EVOLUTION AND COGNITION

General Editor: Stephen Stich, Rutgers University

Published in the Series

Simple Heuristics That Make Us Smart
Gerd Gigerenzer, Peter M. Todd, and the ABC Research Group

Natural Selection and Social Theory: Selected Papers of Robert Trivers
Robert Trivers

Adaptive Thinking: Rationality in the Real World
Gerd Gigerenzer

In Gods We Trust: The Evolutionary Landscape of Religion
Scott Atran

The Origin and Evolution of Cultures
Robert Boyd and Peter J. Richerson

The Innate Mind: Structure and Contents
Peter Carruthers, Stephen Laurence, and Stephen Stich, Eds.

The Innate Mind, Volume 2: Culture and Cognition
Peter Carruthers, Stephen Laurence, and Stephen Stich, Eds.

The Innate Mind, Volume 3: Foundations and the Future
Peter Carruthers, Stephen Laurence, and Stephen Stich, Eds.

Why Humans Cooperate: A Cultural and Evolutionary Explanation
Natalie Henrich and Joseph Henrich

Rationality for Mortals: How People Cope with Uncertainty
Gerd Gigerenzer

Ecological Rationality: Intelligence in the World
Peter M. Todd, Gerd Gigerenzer, and the ABC Research Group

Ecological Rationality

Intelligence in the World

Peter M. Todd
Gerd Gigerenzer
and the ABC Research Group

2012

OXFORD
UNIVERSITY PRESS

12

Why Rare Things Are Precious

How Rarity Benefits Inference

Craig R. M. McKenzie
Valerie M. Chase

> Only with the help of . . . bold conjectures can we hope to discover interesting and relevant truth.
>
> *Karl Popper*

Imagine that you have just moved to a desert town and are trying to determine if the local weather forecaster can accurately predict whether it will be sunny or rainy. The forecaster often predicts sunshine and rarely predicts rain. On one day, you observe that the forecaster predicts sunshine and is correct. On another day, she predicts rain and is correct. Which of these correct predictions would leave you more convinced that the forecaster can accurately predict the weather? According to a variety of information-theoretic accounts, including Bayesian statistics, the more informative of the two observations is the correct prediction of rain (Horwich, 1982; Howson & Urbach, 1989). As we show in more detail later, this is because a correct prediction of sunshine is not surprising in the desert, where it is sunny almost every day. That is, even if the forecaster knew only that the desert is sunny, you would expect her to make lots of correct predictions of sunshine just by chance. Because rainy days are rare in the desert, a correct prediction of rain is less likely to occur by chance and therefore provides stronger evidence that the forecaster can distinguish between future sunny and rainy days. The same reasoning applies to incorrect predictions: Those that are least likely to occur by chance alone are most informative with respect to the forecaster's (in)accuracy.

In short, rarity is valuable. Whether your expectations are confirmed or violated as a result, observing a rare conjunction of events is more revealing than observing a common one. Trying to assess the desert forecaster's accuracy by checking the weather only after she predicts sunshine would be like looking for the proverbial

needle in a haystack: Because nearly every day is sunny, the more informative rainy days would be few and far between. Of course, if you had nothing else to do, you could compare her daily forecasts of rain or sunshine with the actual weather for hundreds of days in succession in order to assess her performance. But in case you do have other things to do, it would be a lot easier just to wait until a rainy day and check whether the forecaster predicted that day's weather correctly. Event rarity matters in the real world because people are boundedly rational; that is, they have limited time, limited opportunities to gather information, and limited cognitive capacities for processing information. Gravitating toward rare events like rainy days in the desert enables people to zero in quickly on the most information-rich regions of their environment. Of course, what is rare depends on the specific setting. For instance, if the forecaster were predicting weather in a rain forest rather than a desert, then, assuming the forecaster usually predicts rain, correctly predicting sunshine would be rarer and therefore more informative than correctly predicting rain.

Given that rare conjunctions of events are more informative than common ones, a question naturally arises: Are people sensitive to event rarity when making inferences? Anecdotal evidence that at least some people are comes from observing scientists, who strive to predict events that are unlikely a priori, presumably because they believe that correct predictions of unlikely events provide relatively strong support for their hypothesis or theory. Consider, for example, Galileo's surprising—and famously correct—prediction that light and heavy objects fall at the same rate. Of course, scientists may be sensitive to rarity when conducting research not because it is intuitive but because it is prescribed by some philosophers of science (e.g., Lakatos, 1978). That is, professional researchers might behave differently from people making inferences in their everyday lives. Are laypeople also influenced by rarity?

In this chapter, we review evidence showing that people are remarkably sensitive to the rarity of events when making inferences. Indeed, people are so attuned to event rarity that their implicit assumptions about rarity guide their thinking even in laboratory tasks where experimenters have implicitly assumed that rarity would not matter. Participants' sensitivity to, and assumptions about, rarity have important implications for understanding lay inference.

Much as physicists study falling objects in a vacuum, psychologists who study intuitive inference typically present participants with tasks that are abstract or unfamiliar in an attempt to eliminate real-world influences that are not of theoretical interest. Viewing the experimental tasks from this perspective, psychologists often turn to content- and context-independent models of inference—such as logic or probability theory—to determine what constitutes optimal, or rational, responses in the task. Because participants' behavior consistently departs from the predictions of these models, it has been generally concluded that people are poor inference makers. Psychologists have only recently begun to realize that, faced with laboratory tasks stripped of content and context, participants fall back on *ecologically rational assumptions*, that is, default assumptions based on their real-world experience. The mismatch between these assumptions and the content- and context-free tasks presented to them in the laboratory can make their *adaptive* behavior in these experiments appear irrational (for reviews, see Funder, 1987; Hogarth, 1981; McKenzie, 2005). When observed in laboratory tasks in which, unbeknownst to participants, these assumptions are violated, lay inference can look maladaptive.

An important assumption about task environments that is made by experts and laypeople alike is that events that stand out and are therefore spoken and thought about—in the context of weather forecasting, personal health, corporate performance, or any other realm—are generally rare rather than common (see this also in the context of recognized vs. unrecognized objects in chapter 5). We argue that it is adaptive for people to make this *rarity assumption* in situations without information to the contrary, because it reflects the ecology of the real world (see also Dawes, 1993; Einhorn & Hogarth, 1986; Klayman & Ha, 1987; Oaksford & Chater, 1994). But we also present a wide range of evidence that people's behavior is *adaptable* in the sense that it is sensitive to violations of the rarity assumption (McKenzie, 2005). In other words, when the rarity assumption clearly does not hold, people's behavior changes largely in accord with Bayesian prescriptions, often erasing inferential "errors" or "biases."

In the next section, we define rarity more precisely and illustrate the normative importance of rarity in inference. In the four sections thereafter, we demonstrate the psychological importance of rarity when people assess covariation between events, evaluate hypotheses after receiving data, and search for information about causes and effects. Hypothesis testing and covariation judgment have been major research topics over the past few decades, but only recently has it become evident that participants' assumptions and knowledge about rarity strongly influence their behavior. After reviewing the evidence, we argue that, despite the computational complexity assumed by a Bayesian analysis, simply being influenced more by rare events than by common ones is a boundedly rational strategy for making inferences that is qualitatively consistent with Bayesian norms.

A Bayesian Analysis of Rarity

What makes an event or observation rare? Because we are concerned with events that either do or do not occur, we define an event as rare if it is absent more often than not, that is, if it has a probability of occurrence of less than .50. Of course, events are more rare (or common) to the extent that they occur with probability closer to 0 (or 1).

Imagine again the desert weather forecaster attempting to predict rain or sunshine. The four possible observations are shown in Figure 12-1: A correct prediction of rain (Cell A), an incorrect prediction of rain (Cell B), an incorrect prediction of sunshine (Cell C), and a correct prediction of sunshine (Cell D). The column marginal probabilities indicate that rain is rare, occurring on 10% of days, and that sunshine is common, occurring on the remaining 90% of days. (We use this relatively high rate of desert rain because using smaller probabilities makes the numbers in our example inconveniently small.) The row marginal probabilities indicate that the forecaster predicts rain just as often as it occurs, that is, on 10% of days (i.e., rarely), and predicts sunshine on 90% of days.

Recall that you are trying to determine whether the forecaster can predict the weather at better than chance-level performance. The values in each cell in the left matrix in Figure 12-1 indicate the probability of each observation, given H0, the "chance-level" hypothesis (i.e., that predictions and events are independent, or that ρ, the true correlation between the forecaster's predictions and actual outcomes, is 0). Under this hypothesis, the probabilities in the cells in the left matrix in Figure 12-1 are the result of simply multiplying the

H0: Prediction and Event Independent (ρ = 0)

Prediction \ Event	Rain	Sun	
Rain	0.01 (A)	0.09 (B)	0.1
Sun	0.09 (C)	0.81 (D)	0.9
	0.1	0.9	

H1: Prediction and Event Dependent (ρ = .5)

Prediction \ Event	Rain	Sun	
Rain	0.055 (A)	0.045 (B)	0.1
Sun	0.045 (C)	0.855 (D)	0.9
	0.1	0.9	

Figure 12-1: Cell proportions when predictions and events are independent (H0; left matrix) and when there is a moderate correlation ρ between them (H1; right matrix). In both cases, rain is predicted to occur, and rain actually occurs, 10% of the time.

respective marginal probabilities, which is appropriate if the predictions and events are assumed to be independent. For example, if the forecaster merely guesses that it will rain on 10% of days and it does rain on 10% of days, the forecaster would be expected to correctly predict rain (by chance) on 1% of days (Cell A). Let the competing hypothesis, H1, be that there is a positive relationship between predictions and events (say, ρ = .5; details about computing correlations for 2 × 2 matrices can be found later in the section on covariation assessment). In this case you would expect that there is a moderate contingency between the forecaster's predictions and events rather than no contingency. The right matrix in Figure 12-1 shows the probabilities under H1.

Now we can ask how informative each of the four possible observations, or event conjunctions, is given these hypotheses. From a Bayesian perspective, data are *informative*, or *diagnostic*, to the extent that they help distinguish between the hypotheses under consideration. Informativeness can be captured using likelihood ratios. In this chapter, we concentrate on how informative a given observation is—regardless of the hypothesis it favors—in situations where the qualitative impact of each observation is clear (A and D observations always favor one hypothesis, and B and C observations always favor the other).

Let the numerator of the ratio be the probability of observing the data assuming that H1 is true, and let the denominator be the probability of observing the same data assuming that H0 is true. A datum is diagnostic to the extent that its likelihood ratio differs from 1. In this example, the likelihood ratio for a Cell A observation is $p(A \mid H1)/p(A \mid H0) = .055/.01 = 5.5$. That is, a correct prediction of rain is 5.5 times more likely if there is a moderate contingency between the forecaster's predictions and the actual events than if the forecaster is merely guessing. For the remaining cells, $p(B \mid H1)/p(B \mid H0) = p(C \mid H1)/p(C \mid H0) = .045/.09 = .5$, and $p(D \mid H1)/p(D \mid H0) = .855/.81 = 1.06$. The fact that the likelihood ratios for A and D observations (correct predictions) are greater than 1 indicates that they are evidence in favor of H1, and the likelihood ratios of less than 1 for B and C observations (incorrect predictions) show that they are evidence in favor of H0. The log likelihood ratio (LLR) is a traditional Bayesian measure that converts the likelihood ratio into bits of information: $LLR_j = \mathrm{Abs}(\log_2 |p(j \mid H1)/p(j \mid H0)|)$, where j corresponds to Cell A, B, C, or D (e.g., Evans & Over, 1996; Good, 1983; Klayman & Ha, 1987). The measure is bounded below by zero and unbounded above. For the A through D observations in this example, LLR equals 2.46, 1.0, 1.0, and 0.08 bits, respectively.

Consider first the relationship between the correct predictions of rain and sunshine, Cells A and D, respectively. Consistent with

the intuitive analysis offered earlier, the correct prediction of rain is much more informative than the correct prediction of sunshine. Indeed, the correct prediction of sunshine is virtually uninformative. Several assumptions were made in the above analysis, however, including that H1 was $\rho = .5$, H0 was $\rho = 0$, and p(predict rain) = p(rain) = .1. How sensitive to these assumptions is the result that $LLR_A > LLR_D$? As it turns out, the competing hypotheses are irrelevant. If the marginal probabilities are the same under the competing hypotheses, all that is necessary for the correct prediction of rain to be more informative than the correct prediction of sunshine is that p(predicted rain) < 1−p(rain) (McKenzie & Mikkelsen, 2007, provide a proof; see also Horwich, 1982; Mackie, 1963; McKenzie & Mikkelsen, 2000; Forster, 1994, provides a non-Bayesian account of inference in which rarity plays an important role). Thus, if rain and predictions of rain are both rare by our definition—that is, if each has a probability of less than .50—Cell A is more informative than Cell D.

What about the informativeness of the two wrong predictions? Under the assumptions outlined earlier, the two wrong predictions are equally informative. All that matters is the relationship between p(predicted rain) and p(rain). (Again, the competing hypotheses are irrelevant.) Because these two probabilities are equal in the above example, $LLR_B = LLR_C$. However, if p(predict rain) < p(rain), the wrong prediction of rain is more informative, and if p(predict rain) > p(rain), the wrong prediction of sunshine is more informative. Put differently, if the forecaster is biased to predict sunshine, then a wrong prediction of rain is the strongest disconfirming outcome, and if the forecaster is biased to predict rain, then a wrong prediction of sunshine is the strongest disconfirming outcome.

The four panels in Figure 12-2 show each of the four cells' informativeness (LLR_j) as a function of p(predict rain) and p(rain), which were orthogonally varied between .1 and .9 in steps of .1 (resulting in 81 data points in each panel). H1 was $\rho = .1$ and H0 was $\rho = 0$. [The low ρ value for H1 was used because there are low upper bounds on positive values of ρ when p(predict rain) or p(rain) is low and the other is high.] The top left panel shows that a Cell A observation is most informative when both probabilities are low; the top right panel shows that Cell B is most informative when p(predict rain) is low and p(rain) is high; the bottom left panel shows that Cell C is most informative when p(predict rain) is high and p(rain) is low; and the bottom right panel shows that Cell D is most informative when both probabilities are high.

The important point is that rarity—how often the forecaster predicts rain versus sunshine and how often it is rainy versus sunny—almost single-handedly determines the informativeness of the

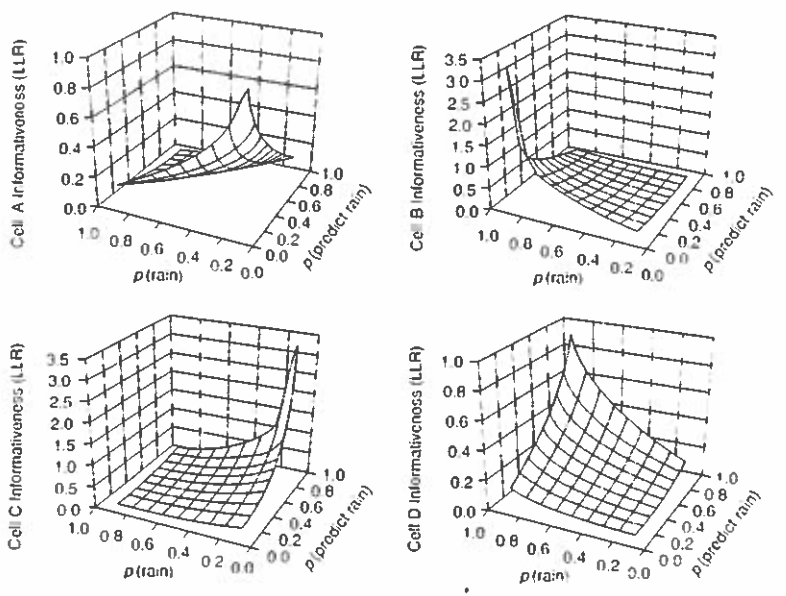

Figure 12-2: The log likelihood ratio (LLR) of a datum in each of the four cells (A, B, C, and D) as a function of p(predict rain) and p(rain). The informativeness measure used is $LLR_j = \text{Abs}(\log_2[p(j|H1)/p(j|H0)])$, where j corresponds to the particular cell. To generate the data in the figure, hypothesis H1 was that $\rho = .1$ (i.e., there was a weak positive relationship between predictions of rain and actual rain) and H0 was that $\rho = 0$ (i.e., predictions of rain and actual rain were independent).

different outcomes. Of course, this analysis generalizes beyond the forecasting example and is applicable to any situation in which one is trying to ascertain whether two binary variables are related (e.g., a symptom and a disease; handedness and a personality trait; for similar analyses, see Evans & Over, 1996; Nickerson, 1996; Oaksford & Chater, 1994).

Note that this analysis is incomplete in the sense that it considers only likelihood ratios and ignores the inference maker's degree of belief in H1 as opposed to H0 before and after observing particular events, which in Bayesian terminology are referred to as the *prior* probability and the *posterior* probability, respectively. A more complete analysis would take into account, for example, prior beliefs regarding the weather forecaster's ability to predict at better-than-chance-level performance. We emphasize the likelihood ratio

because of our interest in how people perceive data informativeness rather than how they incorporate information, once garnered, into their beliefs. Moreover, as will be shown later, not only the prior and posterior probabilities but the specific dependence hypothesis (the hypothesis specifying that there is a relationship between the variables) under consideration has surprisingly little impact on this and other measures of informativeness.

Furthermore, our analysis thus far has concentrated on the informativeness of passively witnessed outcomes. What about situations in which one must decide how to actively search for information (as discussed for cue search in chapter 10)? If you had to choose between checking whether a prediction of rain is correct and checking whether a prediction of sunshine is correct, for example, which would better help you determine whether or not the forecaster is capable? Because you do not know which outcome will occur (e.g., when checking a prediction of rain, you do not know whether you will find that it subsequently rained or was sunny), considerations of *expected* informativeness come into play. Here, too, event rarity is crucial. We present a more formal analysis of information search in a later section.

We now briefly review several areas of research in which participants' sensitivity to rarity has turned out to be key to understanding their inference-making behavior. In several cases, what have traditionally been interpreted as biases on the part of participants have turned out instead to be adaptive behavior, driven to a large extent by participants' reasonable ecological assumptions about event rarity.

Covariation Assessment

Imagine that, after moving to the desert town, you occasionally experience allergic reactions, but you do not know why. You might attempt to discern which events tend to precede the reactions. That is, you might try to figure out what events *covary*, or tend to go together, with your allergic reactions. One can think of this in terms of the familiar 2 × 2 matrix (e.g., Figure 12-1): For example, when you are around cats, how often do you have a reaction, and how often do you not have a reaction? And when you are *not* around cats, how often do you have a reaction and how often not? Accurately assessing how variables covary is crucial to our ability to learn (Hilgard & Bower, 1975), categorize objects (Smith & Medin, 1981), and judge causation (Cheng, 1997; Cheng & Novick, 1990, 1992; Einhorn & Hogarth, 1986; for reviews, see Allan, 1993; McKenzie, 1994). In a typical covariation task, participants are asked to assess whether (or how strongly) two variables, both of which can

be present or absent, are related. Consider the following scenario used by McKenzie and Mikkelsen (2007). Participants were asked to uncover the factors that determine whether people have personality type X or personality type Y. They were informed that everyone has either one personality type or the other. The factor to be examined was genotype, and participants were told that everyone has either genotype A or genotype B. To find out if there was a relationship between genotype and personality type, participants viewed records that stated whether each person had genotype A (yes or no) and personality type X (yes or no). Note that these records were described in terms of the presence and absence of genotype A and personality type X.

Participants were shown two different (purportedly random) samples of nine people, given at the top of Table 12-1 (Condition 1). The frequencies indicate the number of people falling into each category for each sample. For instance, six of the nine people in Sample 1 had genotype A and personality type X. Participants were asked whether Sample 1 or Sample 2 provided stronger support for a relationship between genotype and personality type. Most (76%) selected Sample 1, in which the large frequency corresponded to the joint presence of the two variables (the yes/yes category), traditionally labeled Cell A in the covariation literature.

In another condition (Condition 2 in Table 12-1), the labeling of the observations in terms of yes and no was reversed without altering the logical identity of each observation. Rather than indicating whether or not each person had genotype A and personality type X, the records showed whether each person had genotype B (yes or no) and personality type Y (yes or no). For example, a person identified in Condition 1 as genotype A/personality type X (Cell A) was instead identified in Condition 2 as not-genotype B/not-personality type Y (Cell D). Participants in this condition were presented with the two samples of nine people shown in Table 12-1, Condition 2.

Note that these two samples are equivalent to their counterparts presented earlier (Condition 1). For example, the two Sample 1s are the same; the categories are simply labeled differently. Nonetheless, Table 12-1 shows that participants' preferences reversed: Now most participants reported that Sample 2 provided stronger evidence of a relationship between genotype and personality type. These results replicate what has been found in numerous previous studies, namely, that the number of Cell A (joint presence) observations has a much larger impact on judgments of covariation than does the number of Cell D (joint absence) observations (Kao & Wasserman, 1993; Levin, Wasserman, & Kao, 1993; Lipe, 1990; Schustack & Sternberg, 1981; Wasserman, Dorner, & Kao, 1990). In terms of impact, the ordering of the cells is often A>B=C>D.

Table 12-1: Composition of Conditions Along With Results From McKenzie and Mikkelsen's (2007) Covariation Study

	Factor present?	Sample 1	Sample 2	Cell
Condition 1 (Abstract)				
Genotype A/Personality X	Yes/Yes	6	1	A
	Yes/No	1	1	B
	No/Yes	1	1	C
	No/No	1	6	D
Participants (%) choosing sample as strongest evidence of relationship		76.3	23.7	
Condition 2 (Abstract)				
Genotype B/Personality Y	No/No	6	1	D
	No/Yes	1	1	C
	Yes/No	1	1	B
	Yes/Yes	1	6	A
Participants (%) choosing sample as strongest evidence of relationship		26.3	73.7	
Condition 3 (Concrete)				
Disturbed/Dropout	Yes/Yes	6	1	A
	Yes/No	1	1	B
	No/Yes	1	1	C
	No/No	1	6	D
Participants (%) choosing sample as strongest evidence of relationship		73.1	26.9	
Condition 4 (Concrete)				
Healthy/Graduate	No/No	6	1	D
	No/Yes	1	1	C
	Yes/No	1	1	B
	Yes/Yes	1	6	A
Participants (%) choosing sample as strongest evidence of relationship		67.1	32.9	

Note. Sample columns indicate number of fictional people in each sample with indicated factors present or absent. Participants considered the sample in which the large frequency corresponded to Cell A (rather than Cell D) to provide the strongest evidence of a relationship—except in Condition 4, where participants knew that Cell A observations were common. In that condition, participants considered the large Cell D sample to provide the strongest support.

A model considered normative by covariation researchers is the phi coefficient: $\phi = (AD-BC)/[(A+B)(C+D)(A+C)(B+D)]^{1/2}$, where A, B, C, and D correspond to the respective cell frequencies. Phi is a special case of Pearson's product-moment correlation coefficient, ranging between −1 and 1. (Whereas ρ, discussed earlier, is a population parameter, ϕ is a sample statistic.) The closer this coefficient is to 1 (−1), the stronger the positive (negative) relationship between the variables: One variable is more (less) likely to be present when the other variable is present rather than absent. When $\phi = 0$, the variables are independent. In Table 12-1, reversing the frequencies in Cells A and D (both of which provide evidence of a positive relationship) leaves ϕ unchanged. Thus, all the samples show the same objective phi correlation, namely, .36. Because the four cells contribute equally to ϕ, their differential impact on perceived correlation has been routinely interpreted as a fallacy in people's reasoning. For example, Kao and Wasserman (1993, p. 1365) stated, "It is important to recognize that unequal utilization of cell information implies that nonnormative processes are at work," and Mandel and Lehman (1998) attempted to explain differential cell utilization in terms of a combination of two reasoning biases.

Note that the traditional normative view of the task is a logical one that leaves no room for ecological variables, such as how rare the events are. Phi is a descriptive statistic that merely summarizes the presented information. No information beyond the four cell frequencies is considered relevant; it would be considered an error if any additional information or beliefs were to influence judgment.

An ecological Bayesian account can explain, in contrast, the larger impact on perceived correlation of joint presence relative to joint absence. If it is assumed that the presence of events is rarer than their absence, then joint presence is more informative than joint absence. The assumption is that (a) the observations in the matrix are sampled from a larger population of interest, and (b) there are competing hypotheses, for example, that there is either a positive relationship ($\rho = .5$) or no relationship ($\rho = 0$) between the variables. Observing the rare observation, Cell A, distinguishes better between the competing hypotheses. If presence of the two variables were rare, then it would not be surprising to see both variables absent, a Cell D observation, even if the variables were independent. In contrast, observing their joint presence would be surprising, *especially* if the variables were independent. Joint presence provides stronger support than joint absence for the hypothesis that the variables are related.

Note, then, that if the presence of the two variables is rare, Cell A is more informative than Cell D. Furthermore, depending on the competing hypotheses, Cells B and C can fall between Cells A and

D in terms of informativeness (see Figure 12-2). Of course, this is consistent with the robust finding that, in terms of participants' reported subjective impact of different cells on judgment, the ordering is A>B=C>D. Thus, assuming that presence is rare, a normative Bayesian account can naturally explain the perceived differences in cell informativeness (see also Anderson, 1990).

Does the presence of an event of interest tend to be rarer than its absence? That is, might it be adaptive to assume that presence is rare? The answer will probably vary across specific domains, but we believe that in the vast majority of cases the answer is yes. Most things are not red, most things are not mammals, most people do not have a fever, and so on. Moreover, most things people bother to remark on—whether "remarking on" something means noticing it or communicating about it—are rare, or else they would not be worth remarking on (see chapters 4 and 15 on such skewed environment distributions and chapter 5 on what people talk about and thus recognize). We are not making a claim about metaphysics, but about how people use language. Imagine two terms, "X" and "not-X" (e.g., red things and non-red things), where there is no simple non-negated term for not-X. If (as we expect) not-X is usually a larger category than X, then it is plausible that people learn early on that the presence of an event of interest is usually rarer than its absence, and furthermore that observing the joint presence of two such events is therefore usually more informative than observing their joint absence. What looks like a bias in the laboratory might reflect deeply rooted tendencies that are highly adaptive in the real world.

Is it possible to get participants to reverse their preference for Cell A over Cell D? That is, might participants' approach to covariation assessment be adaptable as well as generally adaptive? The most likely way to demonstrate adaptability would be to use concrete variables that participants are familiar with. Ideally, participants would already know how common the levels of each variable are. Tapping into participants' real-world knowledge about rarity can have large effects on behavior in the direction predicted by the Bayesian account (McKenzie & Mikkelsen, 2000; see also McKenzie, 2006). To test this idea, McKenzie and Mikkelsen (2007) asked participants in the *concrete* condition of their experiment to imagine that they worked at a large high school and were trying to uncover factors that determine students' "high school outcome": whether they drop out or graduate. The factor being examined was students' "emotional status." All students were said to undergo a thorough psychological examination during their freshman year and to be categorized as either emotionally disturbed or emotionally healthy. Though it was assumed that participants knew that dropping out and being emotionally disturbed are both rare events, this was reinforced in the task instructions.

These concrete participants were told that they had access to the records of former students in order to find out if there was a relationship between students' emotional status and high school outcome. Half of these participants were told that each record listed whether the student was emotionally disturbed (yes or no) and whether the student dropped out (yes or no). Thus, the presence (i.e., the "yes" level) of each variable was rare, making a Cell A observation rarer than a Cell D observation. When presented with the two samples of nine observations (see Condition 3 in Table 12-1), one with many Cell A observations and one with many Cell D observations, the Bayesian account predicts the same results that have been found in earlier covariation studies, including the ones reported above: Because presence is rare in this condition, participants should find the large Cell A sample as providing stronger evidence of a relationship between emotional health and high school outcome. Indeed, this is what McKenzie and Mikkelsen (2007) found: Table 12-1 shows that more than 70% of participants selected the large Cell A sample.

The key condition was the one remaining: Some participants were presented with the same concrete scenario but simply had the labeling reversed, just as in the abstract condition (see Condition 4 in Table 12-1). Rather than indicating whether each student was emotionally disturbed and dropped out, the records indicated whether each was emotionally healthy (yes or no) and whether each graduated (yes or no). Thus, the *absence* of each of these variables was rare, making Cell A more common than Cell D. The Bayesian perspective leads to a prediction for this condition that is the opposite of all previous covariation findings: Participants will find Cell D information most informative. McKenzie and Mikkelsen (2007) again found that the results were consistent with the Bayesian account. As shown in Table 12-1, only 33% of these participants selected the sample with the large Cell A frequency as providing stronger support; that is, most found the large Cell D sample most supportive.

This is the first demonstration of such a reversal of which we are aware. The results provide strong evidence for the hypothesis that the robust Cell A bias demonstrated over the past four decades stems from (a) participants' ecological approach to the task (consistent with the Bayesian perspective), and (b) their default assumption (perhaps implicit) that presence is rare. When there is good reason to believe that absence is rare, Cell D is deemed more informative, just as the Bayesian approach predicts. Note that the behavior of both the concrete and the abstract groups is explained in terms of their sensitivity to rarity: The former exploited real-world knowledge about which observations were rare, and the latter exploited knowledge about how labeling indicates what is (usually) rare (see also McKenzie, 2004a).

Hypothesis Evaluation With Passive Observation

Suppose you are at an art museum with a friend who is unfamiliar with art, and she occasionally remarks that she likes particular pieces. Based on this information, you try to figure out what she likes, and you are beginning to think, or hypothesize, that she likes modern art. The next piece you encounter is from the Renaissance, and your friend says nothing. Would this affect your confidence in your hypothesis that your friend likes modern art?

In this example, you passively receive data and update confidence in your hypothesis. Such hypothesis *evaluation* is a passive form of hypothesis testing, to be distinguished from active hypothesis testing (discussed in the next section), where you actively choose which information to gather (e.g., you would decide which pieces to ask your friend about; for reviews, see Klayman, 1995; McKenzie, 2004b; Poletiek, 2001). Like covariation assessment, hypothesis evaluation is concerned with the passive receipt of information and can be thought of in terms of a 2×2 matrix. Your friend not commenting on a Renaissance piece could be seen as a Cell D observation, and her announcement of liking a piece of modern art could be seen as a Cell A observation. Despite some similarities, hypothesis evaluation and covariation assessment tasks differ in potentially important ways. One is that the levels of the variables in hypothesis evaluation are often symmetrical (e.g., introvert/extrovert), whereas in covariation assessment they are traditionally asymmetrical (e.g., treatment/no treatment). In addition, the task instructions are different. In hypothesis evaluation, participants are often asked to evaluate "If X, then Y" statements, whereas in covariation assessment, participants are asked to assess a relationship between variables.

Now imagine that you are a researcher investigating a possible relationship between genetics and personality type. Assume that everyone has either genotype A or genotype B and that everyone has either personality type X or personality type Y. You are evaluating the following hypothesis: "If a person has personality type Y, then he/she has genotype B" (or "Y → B"). Of the first two people you observe, one has genotype A and personality type X (which we will call AX) and one has genotype B and personality type Y (BY). Both of these observations support the hypothesis, but which do you think provides stronger support?

When McKenzie and Mikkelsen (2000) presented this unfamiliar, rather abstract task to participants, more than 70% of them chose the BY observation as most supportive when forced to choose between the BY and AX observations. Of participants asked to evaluate the hypothesis "If a person has genotype A, then he/she has personality type X" (or "A → X"), almost 80% selected the AX

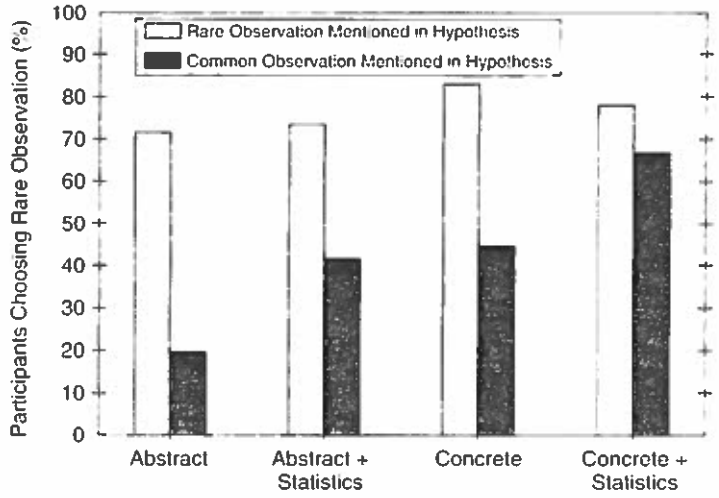

Figure 12-3: Results for the hypothesise study (McKenzie & Mikkelsen, 2000). Shown is the percentage of participants selecting the rare observation as a function of whether the task was abstract or concrete, whether statistical information about rarity/commonality was provided, and whether the rare observation was mentioned in the hypothesis. (The "abstract" group had no information about rarity.) Generally, participants were more likely to correctly select the rare observation as more informative when the task was concrete, statistical information was provided, and the rare observation was mentioned in the hypothesis. Most interesting is that participants in the "concrete + statistics" group (far right) often selected the rare observation regardless of whether it was mentioned in the hypothesis, which is in contrast to the traditional finding that observations mentioned in hypotheses are considered most informative.

observation as most supportive. The results for these two "abstract" groups are illustrated on the left side of Figure 12-3. The tall first column shows that most participants selected the BY observation when testing Y → B, and the short second column shows that few selected the BY observation when testing A → X. (Although the abstract groups had no information regarding rarity, the BY observation is referred to as the "rare observation" in Figure 12-3 for reasons that will become clear shortly.)

From the perspective of the logical approach to this problem, these participants' behavior is peculiar. The two hypotheses are logically equivalent (one is the contrapositive of the other), and therefore whichever observation supports one hypothesis most strongly must also support the other hypothesis most strongly.

Nonetheless, participants selected different observations depending on which logically equivalent hypothesis was presented to them. In particular, note that the observation mentioned in the hypothesis is usually considered most supportive in each case. That is, when testing Y → B, the BY observation is seen as most supportive, and when testing A → X, the AX observation is seen as most supportive. First demonstrated decades ago, this phenomenon is known, depending on the inferential context, as "confirmation bias," "matching bias," or "positive testing" (Evans, 1989; Fischhoff & Beyth-Marom, 1983; Klayman & Ha, 1987; McKenzie, 1994; Mynatt, Doherty, & Tweney, 1977; Wason, 1960; see also McKenzie, 1998, 1999; McKenzie, Wixted, Noelle, & Gyurjyan, 2001). It is perhaps the most commonly reported finding in the hypothesis-testing literature.

Note that the logical view of the task leaves no room for ecological variables, such as how rare the events mentioned in the hypothesis are. When testing P → Q, the logical perspective considers irrelevant what P and Q are and any knowledge the tester has about P and Q. An ecological Bayesian perspective, by contrast, leaves room for considerations such as rarity. Are lay hypothesis testers influenced by rarity information? To address this question, McKenzie and Mikkelsen (2000) told additional participants evaluating each of the above hypotheses that few people have genotype B (and most have genotype A) and few have personality type Y (and most have personality type X)—information that, from a Bayesian perspective, makes the BY observation most supportive because it is rare. As shown in Figure 12-3, these "abstract + statistics" participants were about as likely as the "abstract" participants to select the BY observation when testing Y → B (compare the first two light gray columns). However, this is not too surprising because the BY observation is mentioned in the hypothesis in both cases. More interesting are the results when the BY observation was not mentioned in the hypothesis (dark gray columns). As can be seen, the "abstract + statistics" group was about twice as likely as the "abstract" group to select the BY observation when testing A → X (compare the first two dark gray columns). That is, participants were more likely to select the *unmentioned* observation if they were told that it was rare rather than told nothing.

The above results were for abstract, unfamiliar hypotheses. Even the rarity information provided was arbitrary and probably had little meaning for participants. One might expect that sensitivity to rarity would increase when participants are presented with familiar variables that tap into their real-world knowledge regarding rarity. To this end, additional participants were told that they were researchers examining a possible relationship between mental health and AIDS. These participants tested one of two *concrete* hypotheses: "If a person is HIV-positive (HIV+), then he/she is psychotic" or "If a person is mentally healthy, then he/she is HIV-negative (HIV–)." They then selected whether a person who is HIV+ and psychotic or a person who is HIV– and mentally healthy provided stronger support for the hypothesis they were evaluating. Again, the two hypotheses are logically equivalent and both observations support both hypotheses. However, the HIV+/psychotic observation is relatively rare—and participants presumably knew this. Figure 12-3 shows that when these "concrete" participants tested "mentally healthy → HIV–" almost half of them selected the rare HIV+/psychotic person (dark gray column). That is, the unmentioned observation was often seen as most supportive if it was rare.

A final group of participants was given one of the two concrete hypotheses to evaluate but was "reminded" that few people are HIV+ (and most are HIV–) and that few are psychotic (and most are mentally healthy). Figure 12-3 shows that almost 70% of these "concrete + statistics" participants testing "mentally healthy → HIV–" selected the HIV+/psychotic person—the unmentioned observation—as most supportive (dark gray column). Regardless of which hypothesis they were testing, "concrete + statistics" participants were about equally likely to select the HIV+/psychotic person. When real-world knowledge was combined with a validation of their beliefs about rarity, participants preferred the normatively more supportive rare observation, regardless of whether it was mentioned in the hypothesis.

In short, then, these results indicate that when participants evaluate abstract, unfamiliar variables and there is no explicit information about rarity—that is, in the usual laboratory task—participants deem the mentioned confirming observation most informative. However, the unmentioned confirming observation was more likely to be chosen (a) when concrete hypotheses were used, which allowed participants to exploit their knowledge about rarity, and (b) when explicit information about rarity was provided. The combination of the concrete hypothesis and rarity "reminder" led most participants to correctly select the rare confirming observation, regardless of whether it was mentioned in the hypothesis. Knowledge about rarity—which is traditionally considered irrelevant to the task but is crucial in an ecological framework—virtually erased the common bias found in hypothesis testing.

One question remains. The above findings show that participants' hypothesis-testing strategies are adaptable in that they change in a qualitatively appropriate manner when information about rarity is provided. However, what about the apparent default strategy of deeming the mentioned confirming observation most informative? Why is this the default strategy? Is it adaptive, reflecting how the world usually works?

Indeed, one can make normative sense out of the default strategy if, when testing X1 → Y1, X1 and Y1 (the mentioned events) are assumed to be rare relative to X2 and Y2 (the unmentioned events). If this were so, then the mentioned confirming observation would be *normatively* more informative than the unmentioned confirming observation. In other words, it would be adaptive to treat mentioned observations as most informative if hypotheses tend to be phrased in terms of rare events. Do laypeople tend to phrase conditional hypotheses in terms of rare events?

Consider the following scenario: A prestigious college receives many applications but admits few applicants. Listed in Table 12-2 is information regarding five high school seniors who applied last year. Next to each applicant is a rating from the college in five categories. In each category, one candidate was rated "high" and the other four were rated "low." On the far right is shown that only one of the five candidates was accepted. Given the information, how would you complete the statement: "If applicants _____, then _____"?

You probably noticed that only SAT scores correlate perfectly with whether the applicants were rejected or accepted. Importantly, however, a choice still remains as to how to complete the statement. You could write, "If applicants have high SAT scores, then they will be accepted" or "If applicants have low SAT scores, then they will be rejected." Both are accurate, but the former phrasing targets the rare events, and the latter targets the common ones.

McKenzie, Ferreira, Mikkelsen, McDermott, and Skrable (2001) presented such a task to participants, and 88% filled in the conditional with, "If applicants have high SAT scores, then they will be accepted"—that is, they mentioned the rare rather than the common events. Another group was presented with the same task, but the college was said to be a local one that did not receive many applications and admitted most applicants. "Accepted" and "rejected" were merely reversed in the above scenario, as were "high" and "low." Everything else was the same. Now only 35% filled in the conditional with "If applicants have high SAT scores, then they will be accepted." Most participants targeted the rare events, "If applicants have low SAT scores, then they will be rejected." Thus, whether particular events were mentioned depended on whether they were rare. Virtually identical results were found using other scenarios with different content.

Table 12-2: Example of a Scenario Used to Study How People Phrase Conditional Hypotheses (McKenzie, Ferreira, et al., 2001)

	GPA	SAT scores	Letters of recommendation	Interview	Extracurricular activities	Application outcome
Alice	Low	Low	High	Low	Low	Rejected
Bill	Low	High	Low	Low	Low	Accepted
Cindy	Low	Low	Low	Low	High	Rejected
Dennis	Low	Low	Low	High	Low	Rejected
Emily	High	Low	Low	Low	Low	Rejected

Thus, people appear to have a tendency—often a very strong one—to phrase conditional hypotheses in terms of rare rather than common events. We believe this answers the question of why people consider mentioned confirming observations to be more informative than unmentioned confirming observations: Mentioned observations generally *are* more informative because they are rare.

The findings discussed earlier in this section indicate that people are sensitive to rarity when evaluating hypotheses, that is, that people's intuitive hypothesis-evaluation strategies are *adaptable.* The findings discussed immediately above indicate that a default strategy of deeming the mentioned confirming observation most informative is also *adaptive* because such observations usually are most informative in the real world (see also McKenzie, 2004b). Understanding the environmental conditions under which people typically operate, together with normative principles that make sense given these conditions, thus can help explain why people behave as they do.

Hypothesis Testing With Active Search

Suppose you think that hormone replacement therapy, which is administered to some postmenopausal women, causes breast cancer. How should you go about gathering information to test this hypothesis? For example, would it be more useful to find out what percentage of women who receive hormone replacement therapy develop breast cancer or what percentage of women with breast cancer have received hormone replacement therapy?

As every statistics textbook impresses on its readers, correlation is not causation. But experts and laypeople alike take covariation information such as that presented in Figure 12-1 into account when making inferences. Whereas previous sections have examined how people make use of passively received data, the topic of this section is how people should and do search *actively* for covariation data in testing hypotheses about cause–effect relationships.

From an ecological perspective, rarity matters as much when people actively search for information as when they observe it passively (e.g., in the hypothesis-evaluation case from the previous section).

In fact, because searching for information is costlier than merely registering it (see chapter 10 on inferences using search vs. inferences from givens), sensitivity to the relationship between informativeness and rarity would seem to be even more advantageous in active search contexts. Why expend effort looking for relatively nondiagnostic data if more diagnostic data are available? Below we explore whether people are sensitive to rarity under conditions of active information search.

Hypothesis Testing the Hard Way

Adapted from a classic reasoning problem designed by Wason (1968), the *causal selection task* simulates real-world information search in a laboratory context. In its most common form, it allows participants to perform up to four tests of a causal hypothesis relating a possible cause to an effect by examining up to four samples of events: (a) events in which the cause is known to be present (*cause test*), (b) events in which the effect is known to be present (*effect test*), (c) events in which the cause is known to be absent (*not-cause test*), and (d) events in which the effect is known to be absent (*not-effect test*). In each case, the participant will find out about the unspecified information (presence of cause or effect) in the items in the sample. If in testing your hypothesis that hormone replacement therapy causes breast cancer you chose the cause test, you could ask, say, 100 women who received long-term hormone replacement therapy (cause) whether they now have breast cancer (effect). If you chose the not-effect test, you could ask 100 women who do not have breast cancer whether they ever received long-term hormone replacement therapy. When choosing what data to gather, it is not informativeness but *expected* informativeness that you should maximize.

To understand where the "expected" in expected informativeness comes from, let us first flesh out our ecological Bayesian analysis by mapping the data relevant in causal hypothesis testing onto a 2 × 2 matrix. In each matrix depicted in Figure 12-4, the top and bottom rows correspond to the presence and absence of the cause, respectively, while the left and right columns correspond to the presence and absence of the effect, respectively. The cells representing the four possible conjunctive pairs of these events can be denoted A, B, C, and D and expressed as joint probabilities, as in Figure 12-1. Let the effect be 10 times rarer than the cause: p(cause) = .1 and p(effect) = .01. Assume for the moment that the hypothesis under test, H0, corresponds to $\rho = 0$ (see the left panel of Figure 12-4) and that the hypothesis against which it is being compared, H1, corresponds to $\rho = .1$ (right panel in Figure 12-4). (Because the cause occurs 10 times more often than the effect

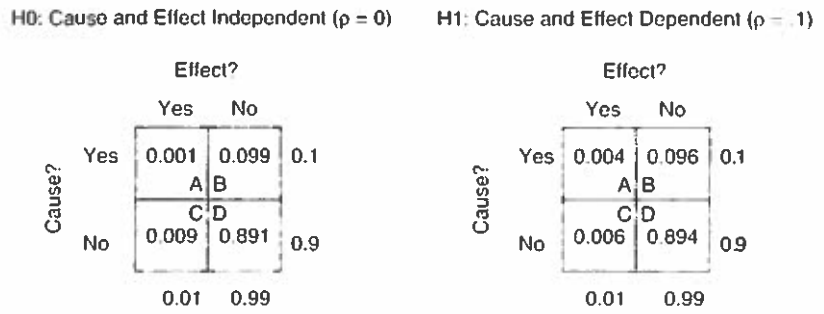

Figure 12-4: Joint probability distributions representing a causal hypothesis (H1) and its alternative (H0). In this example, hypothesis H0 is that $\rho = 0$, and H1 is that $\rho = .1$. Note that p(cause) = 0.1 and p(effect) = 0.01 regardless of which hypothesis holds.

and a correlation of 1 would mean the cause and effect always co-occur, the highest possible correlation between them is considerably less than 1—specifically, $\rho = .3$. Relative to this maximum, the correlation under H1 is thus fairly strong.) To represent the fact that the hypothesis tester has a sense of the marginal event probabilities—from sources including daily experience and media coverage—these remain the same regardless of which hypothesis holds.

Comparison of the tables in Figure 12-4 makes it clear that the data in the four cells discriminate between the hypotheses to different degrees. For example, because both the cause and the effect are rare ($p < .5$), Cell A is more informative than Cell D. The LLR for Cell A is 2, whereas that for Cell D is 0.0049.

Not only do the four cells in the 2 × 2 matrix differ with respect to how well they discriminate between hypotheses, but the four tests in the causal selection task differ with respect to the probability of revealing cases in each cell. A cause test, for example, can only reveal a case in Cell A or Cell B, whereas an effect test can only reveal a case in Cell A or Cell C; neither can uncover a case in Cell D. Moreover, although either test can turn up a case in Cell A, the probabilities of observing a case in Cell A differ between them. How can we express the probability of observing a case in Cell A—that is, the conjunction of the cause and the effect—for each test given that we do not know whether hormone replacement contributes to breast cancer (i.e., whether H0 or H1 is correct)? The answer, as always in an information-theoretic analysis of a decision problem, is to calculate an average across the hypotheses weighted by their prior probabilities.

The probability of observing a case in Cell A given that one performs a cause test, $p(\text{A} \mid \text{cause test})$, is captured by the following equation:

$$p(\text{A} \mid \text{H0} \cap \text{cause test})\, p(\text{H0}) + p(\text{A} \mid \text{H1} \cap \text{cause test})\, p(\text{H1})$$

Assuming for the moment that the prior probabilities of H0 and H1 are both .5, we obtain (1/100)(.5) + (4/100)(.5), or .025. The probability of observing a case in Cell B given that one performs a cause test is computed in the same way: (.99)(.5) + (.96)(.5), or .975. The probabilities of observing a case in Cell A and a case in Cell C given that one performs an effect test are .25 and .75, respectively.

Using the probabilities of each datum given each test and the definition of informativeness already presented, we can now compute the expected LLR of the cause test using this equation:

$$\text{Expected LLR}_{\text{cause test}} =$$
$$p(\text{A} \mid \text{cause test}) \left| \log_2 \left(\frac{p(\text{A} \mid \text{H1} \cap \text{cause test})}{p(\text{A} \mid \text{H0} \cap \text{cause test})} \right) \right|$$
$$+ p(\text{B} \mid \text{cause test}) \left| \log_2 \left(\frac{p(\text{B} \mid \text{H1} \cap \text{cause test})}{p(\text{B} \mid \text{H0} \cap \text{cause test})} \right) \right|$$

Substituting in the appropriate values, we find that the expected LLR of the cause test (which reveals either Cell A or B) is 0.093. By the same procedure, the expected LLR of the effect test (which reveals either Cell A or C) is 0.939. In this example, then, a Bayesian hypothesis tester should prefer to perform the effect test because it will reveal an average of 10 times as many bits of information as the cause test (for alternative, but nonetheless similar, measures of a test's informativeness, see Baron, 1985, chapter 4; Klayman & Ha, 1987; Nelson, 2005; Oaksford & Chater, 1994). As the equation above illustrates, the expected LLR is harder to calculate than the LLR because the expected LLR takes into account the probabilities of a hypothesis test's possible outcomes (e.g., observing a case in Cell A) as well as the informativeness of those outcomes.

We have already presented considerable evidence that people are sensitive to the relative informativeness of known data. Are they also sensitive to the relative informativeness of unknown data, for which the Bayesian calculations are considerably more complex? And if so, what processes—complex Bayesian formulae or simple heuristics—might people use to guide their information search in this setting?

Hypothesis Testing Using Rarity-Sensitive Heuristics

To address this question, participants in a series of studies by Chase (1999) were presented with scenarios involving events hypothesized to cause health risks. In each scenario, the probabilities of the effect and the possible cause were given in numerical form and manipulated between participants. In most cases, participants had to choose between performing a cause test and an effect test. The measure of primary interest was the proportion of participants who selected the cause test when the cause test had the higher expected LLR minus the proportion of participants who did so when the effect test had the higher expected LLR. A positive difference indicates sensitivity to changes in expected informativeness; a negative difference suggests a form of sensitivity to expected informativeness that departs systematically from information-theoretic prescriptions; and a difference of zero indicates insensitivity to informativeness. Chase predicted that the difference would be positive—that is, that people would be sensitive to the expected informativeness of the cause test relative to that of the effect test.

We use the results from the first of Chase's studies to illustrate the broader set of findings. Each participant received the same two scenarios, one of them involving the possible relationship between doing shift work and suffering from insomnia and the other between drinking a specific beverage and having high blood pressure. The expected LLR of the cause and effect tests was manipulated between participants such that, for each scenario, some participants received a version in which the cause test had the higher expected LLR and other participants received a version in which the effect test had the higher expected LLR. As already indicated, the expected LLR was manipulated by varying the cause and effect probabilities provided in the scenario. For example, some participants were told that the probability of shift work was .1 and the probability of insomnia was .01, while others were told that the probability of shift work was .01 and the probability of insomnia was .1. Thus, there were four unique problems, two of which were seen by each participant. Consistent with our argument that people are sensitive to the (expected) informativeness of data, the proportion of participants who chose the cause test when it had the higher expected LLR was 29 percentage points higher than when the effect test had the higher expected LLR in the shift work–insomnia scenario; in the other scenario (where it was predicted that the difference would be smaller, but still positive), the difference was 18 percentage points.

Other studies of causal hypothesis testing have likewise indicated that lay hypothesis testing reflects an at least implicit understanding of expected informativeness (for a theoretical analysis of the causal context, see Over & Jessop, 1998). In a causal selection

task similar to those used by Chase (1999), for example, Green and Over (2000) asked participants to test the hypothesis that drinking from a particular well causes cholera. Participants could choose one or more of all four tests: the cause test, the effect test, the not-cause test, and the not-effect test. The probabilities of people's drinking from the well and having cholera were manipulated between participants with the verbal labels "most" and "few" (e.g., "Most people drink from the well"). Consistent with the evidence already reviewed, Green and Over found that participants' likelihood of choosing a test increased with the test's expected informativeness. Taken together, the results indicate that people are sensitive to rarity not only when making inferences on the basis of known data, but also when deciding what data to seek.

How Boundedly Rational Minds Can Act Like Ecological Bayesians

Earlier in the chapter, we argued that (a) in the absence of knowledge about rarity, people are justified in behaving as if the events mentioned in a hypothesis are rare (McKenzie, Ferreira, et al., 2001); and (b) in the presence of knowledge about rarity, they should abandon that rarity assumption, instead searching for and weighting most heavily whatever events are rarest and therefore most diagnostic. The literature review at the beginning of this chapter indicates that people indeed seem to make a rarity assumption but that they can also adapt their behavior in contexts where it is clear that the assumption is violated (see also McKenzie & Mikkelsen, 2007). Adaptability in the causal selection task is particularly impressive because it seems to call for highly complex Bayesian calculations. But can we instead account for this within the framework of bounded rationality, as the outcome of using simple heuristics from the adaptive toolbox?

The most plausible explanation, in our view, is that people make their choice of information to seek (in test cases) using rarity as a cue to informativeness (Chase, 1999). Indeed, this behavior is consistent with philosophies of science and formal models of hypothesis testing for which the rarity of data is crucial (Poletiek, 2001, chapters 1 and 2). In the case of passively observing data in order to discriminate between competing hypotheses, one need only give more weight to rare conjunctions of events (e.g., a rare prediction of a rare outcome). Recall, for example, that joint presence provides stronger evidence of a relationship between two variables than does joint absence if the presence of the variables is rare ($p<.5$). In the case of deciding which of two hypothesis tests is more likely to reveal informative data, people need only compare the probabilities of the tests' *conditioning events*, where the conditioning event is that known to have occurred. In the causal context, for example, the conditioning event of a cause test is the cause itself: The hypothesis tester knows that the cause has occurred, and what remains to be discovered is whether the effect has also occurred. Thus, people using the *rarity heuristic* to estimate the relative informativeness of causal hypothesis tests need only look at how rare the conditioning events are relative to one another—p(cause)/p(effect)—to choose the test(s) with the highest expected informativeness (Chase, 1999). This simple heuristic enables them to behave in a way loosely consistent with the complex Bayesian calculations shown in the previous section.

Conclusion

Our review of several areas of research indicates that the rarity of events in the environment is an important factor in inference, although logical (traditionally normative) and descriptive approaches have ignored this crucial ecological variable. When assessing covariation, participants usually deem the joint presence of two events to be more informative than their joint absence. Because the four cells of a 2×2 matrix contribute equally to calculations of correlation widely considered normative, the stronger influence of joint presence has been routinely interpreted as nonnormative. A focus on joint presence makes sense from an ecological perspective, however, if the presence of events is assumed to be rarer than their absence (Anderson, 1990). Indeed, when it is made clear to participants that presence is common rather than rare, their preference for joint presence over joint absence can be reversed (McKenzie & Mikkelsen, 2007). Thus, covariation assessment is *adaptable* in that behavior changes when it is clear that presence is common, and it is *adaptive* in that, in the absence of evidence to the contrary, the rarity assumption is reasonable (see also Klayman & Ha, 1987).

When testing hypotheses, participants typically find the mentioned confirming observation to be more informative than the unmentioned confirming observation, and the traditional conclusion has been that this is an error of logic. Guided by an ecologically informed Bayesian framework, we showed that this tendency is drastically reduced when participants know that the unmentioned observation is rare (McKenzie & Mikkelsen, 2000). Thus, lay hypothesis evaluation is also adaptable: People's testing behavior changes in a qualitatively normative manner when it is clear to them that the rarity assumption (Oaksford & Chater, 1994)—which takes for granted that the observations mentioned in hypotheses are rare—is violated. In addition, the results of McKenzie,

Ferreira, et al. (2001) indicate that, as a default strategy, assuming that mentioned observations are more informative than unmentioned observations is adaptive: Hypotheses tend to be phrased in terms of rare events, and therefore mentioned observations usually *are* more informative (Grice, 1975). Finally, when testing hypotheses about cause–effect relationships, people tend to choose tests conditioned on the rarest events available. In performing those causal hypothesis tests that are most likely to discriminate between the competing hypotheses (Chase, 1999; Green & Over, 2000), people thus behave roughly in accord with much more complex Bayesian calculations.

Although the explanations and predictions of behavior in the tasks reviewed here were derived from a Bayesian perspective (together with the rarity assumption), there are good empirical reasons (e.g., McKenzie, 1994) as well as good theoretical reasons (e.g., Charniak & McDermott, 1985; Dagum & Luby, 1993) to doubt that people perform the computations prescribed by Bayesian analyses. However, to behave in a way that is qualitatively consistent with Bayesian norms, participants need only consider rare events more informative than common ones when interpreting data and seek rare over common data when choosing among conditional observations. In other words, mere sensitivity to rarity leads to behavior that is qualitatively Bayesian.

Rarity is a factor in inference that can no longer be ignored by experimenters. The results reviewed here have shown that using abstract and unfamiliar materials in an attempt to eliminate real-world interference simply leads participants to fall back on default assumptions about rarity based on how the world usually works (e.g., hypotheses mention rare events; the presence of events is rarer than their absence). Lack of awareness of this problem has led many experimenters to misinterpret adaptive responses as irrational.

Finally, our ecological account of human inference shows the importance of taking context into account, something that is unnecessary from a logical point of view. People's knowledge or assumptions about rarity are crucial, and what is perceived as rare depends on the particular decision environment. A deeper understanding of human inference can thus be achieved only through a better understanding of environmental and task structures in conjunction with decision mechanisms that make sense in a world where rare things are precious.

13

Ecological Rationality for Teams and Committees

Heuristics in Group Decision Making

Torsten Reimer
Ulrich Hoffrage

> Good decision *processes* are the best hope for good decision *outcomes*.
>
> *Jay Edward Russo and Paul Shoemaker*

When was your last meeting? By one estimate, the number of meetings held per day in the United States is more than 25 million (Massachusetts Institute of Technology, 2003), and during the 1980s executives spent, on average, as much as 40–50% of their professional time in meetings (Monge, McSween, & Wyer, 1989). In fact, meetings play a key role in today's world of business and politics, in which many decisions are formed by work teams and committees. Yet, meetings often have a bad reputation. "Meeting's over, let's get back to work"—who has not heard or made such a comment at the end of a session? Participants in meetings often report that too much of the time during their meetings is wasted. Their estimates range from a third (Green & Lazarus, 1991) to half (Monge et al., 1989; Mosvick & Nelson, 1987) of the time. Typical complaints include that meetings are often called with too short notice, last too long, and too often end without concrete results (Romano & Nunamaker, 2001).

In this chapter, we ask whether there are efficient and effective decision strategies that can be used by committees and groups to come to a joint decision. We address this question in a series of simulation studies, in which we compare the accuracy of information-laden strategies that require intense processing with the accuracy of frugal heuristics that limit information processing (Reimer & Hoffrage, 2005, 2006). We focus on one of the most popular decision rules that is often used as a default strategy when groups