system that gives rise to these particularities. One can analyze possible designs and test for them. Given that different biases are produced by the same information processing system, the overall design of the system should reveal how the diverse biases are related, and vice versa. If part of our irrationality is understood as a consequence of the peculiar design of our noisy human information processing system, a better understanding of its properties will eventually provide us with insights to normatively improve our judgment and decision-making.

References

Adams, P. A., & Adams, J. K. (1960). Confidence in the recognition and reproduction of words difficult to spell. *The American Journal of Psychology*, *73*(4), 544-552.

Albert, R., & Barabasi, A.-L. (2002). Statistical mechanics of complex networks. *Reviews of Modern Physics*, *74*(1), 47. doi:10.1103/RevModPhys.74.47

Ariely, D. (2009). *Predictably irrational: The hidden forces that shape our decisions* (1st ed.). New York: HarperCollins.

Attneave, F. (1953). Psychological probability as a function of experienced frequency. *Journal of Experimental Psychology*, *46*(2), 81-86.

Baddeley, A., Eysenck, M. W., & Anderson, M. C. (2009). *Memory* (1st ed.). New York: Psychology Press.

Bar-Hillel, M. (1980). The base-rate fallacy in probability judgments. *Acta Psychologica*, *44*(3), 211-233. doi:10.1016/0001-6918(80)90046-3

Baron, J. (2007). *Thinking and deciding* (4th ed.). New York: Cambridge University Press.

Bar-Tal, D., Graumann, C. F., Kruglanski, A. W., & Stroebe, W. (1989). *Stereotyping and prejudice: Changing conceptions* (1st ed.). New York: Springer.

Bearden, N., & Wallsten, T. (2004). MINERVA-DM and subadditive frequency judgments. *Journal of Behavioral Decision Making*, *17*(5), 349.

Berger, T., & Levy, W. B. (2010). A mathematical theory of energy efficient neural computation and communication. *Information Theory, IEEE Transactions on*, *56*(2), 852-874. doi:10.1109/TIT.2009.2037089

Berger, T. (2003, March). Living information theory: The 2002 Shannon lecture. *IEEE information theory society newsletter*, *53*(1). Retrieved from http://www.itsoc.org/publications/newsletters/past-newsletters/itNL0303.pdf/view

Borst, A., & Theunissen, F. E. (1999). Information theory and neural coding. *Nat Neurosci*, 2(11), 947-957. doi:10.1038/14731

Brunswik, E. (1952). The conceptual framework of psychology. *International Encyclopedia of Unified Science*, Chicago: University of Chicago Press, *1*(10), IV + 102.

Budescu, D. V., Wallsten, T. S., & Au, W. T. (1997c). On the importance of random error in the study of probability judgment. Part II: Applying the Stochastic Judgment Model to Detect Systematic Trends. *Journal of Behavioral Decision Making*, *10*(3), 173-188. doi:10.1002/(SICI)1099-0771(199709)10:3<173::AID-BDM261>3.0.CO;2-6

Budescu, D., Erev, I., & Wallsten, T. (1997b). On the Importance of Random Error in the Study of Probability Judgment. Part I: New Theoretical Developments. *Journal of Behavioral Decision Making*, *10*(3), 157-171. doi:10.1002/(SICI)1099-0771(199709)10:3<157::AID-BDM260>3.0.CO;2-#

Budescu, D., Erev, I., Wallsten, T., & Yates, F. (1997a). Introduction to this Special Issue on stochastic and cognitive models of confidence. *Journal of Behavioral Decision Making*, *10*(3), 153-155. doi:10.1002/(SICI)1099-0771(199709)10:3<153::AID-BDM279>3.0.CO;2-K

Cooper, A. C., Woo, C. Y., & Dunkelberg, W. C. (1988). Entrepreneurs' perceived chances for success. *Journal of Business Venturing*, Journal of Business Venturing, *3*(2), 97-108.

Cover, T. M., & Thomas, J. A. (2006). *Elements of information theory* (2nd ed.). Hoboken, NJ: Wiley-Interscience.

Dougherty, M. R. (2001). Integration of the ecological and error models of overconfidence using a multiple-trace memory model. *Journal of Experimental Psychology. General, 130*(4), 579-599.

Dougherty, M. R. P., Gettys, C. F., & Ogden, E. E. (1999). MINERVA-DM: A memory processes model for judgments of likelihood. *Psychological Review, 106*(1), 180-209.

DuCharme, W. M. (1970). Response bias explanation of conservative human inference. *Journal of Experimental Psychology, 85*(1), 66-74.

Edwards, W. (1954). The theory of decision making. *Psychological Bulletin, 51*(4), 380-417. doi:10.1037/h0053870

Edwards, W. (1968). Conservatism in human information processing. In B. Kleinmuntz (Ed.), *Formal representation of human judgment,* (pp. 17-52). New York: Wiley.

Erev, I., Wallsten, T. S., & Budescu, D. V. (1994). Simultaneous over- and underconfidence: The role of error in judgment processes. *Psychological Review, 101*(3), 519-527.

Fiedler, K. (1991). The tricky nature of skewed frequency tables: An information loss account of distinctiveness-based illusory correlations. *Journal of Personality and Social Psychology, 60*(1), 24-36.

Fiedler, K. (1996). Explaining and simulating judgment biases as an aggregation phenomenon in probabilistic, multiple-cue environments. *Psychological Review, 103*(1), 193-214.

Fiedler, K., & Armbruster, T. (1994). Two halfs may be more than one whole: Category-split effects on frequency illusions. *Journal of Personality and Social Psychology, 66*(4), 633-645.

Fiedler, K., Unkelbach, C., & Freytag, P. (2009). On splitting and merging categories: a regression account of subadditivity. *Memory & Cognition, 37*(4), 383-393. doi:10.3758/MC.37.4.383

Fischer, I., & Budescu, D. V. (2005). When do those who know more also know more about how much they know? The development of confidence and performance in categorical decision tasks. *Organizational Behavior and Human Decision Processes, 98*(1), 39-53.

Fischhoff, B. (2003). Hindsight ≠ foresight: the effect of outcome knowledge on judgment under uncertainty. *Quality and Safety in Health Care, 12*(4), 304 -311. doi:10.1136/qhc.12.4.304

Fischhoff, B., Slovic, P., & Lichtenstein, S. (1977a). Knowing with certainty: The appropriateness of extreme confidence. *Journal of Experimental Psychology: Human Perception and Performance, 3*(4), 552-564.

Fischhoff, B., Slovic, P., & Lichtenstein, S. (1977b). Knowing with certainty: The appropriateness of extreme confidence. *Journal of Experimental Psychology: Human Perception and Performance, 3*(4), 552-564. doi:10.1037/0096-1523.3.4.552

Fox C.R., & Levav J. (2000). Familiarity bias and belief reversal in relative likelihood judgment. *Organizational Behavior and Human Decision Processes, 82*, 268-292. doi:10.1006/obhd.2000.2898

Freedman, D., Pisani, R., & Purves, R. (2007). *Statistics, 4th Edition* (4th ed.). New York: W.W. Norton & Co.

Gallileo, G. (1623). *The Assayer*, translated by Stillman Drake, Discoveries and Opinions of

Galileo (New York: Doubleday & Co., 1957), 231-280 ed. original Rome. Retrieved from http://www.princeton.edu/~hos/h291/assayer.htm

Gell-Mann, M. (1995a). What is complexity? *Complexity*, *1*(1). Retrieved from http://www.scribd.com/doc/7887206/COMPLEXITY-by-Murray-Gell-Mann

Gell-Mann, M. (1995b). *The quark and the jaguar: Adventures in the simple and the complex*. New York: St. Martin's Griffin.

Goldstein, D. G., & Gigerenzer, G. (2002). Models of ecological rationality: the recognition heuristic. *Psychological Review*, *109*(1), 75-90.

Goldstein, E. B. (2002). *Sensation and perception, 6th (Sixth) Edition* (6th ed.). Belmont: Wadsworth.

Goldstein, E. B. (2005). *Cognitive psychology - Connecting mind, research, and everyday experience.* Belmont: Thomson/Wadsworth.

Greene, R. L. (1988). Generation effects in frequency judgment. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *14*(2), 298-304.

Greenspan, A. (2008). *The financial crisis and the role of federal regulators: Hearing before the committee on oversight and government reform, statement of Alan Greenspan, former chairman of the Federal Reserve Board* (No. 110-209). House of Representatives, One hundret tenth Congress, second session. Retrieved from http://frwebgate.access.gpo.gov/cgi-bin/getdoc.cgi?dbname=110_house_hearings&docid=f:55764.pdf

Hamilton, D. L., Dugan, P. M., & Trolier, T. K. (1985). The formation of stereotypic beliefs: Further evidence for distinctiveness-based illusory correlations. *Journal of Personality and Social Psychology. Vol. 48(1), 48*(1), 5-17.

Hamilton, D. L., & Gifford, R. K. (1976). Illusory correlation in interpersonal perception: A cognitive basis of stereotypic judgments. *Journal of Experimental Social Psychology*, *12*(4), 392-407. doi:10.1016/S0022-1031(76)80006-6

Hasher, L., Goldstein, D., & Toppino, T. (1977). Frequency and the conference of referential validity. *Journal of Verbal Learning and Verbal Behavior*, *16*(1), 107-112. doi:10.1016/S0022-5371(77)80012-1

Hasher, L., & Zacks, R. (1984). Automatic processing of fundamental information: The case of frequency of occurrence. *American Psychologist*, *39*(12), 1372-1388. doi:10.1037/0003-066X.39.12.1372

Hastie, R., & Dawes, R. M. (2001). *Rational choice in an uncertain world: the psychology of judgment and decision making*. Thousand Oaks: SAGE.

Heeger, D. (1997). *Signal detection theory*. New York. Retrieved from http://www.cns.nyu.edu/~david/handouts/sdt/sdt.html

Hintzman, D. L. (1969). Apparent frequency as a function of frequency and the spacing of repetitions. *Journal of Experimental Psychology*, *80*(1), 139-145. doi:10.1037/h0027133

Hintzman, D. L. (1984). MINERVA 2: A simulation model of human memory. *Behavior Research Methods, Instruments, & Computers*, *16*, 96-101.

Hintzman, D. L. (1988). Judgments of frequency and recognition memory in a multiple-trace memory model. *Psychological Review*, *95*(4), 528-551. doi:10.1037/0033-295X.95.4.528

Hockley, W. E. (1984). Retrieval of item frequency information in a continuous memory task.

Memory & Cognition, *12*(3), 229-242. doi:http://www.psychonomic.org/backissues/9947/mc/vol12-3/mc-12-229.pdf

Howell, W. C. (1973). Storage of events and event frequencies: A comparison of two paradigms in memory. *Journal of Experimental Psychology*, *98*(2), 260-263. doi:10.1037/h0034380

Jaynes, E. T. (1957a). Information theory and statistical mechanics. *Physical Review*, *106*(4), 620. doi:10.1103/PhysRev.106.620

Jaynes, E. T. (1957b). Information theory and statistical mechanics. II. *Physical Review*, *108*(2), 171. doi:10.1103/PhysRev.108.171

Jones, R., Scott, J., Solernou, J., Noble, A., Fiala, J., & Miller, K. (n.d.). Availability and formation of stereotypes. *Perceptual and Motor Skills*, *44*, 631-638. doi:10.2466/PMS.44.2.631-638

Juslin, P, Winman, A., & Olsson, H. (2000). Naive empiricism and dogmatism in confidence research: a critical examination of the hard-easy effect. *Psychological Review*, *107*(2), 384-396.

Juslin, P., Olsson, Henrik, & Bjorkman, M. (1997). Brunswikian and Thurstonian origins of bias in probability assessment: On the interpretation of stochastic components of judgment. *Journal of Behavioral Decision Making*, *10*(3), 189-209. doi:10.1002/(SICI)1099-0771(199709)10:3<189::AID-BDM258>3.0.CO;2-4

Kahneman, D., Slovic, P., & Tversky, A. (1982). *Judgment under uncertainty: Heuristics and biases* (1st ed.). Cambridge University Press.

Kahneman, D., & Tversky, A. (1973). On the psychology of prediction. *Psychological Review*, *80*(4), 237-251.

Kahneman, D., & Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica*, *47*(2), 263-291.

Kahneman, D., & Tversky, A. (1982). The simulation heuristic. In D. Kahneman, P. Slovic, & A. Tversky (Eds.), *Judgment under uncertainty: Heuristics and biases* (pp. 201-208). Cambridge and New York: Cambridge University Press.

Kahneman, D., & Tversky, A. (1996). On the reality of cognitive illusions. *Psychological Review*, *103*(3), 582-591.

Kanouse, D., & Hanson, R. (1972). *Negativity in evaluations*. Morristown, N.J: General Learning Press.

Kaufman, E. L., Lord, M. W., Reese, T. W., & Volkmann, J. (1949). The discrimination of visual number. *The American Journal of Psychology*, *62*(4), 498-525.

Keren, G. (1988). On the ability of monitoring non-veridical perceptions and uncertain knowledge: Some calibration studies. *Acta Psychologica*, *67*(2), 95-119. doi:10.1016/0001-6918(88)90007-8

Keren, G. (1997). On the calibration of probability judgments: Some critical comments and alternative perspectives. *Journal of Behavioral Decision Making*, *10*(3), 269-278. doi:10.1002/(SICI)1099-0771(199709)10:3<269::AID-BDM281>3.0.CO;2-L

Kruger, J. (1999). Lake Wobegon be gone! The "below-average effect" and the egocentric nature of comparative ability judgments. *Journal of Personality and Social Psychology*, *77*(2), 221-232.

Kruger, J., & Dunning, D. (1999). Unskilled and unaware of it: How difficulties in recognizing

one's own incompetence lead to inflated self-assessments. *Journal of Personality and Social Psychology*, *77*(6), 1121-1134. doi:10.1037/0022-3514.77.6.1121

Labouvie, E. W. (1982). The concept of change and regression toward the mean. *Psychological Bulletin*, *92*(1), 251-257. doi:10.1037/0033-2909.92.1.251

Landauer, R. (1991). Information is physical. *Physics Today*, *44*(5), 23-29. doi:10.1063/1.881299

Li, M., & Vitányi, P. M. B. (2008). *An introduction to kolmogorov complexity and its applications* (3rd ed.). New York: Springer.

Liberman, V., & Tversky, A. (1993). On the evaluation of probability judgments: Calibration, resolution, and monotonicity. *Psychological Bulletin, 114*(1), 162-173.

Lichtenstein, S., & Fischhoff, B. (1977). Do those who know more also know more about how much they know? *Organizational Behavior and Human Performance*, *20*(2), 159-183. doi:10.1016/0030-5073(77)90001-0

Lichtenstein, S., Fischhoff, B., & Phillips, D. (1982). Calibration of probabilities: The state of the art to 1980. In D. Kahneman, P. Slovic, & A. Tversky (Eds.), *Judgment under uncertainty: Heuristics and biases*. Cambridge and New York: Cambridge University Press.

Linville, P. W., Fischer, G. W., & Salovey, P. (1989). Perceived distributions of the characteristics of in-group and out-group members: Empirical evidence and a computer simulation. *Journal of Personality and Social Psychology, 57*(2), 165-188.

Loewenstein, G. F., Weber, E. U., Hsee, C. K., & Welch, N. (2001). Risk as feelings. *Psychological Bulletin*, *127*(2), 267-286. doi:10.1037/0033-2909.127.2.267

Luce, R. D. (2003). Whatever happened to information theory in psychology? *Review of General Psychology, 7*(2), 183-188.

MacGregor, D., Lichtenstein, S., & Slovic, P. (1988). Structuring knowledge retrieval: An analysis of decomposed quantitative judgments. *Organizational Behavior and Human Decision Processes*, *42*(3), 303-323. doi:10.1016/0749-5978(88)90003-9

Massey, J. (1998). *Applied digital information theory: Lecture notes by Prof. em. J. L. Massey*. Swiss Federal Institute of Technology. Retrieved from http://www.isiweb.ee.ethz.ch/archive/massey_scr/

McClelland, A., & Bolger, F. (1994). The calibration of subjective probabilities: Theories and models 1980-94. In G. Wright & P. Ayton (Eds.), *Subjective Probability* (1st ed., pp. 453-482). John Wiley & Sons.

McClelland, J. L., & Rumelhart, D. E. (1985). Distributed memory and the representation of general and specific information. *Journal of Experimental Psychology: General. Vol. 114(2), 114*(2), 159-188.

Medin, D. L., & Schaffer, M. M. (1978). Context theory of classification learning. *Psychological Review, 85*(3), 207-238.

Meehl, P. E. (1986). Causes and effects of my disturbing little book. *Journal of Personality Assessment*, *50*(3), 370. doi:10.1207/s15327752jpa5003_6

Merkle, E. C. (2009). The disutility of the hard-easy effect in choice confidence. *Psychonomic Bulletin & Review*, *16*(1), 204-213. doi:10.3758/PBR.16.1.204

Metcalfe, J. (1990). Composite holographic associative recall model (CHARM) and blended memories in eyewitness testimony. *Journal of Experimental Psychology: General, 119*(2), 145-160.

Meyer, C. D. (2001). *Matrix analysis and applied linear algebra*. SIAM: Society for Industrial and Applied Mathematics.

Miller, G. A. (1956). The magical number seven, plus or minus two: some limits on our capacity for processing information. *Psychological Review, 63*(2), 81-97.

Moore, D. A., & Healy, P. J. (2008). The trouble with overconfidence. *Psychological Review, 115*(2), 502-517.

Moore, D. A., & Cain, D. M. (2007). Overconfidence and underconfidence: When and why people underestimate (and overestimate) the competition. *Organizational Behavior and Human Decision Processes, 103*(2), 197-213. doi:10.1016/j.obhdp.2006.09.002

Mullen, B., & Johnson, C. (n.d.). Distinctiveness-based illusory correlations and stereotyping: A meta-analytic integration. *British Journal of Social Psychology*, *29*, 11-28.

Newman, M. E. J. (2003). The structure and function of complex networks. *SIAM REVIEW, 45*, 167--256.

Newman, M. (2010). *Networks: An introduction*. Oxford University Press, USA.

Pfister, H.-R., & Böhm, G. (2008). The multiplicity of emotions: A framework of emotional functions in decision making. *Judgment and Decision Making*, *3*, 5-17.

Phillips, L.D., Hays, W. L., & Edwards, W. (1966). Conservatism in complex probabilistic inference. *Human Factors in Electronics, IEEE Transactions on*, *HFE-7*(1), 7-18.

Phillips, Lawrence D, & Edwards, Ward. (1966). Conservatism in a simple probability inference task. *Journal of Experimental Psychology, 72*(3), 346-354.

Pierce, J. R. (1980). *An introduction to information theory* (2nd ed.). Dover Publications.

Pryor, J. B. (1986). The influence of different encoding sets upon the formation of illusory correlations and group impressions. *Personality and Social Psychology Bulletin*, *12*(2), 216 -226. doi:10.1177/0146167286122008

Redelmeier, D. A., Koehler, D. J., Liberman, V., & Tversky, A. (1995). Probability judgment in medicine: Discounting unspecified possibilities. *Medical Decision Making*, *15*(3), 227-230. doi:10.1177/0272989X9501500305

Rubenstein, A. (1998). *Modeling bounded rationality* (free online edition.). Cambridge, MA: MIT Press. Retrieved from http://arielrubinstein.tau.ac.il/book-br.html

Sande, G. N., Goethals, G. R., & Radloff, C. E. (1988). Perceiving one's own traits and others': The multifaceted self. *Journal of Personality and Social Psychology*, *54*(1), 13-20.

Shah, A. K., & Oppenheimer, D. M. (2008). Heuristics made easy: An effort-reduction framework. *Psychological Bulletin*, *134*(2), 207-222. doi:10.1037/0033-2909.134.2.207

Shannon, C. (1948). A mathematical theory of communication. *Bell System Technical Journal*, *27*, 379-423, 623-656. doi:10.1145/584091.584093

Sieck, W. R., & Yates, J. F. (2001). Overconfidence effects in category learning: a comparison of connectionist and exemplar memory models. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *27*(4), 1003-1021.

Simon, H. A. (1955). A behavioral model of rational choice. *The Quarterly Journal of Economics*, *69*(1), 99 -118. doi:10.2307/1884852

Simon, H. A. (1956). Rational choice and the structure of the environment. *Psychological Review, 63*(2), 129-138.

Smith, E. R. (1991). Illusory correlation in a simulated exemplar-based memory. *Journal of Experimental Social Psychology*, *27*(2), 107-123. doi:10.1016/0022-1031(91)90017-Z

Spears, R., Van der Pligt, J., & Eiser, J. R. (1985). Illusory correlation in the perception of group attitudes. *Journal of Personality and Social Psychology*, *48*(4), 863-875.

Spears, R., Van der Pligt, J., & Eiser, J. R. (1986). Generalizing the illusory correlation effect. *Journal of Personality and Social Psychology*, *51*(6), 1127-1134.

Strogatz, S. H. (2001). *Nonlinear dynamics and chaos: With applications to physics, biology, chemistry, and engineering* (1st ed.). Cambridge: Westview Press.

Suantak, L., Bolger, F., & Ferrell, W. R. (1996). The hard-easy effect in subjective probability calibration. *Organizational Behavior and Human Decision Processes*, *67*, 201-221.

Swets, J. (1964). *Signal detection and recognition by human observers*. New York: Wiley.

Tversky, A., & Kahneman, D. (1973). Availability: A heuristic for judging frequency and probability. *Cognitive Psychology*, *5*(2), 207-232.

Tversky, A., & Kahneman, D. (1981). The framing of decisions and the psychology of choice. *Science*, *211*(4481), 453-458. doi:10.1126/science.7455683

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science*, *185*(4157), 1124 -1131. doi:10.1126/science.185.4157.1124

Tversky, A., & Kahneman, D. (1983). Extensional versus intuitive reasoning: The conjunction fallacy in probability judgment. *Psychological Review*, *90*(4), 293-315.

Tversky, A., & Kahneman, D. (1992). Advances in prospect theory: Cumulative representation of uncertainty. *Journal of Risk and Uncertainty*, *5*(4), 297-323. doi:10.1007/BF00122574

Tversky, A., & Koehler, D. J. (1994). Support theory: A nonextensional representation of subjective probability. *Psychological Review*, *101*(4), 547-567.

Wagenaar, W. A., & Keren, G. B. (1985). Calibration of probability assessments by professional blackjack dealers, statistical experts, and lay people. *Organizational Behavior and Human Decision Processes*, *36*(3), 406-416.

Wallsten, T. S., & González-Vallejo, C. (1994). Statement verification: A stochastic model of judgment and response. *Psychological Review*, *101*(3), 490-504.

Wang, X. T., Simons, F., & Brédart, S. (2001). Social cues and verbal framing in risky choice. *Journal of Behavioral Decision Making*, *14*(1), 1-15. doi:10.1002/1099-0771(200101)14:1<1::AID-BDM361>3.0.CO;2-N

Whittlesea, B. W. A. (1983). *Representation and generalization of concepts: The abstractive and episodic perspectives evaluated*. Unpublished doctoral dissertation, MacMaster University.

Wilkinson, N. (2007). *An introduction to behavioral economics: A guide for students*. New York: Palgrave Macmillan.

Witteman, C. L., Renooij, S., & Koele, P. (2007). Medicine in words and numbers: a cross-sectional survey comparing probability assessment scales. *BMC Medical Informatics and Decision Making*, *7*, 13-13. doi:10.1186/1472-6947-7-13

Yechiam, E., Druyan, M., & Ert, E. (2008). Observing others' behavior and risk taking in decisions from experience. *Judgment and Decision Making*, *3*(7), 493-500.

Zuroff, D. C. (1989). Judgments of frequency of social stimuli: How schematic is person memory? *Journal of Personality and Social Psychology*, *56*(6), 890-898.

Appendices

Appendix A: introductory analogy to memory-channel based decision-making schematizations

This appendix gives an introduction to the logic of noise influenced memory-based decision making models and to the kinds of schematizations used to present them. It will help us to set the stage for what is to follow in the article.

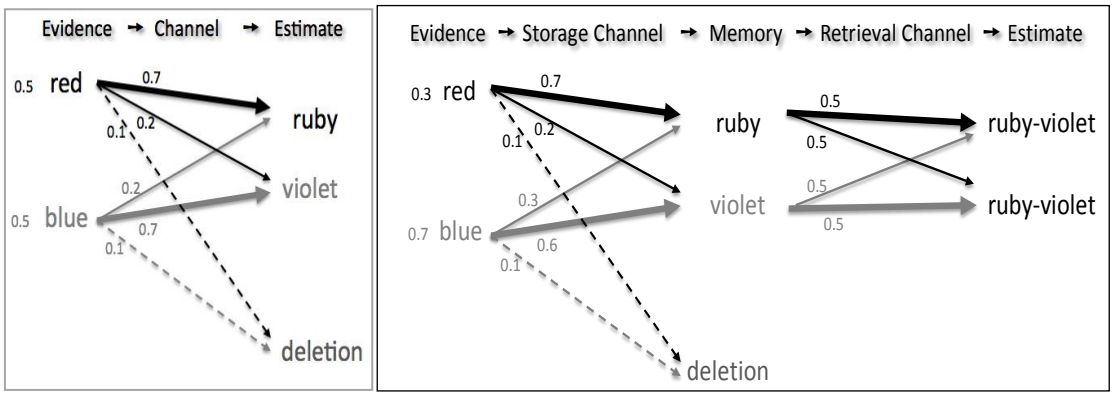
Let us start with a memory-based task. Suppose we would like to make a decision about the "redness" of a red object that we are given. Our strategy would consist in collecting and storing color traces which are purely red and once we are asked to judge redness, we would go to our storage room and pick up some of our prototypical red traces and compare their color with the object to be judged. In case our storage and retrieval processes would be perfect, we would be able to make a perfect judgment about the redness of the object. If this process is not perfect, we might erroneously end up with a sample that is not completely red and our judgment will be biased. The analysis of our storage and retrieval habits will show us the nature of this bias and be a first step in looking for strategies to minimize it.

In other words, when confronted with a memory-based decision-making task, the judge sends a cognitive probe to memory and compares it to existing memory traces. The content of what is found in memory will provide the judge with the answer to the decision problem. The process is not perfect. We will refer to the "confusion" and "mistakes" in this process as "noise". The bias of the judgment can be traced back to two possible sources: one is a biased sample in memory (which results from the noise in the storage channel); and the other one is biased sampling from memory (which results from the noise in the retrieval channel). The combination of the storage and retrieval channels constitutes the overall memory channel. We assume that the channel has certain properties that we would like to define.

In information theory it is customary to present these kinds of channels in a diagram similar to the ones shown in Figure A.1 (see Massey, 1998, Ch.4; Cover and Thomas, 2006, Ch.7). Figure A.1a essentially tells us that the noisy channel mixes blue and red input evidences. As long as the original still prevails, this will turn the red into a ruby and blue into violet. We depicted the noise with crossover arrows. The little numbers next to the arrows tell us about the respective transition probabilities involved in the process. Besides mixing evidence, we have to consider that not all input might make it. The effect is equal to deleting parts of our sample. We start out with equal amounts of red and blue [50% each]. 70% of each goes straight through the channel (we call this the identify transformation), 20% of each color is mixed with the other color (noise) and 10% of each is deleted.

Figure A.1b opens up the overall memory channel and shows that the overall memory channel actually consists of two different sub-channels. The noisy storage channel is followed by the noisy retrieval channel. In Figure A.1b, we start off with less red than blue [0.3, 0.7], and it is more likely to confuse blue with red [0.3], than red with blue [0.2]. The retrieval channel is in a special state of "highest uncertainty/entropy" (the uniform distribution), which leads to a homogeneous output estimate of ruby-violet, independent of the input evidence.

Figure A.1: Two first examples of the memory channel: (a) overall memory channel; (b) opened up into storage and retrieval subchannels



Source: Author.

It is important to note that the intermediate memory step might be very short. Several perceptual tasks rely on sensory memory that corresponds approximately to the initial 200–500 milliseconds after an item is perceived. Therefore, the process might not appear as schematic as presented here (storing in memory, then sending a probe to memory, etc.), but rather as one process. Notwithstanding, without having anything impregnated in any kind of (whatsoever short and instable) memory, no perception could occur. Therefore, memory (of some kind) always makes part of any kind of judgment and decision-making.

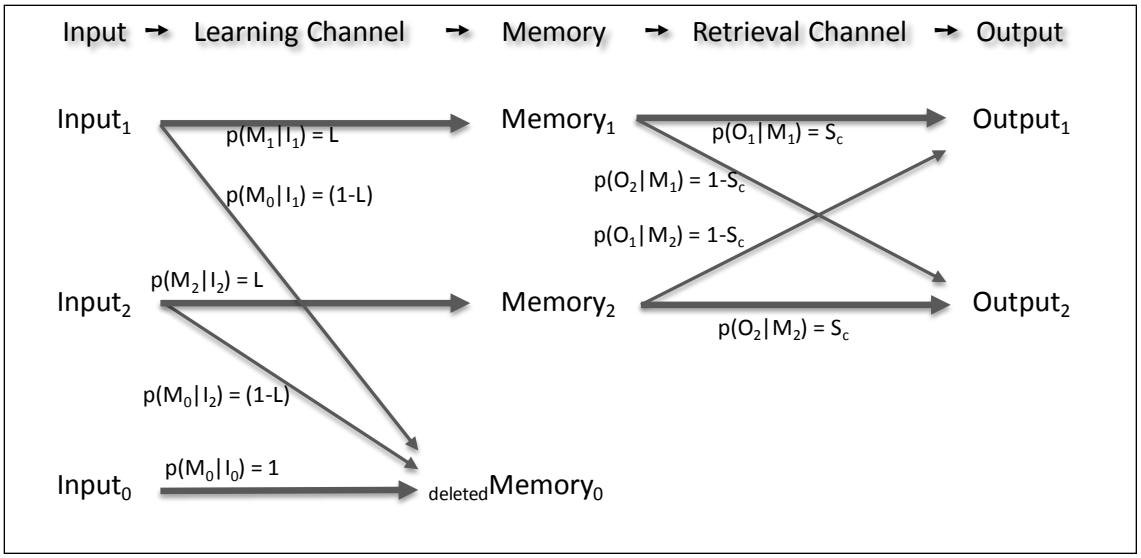
Throughout the article, we will identify which kind of noise is requires in the overall channel (Figure A.1a) to replicate six cognitive biases, and which kind of noise is necessary in the retrieval channel (Figure A.1b) to replicate two additional biases.

Appendix B: the MINERVA-DM channel

In this Appendix, we discuss the essential properties of the MINERVA-DM channel (see Hintzman, 1988; Dougherty, et.al., 1999) (see Figure A.2). The goal is not to replicate the exact nature of the MINERVA-decision-making model, but to model its essential logic

with the help of an information-theoretic channel presentation (see Appendix A; more formal in Massey, 1998, Ch.4; Cover and Thomas, 2006, Ch.7). Hintzman (1988) chooses a ternary input variable of -1, 0, +1, which is basically a binary alphabet, plus the possibility of deletion. The storage channel (in MINERVA called "learning channel") is implemented with what is known as a "Binary Erasure Channel" (BEC) in information theory. The retrieval channel is implemented with what is known as the "Binary Symmetric Channel" (BSC). The channel is symmetric because both identity transitions, and both noise transitions are equal, $p(O_1|M_1) = p(O_2|M_2) = S_c$, and $p(O_2|M_1) = p(O_1|M_2) = 1 - S_c$. Both are very special and important channels in information theory (Massey, 1998, Ch.4; Cover and Thomas, 2006, Ch.7). They are the simplest existing channels and their neat properties enable a straightforward analysis with nice results.

Figure A.2: Rough schematization of the MINERVA-DM model as a memory channel



Source: Author, based on the logic presented in Hintzman, 1988 and Doughterty, et.al, 1999.

The technical details of the specific implementation of MINERVA-DM are more involved than this simplified schematization. One aspect is that Hintzman (1988) chose a multi-trace memory model to implement MINERVA. He also chose not to apply the noise to an entire memory trace, but to its constituents, which he calls features. He uses a ternary code (-1, 0, +1) to represent the value of each feature, which make up the content of specific memories. Each of these features is passed through the channel, which has different probabilities of converting a -1 into a +1, and vice versa, or deleting it, which means converting it to 0. As the features change, the content of the memory trace change and it can even lead to the fact that the memory does not represent anymore what it originally meant to represent. The rate with which the content of the memory traces change, depends on the

transition probabilities that convert the values of the features and on the criterion that defines when a code in a memory matches or not.

MINERVA also includes a particular matching process (which is not further justified by the authors). It basically replicates what is known as the Hamming distance between codewords in information theory. These particular specifications do not change the basic properties of the MINERVA-DM channel, which follows the logic of a BEC followed by a BSC: the storage/learning channel can delete input, and the retrieval channel has the possibility that "false friends" sneak into the final judgment.

Studying the logic of the MINERVA-DM channel, it becomes clear that the only way that the variation of L impacts the output is through a reduction of the sample size of traces in memory, by exactly [1-L], which leads to the well-known channel capacity of the BEC: C_{BEC}= L (see Massey, 1998, Ch.4; Cover and Thomas, 2006, Ch.7). The reason is that L is applied symmetrically to all inputs and that there is no crossover possibility in the storage/learning channel. This prediction was reconfirmed by the MINERVA-DM simulation results of Dougherty, et.al., 1999, shown in their Appendix C. As pointed out by them, the effect of smaller sample sizes is increased variability (the inverse of the law of the large numbers). On contrary, the retrieval channel, which is BSC, is sensible to variations in S_C (this is also in agreement with the simulation of Appendix C, Dougherty, et.al., 1999). The smaller S_C, the larger the crossover probability. The result is that both outputs are "more similar", i.e. they are closer to their "average", which is the uniform distribution (in this binary case: 0.5-0.5). Since the retrieval channel is BSC, its channel capacity is: C_{BSC}= 1-H(S_c) (see Massey, 1998, Ch.4; Cover and Thomas, 2006, Ch.7). MINERVA-DM applies noise of the same distribution to all input evidence. As we will show in Appendixes B and D, this requirement is not necessary to assure conservatism.

Appendix C: Effects of Properties N and S on a bounded noise distribution

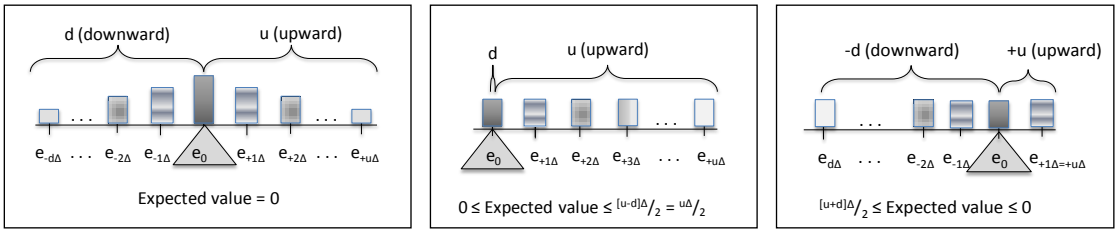
This Appendix shows how Properties N and S lead to the fact that all mean estimates Ê, based on some concrete input evidence $\underline{e_i}$, (Expct.val.[Ê|$\underline{e_i}$]), must lie somewhere in the grey areas of Figure 3b. For likelihood/probability/frequency estimates, the interval variables e_i are replaced with probabilities. In this case, the exercise focuses directly at estimating the value Expct.val.[p(p(Ê)|p($\underline{e_i}$))]. For reasons of clarity of presentation we will treat both cases identically and refer to E instead of P(E).

We use a little trick and scale the identity transition e_i to 0: $\underline{e_0}$ = $\hat{e}_0$ = 0. We denote all estimates to the "positive" side with $\hat{e}_u$, u={0,1,2…u}, and estimates to the "negative" side of the identity transition with $\hat{e}_d$, d={0,1,2…d}. We stick to the assumption of a one-dimensional equidistant interval scale, which results in: $\hat{e}_u = \hat{e}_0 + u\Delta$; $\hat{e}_d = \hat{e}_0 - d\Delta$. The result looks like Figure A.3a. Property N assures that none of the "weights" can be larger than the value assigned to e_0 (identity transition). Property N assures that the weights get smaller

the further they are away from the identity value, and Property S demands that the weights are symmetrical around the identity value. The total number of possible values is: n = u+d+1, whereas +1 counts for the identity value at 0 (note that we do not consider the forgetting / inaccessibility option here. If we would, the number of possible values would be n+1).

Visually the logic of the proof can be seen when playing around with Figures A.3. When moving the balance triangle all the way to the negative extreme (d ≤ u), the minimum value is 0 and the maximum positive value can be achieved by placing the highest possible weight on the largest possible numbers (Figure A.3b). Considering the restrictions of Property N, the uniform distribution achieves this maximum value (in this case: $^{u\Delta}/_{2}$). In Figure A.3c, d ≥ u, and since d is preceded by a minus sign, the expected value can only be negative, with: $^{[u-d]\Delta}/_{2}$ ≤ Expected value ≤ 0.

Figure A.3: (a) representation of an evidence at the middle of the possible scale in the task; (b) representation of evidence at the lowest possible value of the scale; (c) representation of evidence one step from the highest possible value on the scale.



Source: author.

In a more formal proof we first define the limits of the possible expected values (note that EV without hat and no underline, refers to "Expected Value", not to be confused with Ê: estimation; and E: evidence):

If d ≤ u, then:

$$0 \le EV\left[\hat{E}\middle|\underline{e_0}\right] \le \sum_{i=0}^{n} \hat{e}_i/n = \left[\sum_{d=0}^{d} \hat{e}_d + \sum_{u=0}^{u} \hat{e}_u\right]/n = \left[\sum_{d=0}^{d} \hat{e}_0 - d\Delta + \sum_{u=0}^{u} \hat{e}_0 + u\Delta\right]/n =$$

$$= \Delta\left[-\sum_{d=0}^{d} d + \sum_{u=0}^{u} u\right]/n = \Delta\left[-\frac{d(d+1)}{2} + \frac{u(u+1)}{2}\right] / [u+d+1] =$$

$$= \Delta\left[\frac{-d^2-d+u^2+u}{2}\right] / [u+d+1] = \frac{[u-d]\Delta}{2} = \frac{\hat{e}_u+\hat{e}_d}{2} = its\ midrange\ point\ m$$

If d ≥ u, then:

$$0 \ge EV\left[\hat{E}\middle|\underline{e_0}\right] \ge \sum_{i=1}^{n} \hat{e}_i/n = \cdots = \frac{[u-d]\Delta}{2}$$

We now reformulate the EV[Ê|e0], following Properties S and N:

$$EV\left[\hat{E}\middle|\underline{e_0}\right] = \sum_{j=-d}^{u} p(\hat{e}_j\,|\,\underline{e_0}) \times \hat{e}_j = \sum_{d=0}^{d} p(\hat{e}_d\,|\,\underline{e_0}) \times \hat{e}_d + \sum_{u=0}^{u} p(\hat{e}_u\,|\,\underline{e_0}) \times \hat{e}_u =$$

$$= \sum_{d=0}^{d} p(\hat{e}_d\,|\,\underline{e_0}) \times (\hat{e}_0 - d\Delta) + \sum_{u=0}^{u} p(\hat{e}_u\,|\,\underline{e_0}) \times (\hat{e}_0 + u\Delta) =$$

If d ≤ u (group all possible symmetric noises under the same sum and cancel them out):

$EV\left[\hat{E}\middle|\underline{e_0}\right] = \sum_{k=0}^{d} p(\hat{e}_k \mid \underline{e_0}) \times (\hat{e}_0 - k\Delta + \hat{e}_0 + k\Delta) + \sum_{u=d+1}^{u} p(\hat{e}_u \mid \underline{e_0}) \times (\hat{e}_0 + \Delta u) =$
$\sum_{u=d+1}^{u} p(\hat{e}_u \mid \underline{e_0}) \times \Delta u =$

≥ 0 (its minimum), achieved if $p(\hat{e}_u|\underline{e_0})=0$, for all d<u;

≤ with its maximum if $P(\hat{E}|\underline{e_0}) = 1/n$ (limited by Property N to the uniform distribution, which puts most weight on the positive extremes), at:

$\leq \frac{1}{n}\sum_{u=d+1}^{u} \Delta u = \frac{1}{u+d+1}\Delta \sum_{u=d+1}^{d+(u-d)} u = \frac{\Delta}{u+d+1}[(d+1)+(d+2)+\cdots+(d+(u-d)] =$

$= \frac{\Delta}{u+d+1}[d(u-d)+\sum_{t=1}^{u-d} t] = \frac{\Delta}{u+d+1}\left[\frac{2d(u-d)}{2}+\frac{(u-d)[(u-d)+1]}{2}\right] =$

$= \frac{\Delta}{u+d+1} \times \frac{-d^2+u^2+u-d}{2} = \frac{(u-d)\Delta}{2}$

=> if d ≤u, then $0 \leq EV[\hat{E}|\underline{e_0}] \leq [(u-d)\Delta]/2 = [\hat{e}_u+\hat{e}_d]/2$= its midrange point m, see Figure 3b.

If d ≥ u (group all possible symmetric noises under the same sum and cancel them out):

$EV\left[\hat{E}\middle|\underline{e_0}\right] = \sum_{k=0}^{u} p(\hat{e}_k \mid \underline{e_0}) \times (\hat{e}_0 - k\Delta + \hat{e}_0 + k\Delta) + \sum_{d=u+1}^{d} p(\hat{e}_d \mid \underline{e_0}) \times (\hat{e}_0 - \Delta d) =$
$\sum_{d=u+1}^{d} p(\hat{e}_d \mid \underline{e_0}) \times (-\Delta d) =$

≤ 0 (its maximum), if $p(\hat{e}_d|\underline{e_0})=0$, for all d>u;

≥ with its minimum if $P(\hat{E}|\underline{e_0}) = 1/n$ (uniform, limited by Property N), at:

$\geq \frac{1}{n}\sum_{d=u+1}^{d}(-\Delta d) = \frac{-\Delta}{d+u+1}\sum_{d=u+1}^{u+(d-u)} d = following\ the\ same\ steps\ as\ above = \frac{(u-d)\Delta}{2}$

=> if d ≥ u, then $0 \geq EV[\hat{E}|\underline{e_0}] \geq [(u-d)\Delta]/2 = [\hat{e}_u+\hat{e}_d]/2$= its midrange point m, see Figure 3b.

Appendix D: Fitting the Gaussian channel

This Appendix shows how to convert a Gaussian channel into discrete transition probabilities and how to fit it to empirical finding. Usually, the most straightforward way of fitting normal noise to empirical findings is to set up a program (for example in MATLAB) that minimizes the distance between the empirical transition matrix and the modeled transition matrix, which is defined by the cutoff criteria between the variables, and the mean and variables of the normal distribution. Least squares can be used.

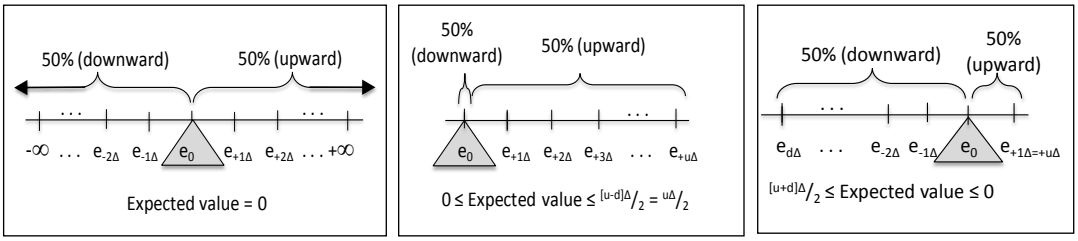
In the following I will go "manually" through the simple process of the ternary exercise of Figure 10. The problem that we face is that we have two degrees of freedom to work with when fitting the curves to the empirical data: the mean and variance of the normal distribution. However, since we suppose an equidistant interval scale (in this case $\Delta \approx 3.83\sigma$), the means of all three normal curves are defined once two means are chosen. The remaining degrees of freedom stem from the adjustable variances. This implies that it is not possible to perfectly fit the three curves to the six degrees of freedom of the empirical finding. We would not have had this problem in a binary decision-making task.

We start by arbitrarily defining that $\underline{e_1}$ fits the standard normal curve with $\mu=0$ and $\sigma=1$. We know that the identity transition of $\underline{e_1}$ is 0.971 and can therefore use the inverse of the cumulative normal distribution to identify the value x: $\Phi(1.896) \approx 0.971$, where $\Phi(x)$ is the cumulative standard normal function. We then have the freedom of choice for the mean and variance of $\underline{e_2}$ (defined by the variables of the normal probability density function: $f(x) = {}^{1}/_{\sqrt{(2\pi)}\sigma}\, e^{-(x-\mu)^2/2\sigma^2}$), as well as for the variance of $\underline{e_3}$ (once we define the mean of $\underline{e_2}$, the equidistance requirement determines the mean of $\underline{e_3}$). Instead of fitting all estimates as good as possible, I took the deliberate decision to "sacrifice" the fit of the noise-transitions $p(\hat{e}_2|\underline{e_1})$ and $p(\hat{e}_3|\underline{e_1})$, since they are very small. It turns out a mean of $\mu_2=3.83$ and variance of $\sigma_2=1.33$ for $\underline{e_2}$ and a variance of $\sigma_3=2.92$ for $\underline{e_3}$ fit to model the remaining transition probabilities of those curves. As expected, the "cost" paid by the mismatch of $\underline{e_1}$ is not very large.

Appendix E: Effects of Properties S and U on an unbounded noise distribution

This Appendix shows how Properties S and U lead to the fact that all average estimates $\hat{E}$, based on some concrete input evidence $\underline{e_i}$, (Expct.val.$[\hat{E}|\underline{e_i}]$), must lie somewhere in the grey areas of Figure 3b. As in Appendix C, we will refer to exercises that focus on estimating absolute numbers, E, but the same argument holds for estimations of discretized probabilities P(E). The basic logic of the effect of Property U can be seen when looking at what happens when we add the overshooting noise to the extremes of a symmetric distribution. A symmetric distribution around $\underline{e_i}$ has expected value = $\underline{e_i}$. We can normalize Expct.val.$[\underline{E}] = 0$ (see Figure A.4a). When the valid scale is limited on the left side at the identity (Figure A.4b), and the weight of the (formerly) negative values is added to the left-extreme value 0, the expected value $\leq {}^{u\Delta}/_{2}$. When the scale is limited to the right (Figure A.4c): Expected value $\geq {}^{[u-d]\Delta}/_{2}$. Following this logic results in the fact that the subjective estimates $\hat{E}$, must lie somewhere within the grey-shaded areas in Figure 3b.

Figure A.4: (a) normalized around 0; (b) left overshoot added to negative extreme; (c) right overshoot added to positive extreme.



Source: author.

The formal proof follows the same notations as in Appendix C, with the addition that the unbounded noise is defined by $v = \{-\infty \ldots -1, 0, +1, \ldots +\infty\}$. As in Appendix C have to proof that:

If $d \le u$ then $0 \le E[\hat{E}|\underline{e_0}] \le {}^{[(u-d)\Delta]}/_2$.

If $d \ge u$ then $0 \ge E[\hat{E}|\underline{e_0}] \ge {}^{[(u-d)\Delta]}/_2$.

We now reformulate the $EV[\hat{E}|\underline{e_0}]$, following Properties S and U:

$$EV\left[\hat{E}\middle|\underline{e_0}\right] = \sum_{j=-d}^{u} p(\hat{e}_j \mid \underline{e_0}) \times \hat{e}_j + \sum_{v<-d}^{-\infty} p(\hat{e}_v \mid \underline{e_0}) \times \hat{e}_d + \sum_{v>u}^{+\infty} p(\hat{e}_v \mid \underline{e_0}) \times \hat{e}_u =$$
$$\sum_{d=0}^{d} p(\hat{e}_d \mid \underline{e_0}) \times (\hat{e}_0 - d\Delta) + \sum_{u=0}^{u} p(\hat{e}_u \mid \underline{e_0}) \times (\hat{e}_0 + u\Delta) + \sum_{v<-d}^{-\infty} p(\hat{e}_v \mid \underline{e_0}) \times \hat{e}_d + \sum_{v>u}^{+\infty} p(\hat{e}_v \mid \underline{e_0}) \times \hat{e}_u =$$

If $d \le u$ (group all possible symmetric noises under the same sum and cancel them out):

$$= \sum_{k=0}^{d} p(\hat{e}_k \mid \underline{e_0}) \times (\hat{e}_0 - k\Delta + \hat{e}_0 + k\Delta) + \sum_{u=d+1}^{u} p(\hat{e}_u \mid \underline{e_0}) \times (\hat{e}_0 - d\Delta + \hat{e}_0 + \Delta u) +$$
$$\sum_{v>u}^{+\infty} p(\hat{e}_v \mid \underline{e_0}) \times (\hat{e}_0 - v + \hat{e}_0 + v) = \sum_{u=d+1}^{u} p(\hat{e}_u \mid \underline{e_0}) \times (u - d)\Delta$$

Whereas d is a constant $\le$ u and u are positive integers.

≥ 0 (its minimum) if $p(\hat{e}_u|\underline{e_0})=0$, for all $d \le u$;

$\le$ its maximum at: $0.5 \times (u\text{-}d)\Delta = {}^{(u-d)\Delta}/_2 = [\hat{e}_u+\hat{e}_d]/_2$= its midrange point, see Figure 3b (given symmetry of unbounded distribution, the maximum possible weight on the positive u-side is 0.5).

If $d \ge u$ (group all possible symmetric noises under the same sum and cancel them out):

$$= following\ t\square e\ same\ steps\ as\ above = \sum_{d=u+1}^{d} p(\hat{e}_d \mid \underline{e_0}) \times (u - d)\Delta$$

Whereas u is a constant $\le$ d and d positive integers.

≤ 0 (its maximum) if $p(\hat{e}_u|\underline{e_0})=0$, for all $d \ge u$;

$\ge$ its minimum at: $0.5 \times (u\text{-}d)\Delta = {}^{(u-d)\Delta}/_2 = [\hat{e}_u+\hat{e}_d]/_2$= its midrange point, see Figure 7 (given symmetry of unbounded distribution, the maximum possible weight on the negative d-side is 0.5).

Appendix F: Effects of Properties D and N

We proof that a single-peak unimodal (Property N) noise distribution that has a doubly stochastic transition matrix (Property D) results in regressive behavior for any kind of input distribution (conservatism). We start with the reformulation of our conservatism requirement, equation (III):

$0 \le cov(\underline{E},\hat{E}) \le Var(\underline{E})$

$0 \le EV[\underline{E}\hat{E}] - EV[\underline{E}] * EV[\hat{E}] \le EV[\underline{E}^2] - (EV[\underline{E}])^2$; (**Error! Bookmark not defined.**)

For n-ary decision-making tasks, scale to:

$EV[\underline{E}] = \sum_{k=1}^{n} p(\underline{e_k})\, \underline{e_k} = 0$; whereas $\underline{e_k}$ are positive numbers (in case of likelihood estimates $\sum_{k=1}^{n} \underline{e_k} = 1$ and $p(\underline{e_k})$ is uniform with $p(\underline{e_k}) = \frac{1}{n}$ (like in Hockley's exercise, or, for likelihood estimates, one can imagine that likelihoods are represented by n memory traces, each representing the likelihood with its respective value $\underline{e_k}$).

$$\Rightarrow\ 0 \le EV[\underline{E}\hat{E}] \le EV[\underline{E}^2];$$

$0 \leq \sum_{k=1}^{n} \sum_{j=1}^{n} p(\underline{e}_k \hat{e}_j)\, \underline{e}_k \hat{e}_j \leq \sum_{k=1}^{n} p(\underline{e}_k)\, \underline{e}_k^2$;

$0 \leq \sum_{k=1}^{n} \sum_{j=1}^{n} p(\underline{e}_k)\, p(\hat{e}_j|\underline{e}_k)\, \underline{e}_k \hat{e}_j \leq \sum_{k=1}^{n} p(\underline{e}_k)\, \underline{e}_k^2$; multiplied with n;

$0 \leq \sum_{k=1}^{n} \sum_{j=1}^{n} p(\hat{e}_j|\underline{e}_k)\, \underline{e}_k \hat{e}_j \leq \sum_{k=1}^{n} \underline{e}_k^2$; whereas $p(\hat{e}_j|\underline{e}_k)$are the doubly stochastic weights of the transition matrix, with $\sum_{j=1}^{n} p(\hat{e}_j|\underline{e}_k) = 1$ and $\sum_{k=1}^{n} p(\hat{e}_j|\underline{e}_k) = 1$.

First, we focus on the left side of the inequality, for which we will show that Properties N and D assure that the resulting correlation cannot be negative:

$0 \leq \sum_{k=1}^{n} \sum_{j=1}^{n} p(\hat{e}_j|\underline{e}_k)\, \underline{e}_k \hat{e}_j = \sum_{k=1}^{n} \underline{e}_k \sum_{j=1}^{n} \hat{e}_j\, [p(\hat{e}_k|\underline{e}_k) - d_j]$; whereas d_j consists of positive numbers which represent how much smaller the noise is than the identity transition $p(\hat{e}_k|\underline{e}_k)$. Note that, according to Property N, $d_j = 0$ at $d_{j=k}$, and increases with j being more distant from k.

$=> \sum_{k=1}^{n} \underline{e}_k \left[\sum_{j=1}^{n} \hat{e}_j\, p(\hat{e}_k|\underline{e}_k) - \sum_{j=1}^{n} \hat{e}_j\, d_j\right] = \sum_{k=1}^{n} \underline{e}_k \sum_{j=1}^{n} \hat{e}_j\ (-d_j)$;

Note that there are positive and negative values of $\hat{e}_j$, since $EV[\underline{\mathrm{E}}] = EV[\hat{\underline{\mathrm{E}}}] = 0$, therefore let $\sum_{j=1}^{n} \hat{e}_j = \left[\sum_{j=-=-j}^{0} \hat{e}_{j-}\right] + \left[\sum_{j+=0}^{+j} \hat{e}_{j+}\right]$, with |-j|+|+j|=n, whereas $\hat{e}_{j-}$ denotes all negative values of $\hat{e}_j$, and $\hat{e}_{j+}$ stands for all positive values of $\hat{e}_j$. Since both parts have the equal weight, it is possible to rearrange both sums and organize them in according to equal distributions ("weigh them against each other", which we call w), with $-[\sum_{j-=-j}^{0} \hat{e}_{j-}] = [\sum_{j+=0}^{+j} \hat{e}_{j+}] = \sum_{m=1}^{m} w_m$, with w always being positive. Let d_{m-} correspond to d_j of $\hat{e}_{j-}$ and d_{m+} correspond to d_j of $\hat{e}_{j+}$:

$=> \sum_{k=1}^{n} \underline{e}_k\, [\sum_{j-=-j}^{0} \hat{e}_{j-}(-d_{j-}) + \sum_{j+=0}^{+j} \hat{e}_{j+}(-d_{j+})] = \sum_{k=1}^{n} \underline{e}_k\, [\sum_{m=1}^{n} w_m(d_{m-}) + \sum_{m=1}^{n} w_m(-d_{m+})] = \sum_{k=1}^{n} \underline{e}_k \sum_{m=1}^{n} w_m(d_{m-} - d_{m+}) = \sum_{k=1}^{n} \underline{e}_k \sum_{m=1}^{n} w_m(d_{+/-})$.

For negative $\underline{e}_k$: $d_{+/-}$ is also negative for all w_m with m≥k (given Property N), but not necessarily for w_m with m<k (since noise is not symmetric around identity m=k). Likewise, for positive $\underline{e}_k$: $d_{+/-}$ is also positive for all w_m with m≤k, but not necessarily for w_m with m>k. However, when rearranging to $\sum_{m=1}^{n} w_m \sum_{k=1}^{n} \underline{e}_k\, (d_{+/-})$, we can see (since $EV[\underline{\mathrm{E}}] = \sum_{k=1}^{n} \underline{e}_k = 0$), that it is impossible that these eventualities drag the second sum into the negative

$=> \sum_{m=1}^{n} w_m \sum_{k=1}^{n} \underline{e}_k\, d_{+/-} \geq 0$.

This shows that the correlation cannot be negative ($0 \leq EV[\underline{\mathrm{E}}\hat{\underline{\mathrm{E}}}]$). The right side of our initial inequality from equation (III) can be shown with similar reformulations, but actually, this proof and its insight are not new. It is very well known in information theory that a doubly stochastic transition matrix converts the channel input in a way that the output is overall closer to its mean ("stochastic mixing increases entropy") (see Cover and Thomas, 2006, Ch.4, p. 88, Exercise 4.1). We have defined conservatism as the output being "closer to the mean" than the input (see equation (II), formulated in variance). The new part is, that in our case, we claim that conservatism also implies a positive correlation between input and output (out estimates have "more to do with the evidence than they don't"). Property N assures this, as shown above.

Footnotes

[1] Information theory is a branch of applied probability theory and is nowadays mainly taught in Electrical Engineering and Communication Departments. It is the rare breed of a branch of science that can almost exclusively be traced back to one single and groundbreaking paper: Claude Shannon's (1948) "A mathematical theory of communication". For an introduction see Pierce (1980). For a more formal approach see the introductory lecture notes of Massey (1998), which might be an easier read than the standard textbook in Engineering Departments, Cover and Thomas (2006), which is more complete.

[2] For example, the transition probability $p(\hat{e}_1|\underline{e}_1)$ is the dot product of the two transition vectors $p(M|\underline{e}_1)$ and $p(\hat{e}_1|M)$. In this ternary case: $p(\hat{e}_1|\underline{e}_1) = [p(m_1|\underline{e}_1) * p(\hat{e}_1|m_1)] + [p(m_2|\underline{e}_1) * p(\hat{e}_1|m_2)] + [p(m_0|\underline{e}_1) * p(\hat{e}_1|m_0)]$. To completely understand the nature of the storage and retrieval channels, it would be necessary to know which way each evidence takes through each sub-channel.

[3] The multiplication rule applies to calculate the probabilities of the random variable $\hat{E}$. For example, $p(\hat{e}_1) = [p(\underline{e}_1)* p(\hat{e}_1|\underline{e}_1)] + [p(\underline{e}_2)* p(\hat{e}_1|\underline{e}_2)] + [p(\underline{e}_0)* p(\hat{e}_1|\underline{e}_0)]$. In words: the probability of ending up with an estimate of event $\underline{e}_1$ after passing through the memory channel, is the probability of $\underline{e}_1$ being transmitted correctly, plus the probability that noise turns $\underline{e}_2$ into $\underline{e}_1$, and the probability that $\underline{e}_1$ is perceived, even though nothing was there.

[4] The case of the outgoing branches from $\underline{e}_0$ is a special case. One would naturally expect that where there is no input, nothing should be stored in memory, and it should not affect our judgment. However, since the early 1900s, research on Gestalt heuristics has shown that our mind makes up (often completes) input where none is originally there (e.g. Goldstein, 2005: p.74). Actually, our sensations are only a small part of our perceptions. We recognize shapes and faces where there are none, and we mentally complete partially concealed objects, even though their unseen parts might be missing. In other words, crossover probabilities $p(\hat{e}_1|\underline{e}_0)$ and $p(\hat{e}_2|\underline{e}_0)$ describe pure products of our imagination, which also influence our estimates. A conceptually similar case is $\hat{e}_0$, which represents the case that an input does not make it. This can have two reasons: forgetting or inaccessibility. In the first case, the input is not (permanently) stored in memory. In the second case, it is "somewhere" in memory, but at the moment of decision making, we cannot access it (see e.g. Baddeley, Eysenck and Anderson, 2009, ch.9). Popular examples of inaccessibility include parking one's car or misplacing one's keys and using some kind of retrieval strategy to access the temporarily lost memory trace.

[5] The concept of "mutual information" is applicable for likelihood/probability estimations and is one of the core concepts of information theory: if knowing the distribution of $P(\underline{E})$ does not tell us anything about the specific values of $P(\hat{E})$, we say that both have 0 mutual information, $I(\underline{E};\hat{E})=0$ (see Massey, 1998: Ch1; Cover and Thomas, 2006: Ch.2).

[6] Note that the presentation does not consider $\underline{e}_0$ (illusion), nor $\hat{e}_0$ (forgetting/inaccessibility) (compare with Figure 2). It assumes that output estimates can only reflect valid input options. In reality, however, the estimates might include options made up by the judge (option: other-than-one-of-the-valid-responses). This is often empirically difficult to obtain, but in this case Hockley did. He mentions (1984: p.230) that of the valid 47,474 judgments, 354 referred to an (inexistent) 4th repetition. However, in line with most studies in decision-making, he unfortunately does not specify their origin and simply takes them out of the sample. In line with this general practice, in the following we will not deal with illusions and forgetting/inaccessibility.

[7] From the data in Figure 6b, the reader might want to verify that the probability of estimating a 1st repetition is 38.7%: $p(\hat{e}_1)= [p(\underline{e}_1)*p(\hat{e}_1|\underline{e}_1) + p(\underline{e}_2)*p(\hat{e}_1|\underline{e}_2) + p(\underline{e}_3)*p(\hat{e}_1|\underline{e}_3)]=0.387$. Likewise, $p(\hat{e}_2)=0.358$ and $p(\hat{e}_3)=0.255$. Equally, given that subjects estimated a 1st repetition $\hat{e}_1$, the probability that it was (originally) a 2nd repetition is 6.1%: $p(\underline{e}_2|\hat{e}_1)= [p(\underline{e}_2)*p(\hat{e}_1|\underline{e}_2)]/p(\hat{e}_1)= [p(\underline{e}_2)*p(\hat{e}_1|\underline{e}_2)]/[p(\underline{e}_1)*p(\hat{e}_1|\underline{e}_1) + p(\underline{e}_2)*p(\hat{e}_1|\underline{e}_2) + p(\underline{e}_3)*p(\hat{e}_1|\underline{e}_3)]= 0.061$. Likewise: $p(\underline{e}_1|\hat{e}_1)=0.920$ and $p(\underline{e}_3|\hat{e}_1)=0.019$.

[8] Proof by counterexample: take a quaternary input, and let $p(\hat{e}_1|\underline{e}_1)=p(\hat{e}_3|\underline{e}_1)=p(\hat{e}_4|\underline{e}_1)=1/3$; $p(\hat{e}_2|\underline{e}_2)=1$; $p(\hat{e}_3|\underline{e}_3)=1$; and $p(\hat{e}_1|\underline{e}_4)=p(\hat{e}_2|\underline{e}_4)= p(\hat{e}_1|\underline{e}_4)=1/3$.

[9] In a binary decision making task (two realizations to choose from, e.g. "right or wrong"), Property N implies that the identity transition, $p(\hat{e}_i|\underline{e}_i)$, is larger or equal to 0.5. In multiary decision-making exercises, it is possible that the sum of all crossover noises is larger than the single probability of the identity transition (e.g. $p(\hat{e}_1|\underline{e}_1) = 0.4$; $p(\hat{e}_2|\underline{e}_1) = 0.3$; $p(\hat{e}_3|\underline{e}_1) = 0.3$). However, Property N states that none of the individual noise transitions can be larger than the identity transition. In an extreme case, Property N allows for equality among all transition probabilities: the identity and noise transitions are uniform with $p(\hat{e}_i|\underline{e}_i) = p(\hat{e}_x|\underline{e}_i) = 1/n$, with n being the number of input realizations. Such channels convert our estimates in homogeneous estimates. In short: uniform noise and identity transitions lead to uniform estimates (proof: $\text{Expct.Val.}[\hat{E}|\underline{E=e}_i] = \sum_j \hat{e}_j\, p(\hat{e}_j|\underline{e}_i) = \sum_j \hat{e}_j\, [1/n\,]$). Input evidence and output estimate are independent and have no mutual information.

[10] The same can be applied to likelihoods, since the probability scale from [0-1] is an interval variable. For a exercises aiming at multiary likelihoods, the one-dimensional scale is a probability distribution between [0-1], with i representing the discrete steps of a discrete probability distribution, $0 < \Delta < 1$ and the sum of all $p(\hat{e}_i)$ summing up to 1. Equidistance implies: $p(\hat{e}_i) = p(\hat{e}_o) + i\Delta$, For example, i could be $i = \{0, 1, 2, 3\}$, $\Delta = 0.1$, and $p(\hat{e}_o) = 0.1$, which results in: $p(\hat{e}_1) = 0.1$; $p(\hat{e}_2) = 0.2$; $p(\hat{e}_3) = 0.3$; $p(\hat{e}_4) = 0.4$, with $\Sigma p(\hat{e}_i)=1$.

[11] $\sum_{i=0} e_i /[n+1] = ([e_o+0\Delta]+[e_o+1\Delta]+[e_o+2\Delta]+…+[e_o+n\Delta])/[n+1]=([n+1]*e_o+(0+1+2+…+n)\Delta]/[n+1] = e_o+\{[n(n+1)/2]\Delta\}/[n+1] = e_o+[n\Delta]/2 = e_o+[e_o+n\Delta-e_o]/2 = [e_o+e_n]/2 = m$.

[12] While typical, this question is actually not very well formulated to detect the full scope of the bias, since it is not exhaustive. A subject might have the opinion that Absinthe is neither a liqueur, nor a

precious stone. Well-formulated binary decision-making exercises should include a complement ("everything else").

[13] Following the pioneering work of Lichtenstein and Fischhoff (1977) and Fischhoff, Slovic, and Lichtenstein (1977), often studies limit the possible confidence rating to a scale of 50%-100%, which is sometimes referred to the "C50 paradigm" in literature, in comparison to a 0-100% scale, sometimes called "NC100" (see Sieck and Yates, 2001). The reasoning behind the 50%-100% scale is that a judge should have at most 50% uncertainty in a binary decision-making task: when in doubt, opt for "fifty-fifty", you cannot do worse. What Lichtenstein, Fischhoff and others intuitively integrated in this setup is what information theorists call the highest entropy state. In case of highest uncertainty, it the natural choice is to opt for the uniform distribution. Shannon (1948) would have said that the uniform distribution contains the most uncertainty (see also Massey, 1998, Ch.1; Cover and Thomas, 2006, Ch.2). The implicit integration of this important insight into the experimental set up can distort the results. For example, a judge might actually think that absinthe is a flower, and therefore could have less than 50% confidence that it is either a liqueur or a stone. The highest entropy state depends on the number of possible variables of the distribution, since the uniform distribution (the state of highest entropy) has probability: $p(\text{uniform})={}^{1}/_{[\text{number of possible values}]}$. Forcing the judge to use a 50%-100% scale can distort the proportions that the judge actually has in memory. Some empirical evidence seems to suggest that in these situations judges "squeeze" the extended 0-100% scale into a 50%-100% scale, which distorts results (e.g. see the data of Liberman and Tversky, 1993, on what they call "designated" and "inclusive" judgments, whereas Figures at p. 169 show that limited scales foster overconfidence and neglect underconfidence; see also data from Sieck and Yates, 2001). This limitation limits the possibility to detect underconfidence, which—especially in a binary decision-making task—is then squeezed onto the 0.5 turning point (this phenomena can already be seen in the original study of Fischhoff, et.al., 1977: their Figure 1).

[14] Juslin, et.al. (2000) quantify the hard-easy effect as a regression weight between the proportion correct (in our notation $P(\hat{E})$) and the difference between confidence and proportion correct (in our notation $[P(M)- P(\hat{E})]$). Empirical studies find that this relationship tends to be negative (see Merkle, 2009). Merkle (2009) follows this reasoning and expands $\text{cov}(P(\hat{E}),[P(M)- P(\hat{E})]) = \sigma_{P(\hat{E})} [r \times \sigma_{P(M)} - \sigma_{P(\hat{E})}]$, which is negative if $\sigma_{P(\hat{E})} > r \times \sigma_{P(M)}$ (see our equation II).

[15] I like to thank Prof. Gerhard Kramer (Technische Universität München, USC's Department of Electrical Engineering, IEEE Information Theory Society) for pointing out this interesting insight to me.

[16] In information theory, similarity of different codes and blockcodes (i.e. in our case, cognitive concepts and cognitive chunks) is often measures with the Hamming distance between two codewords (but any other distance measure can be used). This requires that both are written in the same code. Unfortunately, we do not have an explicit codebook that reveals the coding structure that

we use cognitively. In the long run, it is surely an indispensable undertaking to roughly map out our cognitive codebook. How distant is one concept from another? Coding theory, a branch of information theory, has developed the respective analytical tools. In social science, attribute measures are often used to classify types. A service company is distinct from an agricultural firm because of the distance in their specific attributes. The distance measure in the applied code might or might not be related to the distance in attributes.

[17] According to Bertrand Russell, the actual phrase by William of Ockham (c.1290-c.1349) was: "It is vain to do with more what can be done with fewer". Aristotle (c.384-322B.C.) anticipated Occam's insight: "We may assume the superiority ceteris paribus of the demonstration which derives from fewer postulates or hypothesis—in short, from fewer premises". Isaac Newton (1642-1727) states the principle as rule number 1 for natural philosophy in the Principia: "We are to admit no more causes of natural things than such as are both true and sufficient to explain the appearances. To this purpose the philosophers say that Nature does nothing in vain, and more is in vain when less will serve; for Nature is pleased with simplicity, and affects not the pomp of superfluous causes". Albert Einstein formulated the same idea this way: "It can scarcely be denied that the supreme goal of all theory is to make the irreducible basic elements as simple and as few as possible without having to surrender the adequate representation of a single datum of experience".