Chapter 20

THE COGNITIVE ILLUSION CONTROVERSY: A METHODOLOGICAL DEBATE IN DISGUISE THAT MATTERS TO ECONOMISTS

Ralph Hertwig*
University of Basel, Basel

Andreas Ortmann
Charles University and Academy of Sciences of the Czech Republic, Prague

How do we make decisions? According to subjective expected utility (SEU) theory, which still holds sway throughout much of the social sciences, "decision makers behave *as if* utilities were assigned to outcomes, probabilities were attached states of nature, and decisions were made by taking expected utilities" (Mas-Collel, Whinston, & Green, 1995, p. 205, their emphasis). Although this is an elegant and often useful way to model decision outcomes, it imposes heroic knowledge and rationality requirements, and it clearly does not reflect the way people make decisions most of the time.[1]

Herbert Simon (1956) was the most outspoken critic of the assumption that SEU theory can be applied in any literal way to human choices. In his view, "the SEU model is a beautiful object deserving a prominent place in Plato's heaven of ideas" (1990a, p. 194); real humans, however, "have neither the facts nor the consistent structure of values nor the reasoning power at their disposal that would be required . . . to apply SEU principles" (p. 197). Simon did not limit himself to criticizing the "Olympian model" of SEU theory (Simon, 1990a, p. 198); he also proposed an alternative way to think about decision making, which he called *bounded rationality*.

Simon's vision of bounded rationality has two interlocking components: the limitations of the human mind and the informational structures of the environment in which the mind operates. Simon captured the interplay between these two components thus: "Human rational behavior . . . is shaped by a scissors whose two blades are the structure of task environments and the computational capabilities of the actor" (Simon, 1990b, p. 7). What Simon in effect argued was that rational behavior can only be understood in terms of both scissor blades: the mind and the environment. The cognitive blade requires that models of human judgment and decision-making rest on realistic assumptions about the mind's capacities rather than on idealized

competencies. Due to the mind's limitations, people "*must use approximate methods to handle most tasks*" (Simon, 1990b, p. 6, his emphasis). These methods include recognition processes that often obviate the need for information search and, when information search is necessary, simple rules for guiding and terminating search and for making a decision based on the information obtained. The environmental blade is the statistical structure of the task environment. The extent to which the approximate methods of the cognitive blade are adapted to this statistical structure determines how well they perform.

The idea that environmental and cognitive structures work in tandem is not Simon's alone. Even before Simon coined the term bounded rationality, the psychologist Egon Brunswik (1955) proposed that the processes underlying human perception and cognition are adapted to the uncertain environments in which they evolved and now function. From this premise, he challenged the standard approach to psychological experimentation on ecological grounds (for a review of neo-Brunswikian research, see Dhami, Hertwig, & Hoffrage, in press). In the standard approach, which Brunswik called *systematic design*, experimenters vary one or a few independent variables in isolation and observe resulting changes in the dependent variable(s) while holding other variables constant or allowing them to vary randomly.

Systematic design strongly emphasizes internal validity, that is, the demonstration of causal relationships between variables. Brunswik believed that this approach thereby renders impossible the primary aim of psychological research, that is, to discover probabilistic laws that describe an organism's adaptation to the causal structure of its environment. In pursuit of this aim, experimenters must preserve this structure in the stimuli that they present to participants. If they tamper with this structure, Brunswik argued, they destroy the phenomenon under investigation or at least alter psychological processes such that the experimental findings are no longer representative of people's functioning outside the laboratory.

Brunswik also observed that psychologists followed a double standard in their practice of sampling in experimental research (Brunswik, 1944). Why, he asked, are procedures for sampling participants scrutinized while findings based on stimuli in the laboratory are blithely generalized to stimuli outside the laboratory? He argued that experimental stimuli should be representative of the population of stimuli to which the experimenter intends to generalize the findings in the same way that experimental participants should be representative of the population of people whose behavior the experimenter wishes to study. As an alternative to systematic design, Brunswik proposed *representative* design, which can take any of various forms. The one he seemed to favor is achieved by randomly sampling stimuli from the defined population of stimuli or conditions, or reference class, about which the experimenter aims to make inferences.

Simon's and Brunswik's ecological views of cognition share a methodological corollary: To understand how – and how well – cognitive algorithms work, behavioral researchers need to study them under conditions that are representative of the conditions under which they usually operate. In this chapter, we show how this ecological

approach to experimentation has shed new light on findings from the *heuristics-and-biases* research program in psychology and argue that the resulting insights into cognition have important implications for experimental methods in economics.

COGNITIVE ILLUSIONS

In the early 1970s, Daniel Kahneman and Amos Tversky launched a research program that would strike a powerful blow to SEU theory as a descriptive model of human judgment and choice. The heuristics-and-biases program stresses that people have only limited "reasoning power" at their disposal, implicitly equating bounded rationality with irrationality: "Systematic, predictable differences between normative models of behavior and actual behavior occur because of what Herbert Simon . . . called 'bounded rationality' " (Thaler, 1980, p. 40). On this view, people's cognitive limitations necessitate reliance on cognitive heuristics to make judgments and choices. Although these heuristics are "highly economical and usually effective, . . . they lead to systematical and predictable errors" (Kahneman, Slovic, & Tversky, 1982, p. 20) that are variously referred to as biases, fallacies, or *cognitive illusions*.

In challenging the Olympian model of the human mind on which SEU theory rests, the heuristics-and-biases critique (e.g., Kahneman et al., 1982; Gilovich, Griffin, & Kahneman, 2002) has focused on the premise that the decision maker assigns a consistent joint probability distribution to future sets of events. This premise requires assuming that her inferences conform to the laws of probability (Schoemaker, 1982). In contrast to this premise, the heuristics-and-biases program has shown that people's probabilistic reasoning appears systematically biased and error-prone, and such biases were attributed to flawed cognitive software.

In recent years, the heuristics-and-biases program has attracted the attention of numerous social scientists, including economists (e.g., Barber & Odean, 2001; Camerer, 1995; Hirshleifer, 2001; Odean, 1999) and legal scholars (e.g., Sunstein, 2000). In fact, much of today's work in behavioral economics and behavioral finance draws inspiration and concepts from the heuristics-and-biases program (e.g., Shiller, 2000; Thaler, 1993). This attention is warranted because systematic biases may have important implications for economic behavior. In his analysis of "irrational exuberance" in the stock market during the late 1990s, for example, Shiller (2000) explicitly invoked Kahneman and Tversky's experimental results.

Even as the heuristics-and-biases program gained acceptance outside psychology, it drew criticism within psychology. Some critics suggested that the heuristics-and-biases research strategy has a built-in bias to find cognitive illusions (e.g., Krueger & Funder, 2004). Others claimed that some cognitive illusions were themselves illusory (e.g., Erev, Wallsten, & Budescu, 1994; Koehler, 1996). Perhaps the most influential objections were voiced by Gigerenzer (e.g., 1991, 1996), who argued that the heuristics onto which cognitive illusions were attributed were not precise process models; that the heuristics-and-biases program relied on a narrow definition of rationality; and that cognitive illusions can be reduced or made to disappear by representing statistical information differently than it typically had

been in heuristics-and-biases experiments. A vigorous debate ensued (see Gigerenzer, 1996; Kahneman & Tversky, 1996).

Our concern here is neither the controversy about cognitive illusions nor its implications for rationality. Instead, it is what we see as the important methodological insights that have emerged from the controversy, which can inform the choices that all behavioral experimenters wittingly or unwittingly make when they sample and represent stimuli for their experiments. We have argued elsewhere that psychologists can learn from the experimental practices of economists (e.g., Hertwig & Ortmann, 2001; Ortmann & Hertwig, 2002). In this chapter, we mine the debate in psychology about the reality of cognitive illusions for methodological lessons of relevance to experimental economists. We begin by examining how stimuli are selected from the environment for inclusion in behavioral experiments.

SAMPLING STIMULI

Many kinds of real-world economic failures have been attributed to the *overconfidence bias*. Camerer (1995, p. 594), for example, suggested that the well-documented high failure rate of small businesses may be due to overconfidence, while Barber and Odean (2001; Odean, 1999) argued that overconfidence based on misinterpretation of random sequences of successes leads some investors, typically men, to trade too much. According to Shiller (2000), "[s]ome basic tendency toward overconfidence appears to be a robust human character trait" (p. 142). These conclusions are based on the results of psychological experiments in which confidence is studied using general-knowledge questions like the following:

> Which city has more inhabitants?
> (a) Canberra (b) Adelaide
> How confident are you that your answer is correct?
> 50%, 60%, 70%, 80%, 90%, 100%

Typically, when people say they are 100% confident of their answer, the relative frequency of correct answers is only about 80%. When they are 90% confident, the proportion correct is about 75%, and so on. The size of the bias is measured as the difference between participants' mean confidence and the mean percentage of correct answers. Like many other cognitive illusions, overconfidence bias is thought to be tenacious: "Can anything be done? Not much" (Edwards & von Winterfeldt, 1986, p. 656).

But is there really so little that can be done to undo the overconfidence bias? One implication of Brunswik and Simon's idea that cognitive strategies are adapted to the statistical structure of the task environment is that if the strategies are tested in environments that are unrepresentative of that environment, they will probably perform poorly. Adopting a Brunswikian perspective, Gigerenzer, Hoffrage, and Kleinbölting (1991) argued that this is why people appear overconfident in the

laboratory. In other words, the way in which experimenters sample the questions posed to participants in overconfidence studies helps create the bias.

For illustration, let us assume that a person can retrieve only one piece of knowledge, or cue, pertaining to Australian cities, namely, whether or not a city is the national capital. How good would her inferences be if she inferred the relative population size of two Australian cities based solely on the capital cue? Consider the reference class of the 20 largest cities in Australia. Here the capital cue has an ecological validity of .74.[2] If a person's intuitive estimate of the validity of a cue approximates its ecological validity in the reference class[3] and if she uses the cue's validity as a proxy for her confidence, then her confidence judgments will be well calibrated to her knowledge. This prediction holds as long as the experimenter samples questions such that the cue's validity in the experimental item set reflects its validity in the reference class.

Gigerenzer et al. (1991) conjectured that the overconfidence effect observed in psychology studies stemmed from the fact that the researchers did not sample general-knowledge questions randomly but rather selected items in which cue-based inferences were likely to lead to incorrect choices. Suppose, for example, that an experimenter gives participants only five of the 190 possible paired comparisons of the 20 largest Australian cities: Canberra-Sydney, Canberra-Melbourne, Canberra-Brisbane, Canberra-Perth, and Canberra-Adelaide. In all these comparisons, a person who relies solely on the capital cue, (thus selecting Canberra) will go astray. In fact, if she assigns a confidence of 75% (the approximate ecological validity of the cue) to each pair, she will appear woefully overconfident, although the predictive accuracy of the capital cue is generally high. If the experimenter instead draws the pairs randomly from all possible paired comparisons of the 20 largest Australian cities, the person will no longer appear overconfident.[4] As they predicted, Gigerenzer et al. (1991, Study 1) found that when questions were randomly sampled from a defined reference class (e.g., all paired comparisons of the 83 German cities that have more than 100,000 residents) – that is, in a representative design – participants answered an average of 71.7% of the questions correctly and reported a mean confidence of 70.8%. When participants were presented with a selected set of items, as was typically the case in earlier studies, overconfidence reappeared: Participants answered an average of 52.9% of the questions correctly, and their mean confidence was 66.7%.

Recently, Juslin, Winman, and Olsson (2000) reviewed 130 overconfidence data sets to quantify the effects of representative and selected item sampling. Figure 1 depicts the overconfidence and underconfidence scores (regressed on mean confidence) observed in those studies. The overconfidence effect was, on average, large when participants were given selected samples of questions and close to zero when they were given representative samples of questions. These results hold even when one controls for item difficulty, a variable to which the disappearance of overconfidence in Gigerenzer et al.'s (1991) studies has sometimes been attributed (see Griffin & Tversky, 1992; for a different view see also Brenner, Koehler, Liberman & Tversky, 1996).

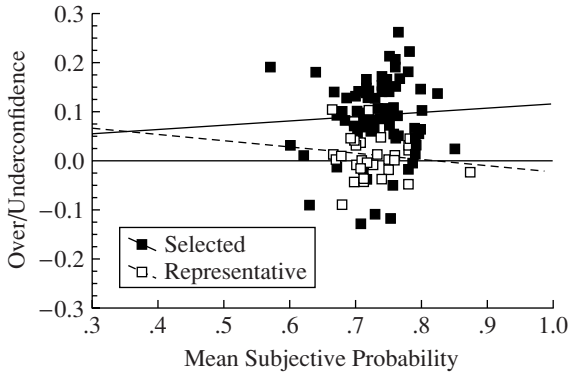

Figure 1. Regression lines relating over/underconfidence scores to mean subjective probability for systematically selected (black squares) and representative samples (open squares) (Reprint of Figure 2B from Juslin et al.*, 2000).*

The impact of item sampling on judgment and decision-making is not restricted to overconfidence. For instance, it has also been shown to affect the hindsight bias, that is, the tendency to falsely believe after the fact that one would have correctly predicted the outcome of an event. Hindsight bias is thought not only to undermine economic decision making (Bukszar & Connolly, 1988) but also to exert tremendous influence on judgments in the legal system (e.g., Sunstein, 2000; for an alternative view of the hindsight bias, see Hoffrage, Hertwig, & Gigerenzer, 2000). Like overconfidence, hindsight has been typically studied in psychology by having participants respond to general-knowledge questions.

To study the impact on hindsight of representative versus selected item sampling, Winman (1997) presented participants with selected or representative sets of general-knowledge questions such as "Which of these two countries has a higher mean life expectancy: Egypt or Bulgaria?" Before they were given an opportunity to respond, participants in the experimental group were told the correct answer (in this case, Bulgaria) and asked to identify the option they would have chosen had they not been told. Participants in the control group were not given the correct answer before they responded. If hindsight biased the responses to a given question, then the experimental group would be more likely to select the correct answer than would the control group. While this was the case, Winman also found that the size of the hindsight bias in the experimental group differed markedly as a function of item sampling: In the selected set, 42% of items elicited the hindsight bias, whereas in the representative set only 29% did so.

Using representative design, researchers have shown that cognitive illusions can be a byproduct of the slices of the world that earlier experimenters happen to take. The lesson is that methods of stimulus sampling can shape participants' performance and, by extension, inferences about human rationality. Experimenters who use selectively chosen or artificially constructed tasks in the laboratory risk altering the very

phenomena that they aim to investigate. The issue is not that selected samples are inherently more difficult to handle but that cognitive strategies are adapted to the informational structure of the environment in which they have been learned (e.g., Gigerenzer, Todd, & the ABC Research Group, 1999; Payne, Bettman, & Johnson, 1993).

DOES STIMULUS SAMPLING MATTER IN EXPERIMENTAL ECONOMICS?

The question of whether and how to sample from the environment has not been of much concern for experimental economists until recently, notwithstanding early calls for "parallelism" (e.g., Plott, 1987). Laboratory environments were typically created to test decision- or game-theoretic predictions derived from (possibly competing) formal models, with a focus on the equilibrium properties of those models. Given this research strategy, little attention was paid to how representative these environments were of their real-world counterparts. Indeed, why should it have been a concern? After all, the theories being tested were formulated to capture the essential characteristics of the world outside the laboratory.

Neglect of representative design in experimental economics was amplified by the practice of using abstract tasks. The rationale behind this methodological choice seemed to be that it would reduce the danger of eliciting participants' responses to field counterparts of the task rather than the task itself. There is now ample evidence that stripping away content and context prevents participants from applying the strategies that they use in their usual habitats. Relying mostly on evidence from psychology, Ortmann and Gigerenzer (1997) argued that experimental economists' convention of stripping the laboratory environment of content and context may be counterproductive and ought to be studied experimentally.

An early demonstration of the importance of representative design in economics was provided by economists Dyer and Kagel (1996) in an experimental investigation of the bidding behavior of executives from the commercial construction industry in one-shot common value auctions. Simple survivorship arguments suggest that such sophisticated bidders should be able to avoid the winner's curse in laboratory-based common value auctions designed to capture the essential characteristics of commercial bidding behavior. Dyer and Kagel (1996) found, however, that a significant number of the executives in their study fell victim to the winner's curse in the laboratory. The authors identified a number of differences between theoretical treatments in the literature – embodied in the experimental design – and practices in the industry that made the experimental design unrepresentative. For example, in the commercial construction industry, it seems to be possible for bidders to void the award of a contract that they realize would cost them dearly by claiming arithmetic errors. The executives' bidding behavior was maladapted to the laboratory situation because that situation failed to capture essential aspects of their natural ecology.[5]

In our view, the issue of representative design lies at the heart of discussions about the existence of altruism, defined here – in line with recent usage – as a form

of unconditional kindness (e.g., Fehr & Gächter, 2004). The debate has revolved around seemingly simple games such as symmetric and simultaneous prisoners' dilemmas (Colman, 1995); public good provision problems (Ledyard, 1995); asymmetric and sequential games such as dictator, ultimatum, and trust games (e.g., Camerer, 2003; Cox, 2004); and closely related gift exchange or principal-agent games. What these games have in common is that tests based on them seem to provide overwhelming evidence that participants are often altruistic, at least by the lights of deductive game theory as it is expounded in textbooks such as Kreps (1990) and Mas-Colell et al. (1995). Indeed, the ultimatum game "is beginning to upstage the PDG prisoner dilemma game in the freak show of human irrationality" (Colman, 2003, p. 147).

Or is it? Recall that the results that precipitated such conclusions are puzzling only if one takes as a benchmark deductive game theory's predictions for one-shot games or for finitely repeated games solvable through backward induction (Mas-Colell et al., 1995, Proposition 9.B.4). As various authors have pointed out (e.g., Hoffman, McCabe, & Smith, 1996), prisoners' dilemma, public good provision, dictator, ultimatum, trust, and gift exchange or principal agent games are typically encountered indefinitely often in the game of life. As observed by Smith (1759/1982) and Binmore (1994, 1997), the game of life is therefore played using cognitive and behavioral strategies with consequences that probably differ markedly from the dire predictions of standard deductive game theory for one-shot and finitely repeated games. In Brunswik's terms, the standard implementations of prisoners' dilemma, public good provision, dictator, ultimatum, trust, and gift exchange or principal agent games in experimental economics are unlikely to capture the conditions under which people usually encounter and make such choices. To the extent that participants perceive these games in the laboratory as some form of social dilemma, they are likely to retrieve experiences and strategies that, unbeknownst to the experimenter, change the nature of the game.

REPRESENTING STIMULI

After stimuli have been sampled, experimenters face another methodological question raised by the controversy about cognitive illusions, namely, how to represent the stimuli to participants. Just as the algorithms of a pocket calculator are tuned to Arabic rather than Roman numerals, cognitive processes are tuned to some information representations and not others (see Marr, 1982). A calculator cannot perform arithmetic operations on Roman numeral inputs, but this fact should not be taken to imply that it lacks an algorithm for multiplication. Similarly, the functioning of cognitive algorithms cannot be evaluated without considering the type of inputs for which the algorithms are designed. In their efforts to convey some aspect of reality to experimental participants, behavioral researchers use all kinds of representations, including words, pictures, and graphs. The choice of representation has far-reaching effects on the computations that a task demands and on the ease with which cognitive algorithms can carry out these operations.

The importance of task representation for cognitive performance has been extensively demonstrated in research on how people update probabilities to reflect new information. Given the importance to the SEU framework of the assumption that this updating process is Bayesian, it is not surprising that researchers in the heuristics-and-biases program have investigated the assumption's psychological plausibility. The results appear devastating for the premise that people are rational Bayesians. Time and again, experimenters found that people failed to make Bayesian inferences, even in simple situations where both the predictor and the criterion are binary. Kahneman and Tversky (1972) left no room for doubt: "Man is apparently not a conservative Bayesian: he is not Bayesian at all" (p. 450).

To get a feel for this research, consider the following study by Eddy (1982) of statistical inferences based on results of mammography tests. In the experiment, physicians received information that can be summarized as follows (the numbers are rounded):

> For a woman at age 40 who participates in routine screening, the probability of breast cancer is 0.01 [base rate, $p(\mathrm{H})$]. If a woman has breast cancer, the probability is 0.9 that she will have a positive mammogram [sensitivity, $p(\mathrm{D}|\mathrm{H})$]. If a woman does not have breast cancer, the probability is 0.1 that she will still have a positive mammogram [false-positive rate, $p(\mathrm{D}|\mathrm{not}-\mathrm{H})$]. Now imagine a randomly drawn woman from this age group with a positive mammogram. What is the probability that she actually has breast cancer?

The posterior probability $p(\mathrm{H}|\mathrm{D})$ that a woman who tests positive actually has breast cancer can be calculated using Bayes' rule, in which H stands for the hypothesis (e.g., breast cancer) and D for the datum (e.g., a positive mammogram):

$$p(\mathrm{H}|\mathrm{D}) = \frac{p(\mathrm{H})p(\mathrm{D}|\mathrm{H})}{p(\mathrm{H})p(\mathrm{D}|\mathrm{H}) + p(\mathrm{not}-\mathrm{H})p(\mathrm{D}|\mathrm{not}-\mathrm{H})}. \quad (1)$$

Inserting the statistical information from the mammography problem into Equation 1 yields:

$$\frac{(.01)(.90)}{(.01)(.90) + (.99)(.10)} \approx .08.$$

In other words, about 9 out of 10 women who receive a positive mammography result do not have breast cancer. Most of the physicians in Eddy's (1982) study overestimated the posterior probability: 95 of 100 physicians gave an average estimate of about .75. Many of them arrived at this estimate because they apparently mistook the sensitivity of the test [$p(\mathrm{D}|\mathrm{H})$] for the posterior probability $p(\mathrm{H}|\mathrm{D})$ or because they subtracted the false positive rate from 100%. Any strategy that, like these two, ignores the base rate of breast cancer can lead to the *base-rate fallacy*.

Although the reality of the base-rate fallacy has been disputed on various grounds (e.g., Koehler, 1996), let us focus on the critique that is most closely related to the ecological approach to experimentation that is the focus of this chapter. Most studies that observed the base-rate fallacy presented information in the form of probabilities or percentages. Mathematically, probabilities, percentages, and frequencies are equivalent representations of statistical information. Psychologically, however, they are not equivalent. Physicist Richard Feynman (1967) described the consequences of information representation for deriving different mathematical formulations of the same physical law thus: "Psychologically they are different because they are completely unequivalent when you are trying to guess new laws" (p. 53). This insight is central to the argument that problems that represent statistical information in terms of *natural frequencies* rather than probabilities, percentages, or relative frequencies are more likely to elicit correct Bayesian inferences from both laypeople and experts (Cosmides & Tooby, 1996; Gigerenzer & Hoffrage, 1995; Hoffrage, Lindsey, Hertwig, & Gigerenzer, 2000). Natural frequencies are absolute frequencies of events that have not been normalized with respect to the base rates of the hypothesis or of its absence. In natural frequencies, the mammography problem would read:

> Of 1,000 women at age 40 who participate in routine screening, 10 women have breast cancer. Nine of those 10 women with breast cancer will test positive and 99 of the 990 women without breast cancer will also test positive. How many of those who test positive actually have breast cancer?

To see how natural frequencies are related to bounded rationality, recall Simon's (1990b) view that human rational behavior arises from the interplay between the structure of task environments and organisms' computational capabilities. In the case of statistical reasoning, this means that one cannot understand people's inferences without taking external representations of statistical information, as well as cognitive algorithms for manipulating that information, into account. For most of their existence, humans and animals have made statistical inferences on the basis of information encoded sequentially through their direct experience. Natural frequencies are the result of this process. The concept of mathematical probability, in contrast, emerged only in the mid-seventeenth century (Daston, 1988). Percentages seem to have become common representations only in the aftermath of the French revolution, mainly for purposes of calculating taxes and interest; only very recently have percentages become a way to represent risk and uncertainty more generally. Based on these observations, Gigerenzer and Hoffrage (1995) argued that minds have evolved to deal with natural frequencies rather than with probabilities.[6]

Independent of evolutionary considerations, Bayesian computations are simpler to perform when the relevant information is presented in natural frequencies than in probabilities, percentages, or relative frequencies because natural frequencies do not require figuring in base rates. Compare, for instance, the computations that an algorithm for computing the posterior probability that a woman has breast cancer given a positive mammogram when the information is represented in probabilities

(shown in Equation 1) with those necessary when the same information is presented in natural frequencies:

$$p(\mathrm{H}|\mathrm{D}) = \frac{pos\ \&\ cancer}{pos\ \&\ cancer + pos\ \&\quad cancer} = \frac{9}{9 + 99} \approx .08. \qquad (2)$$

Equation 2 is Bayes' rule for natural frequencies, where *pos&cancer* is the number of women with breast cancer and a positive test and *pos&¬cancer* is the number of women without breast cancer but with a positive test. In the natural frequency representation, fewer arithmetic operations are necessary, and those required can be performed on natural numbers rather than fractions.

Probabilistic reasoning improves when statistical information is presented in terms of natural frequencies rather than probabilities. Take, for instance, Gigerenzer and Hoffrage's (1995) study of university students' ability to solve a set of 15 Bayesian reasoning problems that included many of the problems in which the base-rate fallacy had been observed (e.g., the mammography problem). Participants received the statistical information in each problem in terms of probabilities or natural frequencies. As Figure 2 shows, in each of the 15 problems, natural frequencies substantially increased the proportion of Bayesian inferences. On average, people reasoned the Bayesian way in only about 1 out of 6 cases given probabilities, whereas in 1 out of 2 cases they did so given natural frequencies. Other studies show that natural frequencies foster Bayesian reasoning among experts who make medical and forensic inferences (e.g., Hoffrage et al., 2000). Moreover, Sedlmeier and Gigerenzer (2001) designed a tutorial computer program that teaches people to translate probability information into natural frequencies (representation training) or, alternatively, to insert probabilities into Bayes' rule (rule training). Rule training resulted in the typical forgetting curve, whereas representation training resulted in robust probabilistic thinking even three months after the training.

Regardless of one's take on the evolutionary argument about natural frequencies,[7] it seems to be widely accepted that the extent to which people obey principle they fall prey to biases such as overconfidence depends on the way in which statistical information is presented.[8]

DOES STIMULUS REPRESENT-ATION MATTER IN EXPERIMENTAL ECONOMICS?

An important example of how information representation matters in economics experiments is the Allais paradox. Together with Ellsberg's paradox, it is the most prominent of the (early) violations of expected utility theory reported in the economics literature (Kreps, 1990; Mas-Colell et al., 1995). According to the independence axiom, aspects that are common to two gambles should not influence choice behavior (Savage, 1954). For any three alternatives X, Y, and Z taken from a set of options S, the independence axiom can be written (Fishburn, 1979):

$$\text{If} \quad pX + (1 - p)Z \succ pY + (1 - p)Z \quad \text{then} \quad X \succ Y \qquad (3)$$

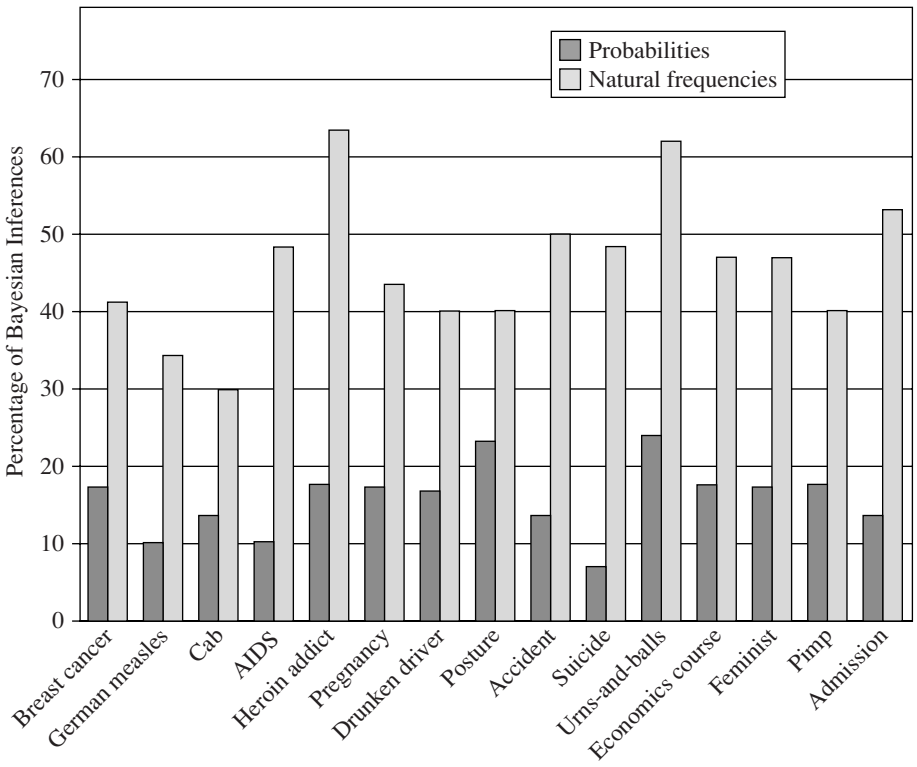

Figure 2. Across 15 Bayesian reasoning problems, statistical information was either presented in probabilities or in natural frequencies. In each problem, probabilistic reasoning improved when statistical information was communicated in natural frequencies (adapted from Gigerenzer and Hoffrage, 1995). To qualify as a Bayesian inference, the participant had to respond with the exact Bayesian estimate, and *the written protocol had to confirm that the response was derived from actual Bayesian reasoning.*

The following choice problems produce violations of the axiom:

A:	100 million	for sure
B:	500 million	$p = .10$
	100 million	$p = .89$
	0	$p = .01$

By eliminating a .89 probability to win 100 million from both gambles *A* and *B*, Allais obtained the following alternatives:

C:	100 million	$p = .11$
	0	$p = .89$
D:	500 million	$p = .10$
	0	$p = .90$

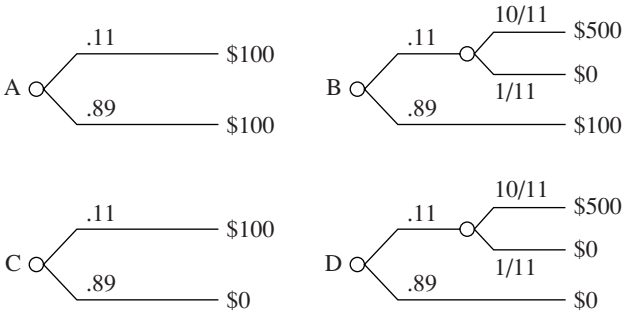

Figure 3. A graphical representation of the gambles involved in the Allais paradox that reduces the proportion of inconsistent choices (adapted from Kreps, 1990).

The majority of people choose *A* over *B* and *D* over *C* (e.g., MacCrimmon, 1968), which constitutes a violation of the axiom.

However, there is evidence that different task representations can lead to considerable reductions in the percentage of inconsistent choices. For example, when the gambles are presented to participants in the graphical form shown in Figure 3 (adapted from Kreps, 1990), then inconsistent behavior decreases sharply (see also Conlisk, 1989, for another example of the impact of task representation on the percentage of inconsistent choices). As with probability representations of Bayesian inference, the problem with the standard representation of the gambles, which coalesces probabilities and makes the payoffs more difficult to compare, is their complexity.

Uncertainty and risk are arguably the dominant theme of modern economics (e.g., Mas-Colell et al., 1995; Kreps, 1990). Probabilities are therefore an essential ingredient of solution concepts such as sequential equilibrium that are used to analyze problems of moral hazard, adverse selection, screening, and signaling that can be conceptualized as games of strategic interaction under incomplete information. Signaling games, the most prominent in this class, go back to Spence's highly influential analysis of informational transfers in hiring and related processes. The basic problem is that workers with higher abilities may not be able to signal this fact credibly to employers. Spence (1974) suggested that, to signal their type, such workers might invest in education. If it is easier for workers with higher abilities to invest in education (as is customarily assumed), then they might be able to distinguish themselves from their less able competitors.

Such signaling games typically have multiple Nash equilibria, the number of which theorists have tried to reduce by imposing various restrictions on out-of-equilibrium beliefs. This is where probabilities come in. Such refinements require the person who uses the model to use Bayes's rule so as to make the strategy profile and the belief system mutually consistent. As every graduate student in economics can attest, this is typically a computational task of a tall order. Not surprisingly, the

experimental evidence from tests of signaling games and refinements indicates that some of the subtler refinements (e.g., beyond sequential equilibrium) overtax participants (e.g., Banks, Camerer, & Porter, 1994). It is important to repeat that these models make heroic knowledge and rationality assumptions as well as assumptions about commonly known identical beliefs. Where do they come from? And what is their ecological validity?

The few experimental tests of signaling models that exist have matched participants repeatedly and observed how they learned. Note that participants in such games, whether or not they know the distribution of types of workers, have to perform belief updating that is likely to be affected by information representation. The results are a mixed bag that shows, among other things, that meaningful context both facilitates learning within a game and across related games (Cooper & Kagel, 2003). In our view, the question of how to represent information is key to the design of such learning experiments.

CONCLUSION

Its implications for human rationality aside, the cognitive illusion controversy in psychology has spawned a body of research with important implications for experimental economics. This research demonstrates that theoretical questions such as how well people are calibrated to their own knowledge and whether people update probabilities in a Bayesian way cannot be disentangled from the methodological questions of how to sample and represent experimental stimuli from the environment. To the extent that cognitive strategies and environmental structures go hand in hand, the world that is realized or represented in the laboratory codetermines how well the strategies perform and, ultimately, experimenters' conclusions.

Germane here is Vernon Smith's (2002) recent discussion of the Duhem-Quine problem in the context of experimentation in economics. The crux of the problem is that any experiment represents a test of two things: the hypotheses derived from the theory of interest and the auxiliary hypotheses necessary to implement the experiment. In psychological and economic experiments, the latter include hypotheses about measurement instruments, participant payments, and instructions. Because of the auxiliary hypotheses, any failure of the experiment to confirm the theoretical hypotheses can be explained in one of three ways: The theory is wrong; one or more of the auxiliary hypotheses are wrong; or both the theory and the auxiliary hypotheses are wrong. Thus, in Lakatos's words (quoted in Smith, 2002, p. 98): "No theory is or can be killed by an observation. Theories can always be rescued by auxiliary hypotheses."

Although experimental outcomes are thus inherently ambiguous, Smith sees no reason for despair. On the contrary, he argues, the Duhem-Quine problem is a driving force behind methodological innovation and scientific progress. Ambiguous results spark not only controversy but also the execution of new experiments designed to narrow the range of tenable interpretations. The results of these experiments, in turn,

illuminate the extent to which the behavior of interest is sensitive to methodological variation. They also suggest new research questions, thus initiating a new cycle of experiments. In Smith's (2002) words, "The bottom line is that good-enough solutions emerge to the baffling infinity of possibilities, as new measuring systems emerge, experimental tools are updated, and understanding is sharpened" (p. 104).

We share Smith's (2002) optimistic pragmatism, although, having observed the tug of war over cognitive illusions for a decade, we are not convinced that more experiments always bring more clarity. Still, the cognitive illusion controversy has yielded profound knowledge about how human reasoning, judgment, and choice are affected by stimulus representation and stimulus sampling. In experimental economics, the auxiliary hypotheses needed to perform an experiment are in themselves substantive theories of, for instance, the interaction between cognitive processes and environmental structures. It is here where psychology has something to contribute to experimental economics.

NOTES

* Ralph Hertwig, University of Basel, Basel, Switzerland; Andreas Ortmann, Charles University and Academy of Sciences of the Czech Republic, Prague. We would like to thank Dirk Engelmann and Pavlo Blavatskyy for many constructive comments and Valerie Chase for valuable editorial input.

Correspondence should be addressed to Ralph Hertwig, Institute for Psychology, University of Basel, Missionsstrasse 60/62, 4055 Basel, Switzerland. Electronic mail may be sent to ralph.hertwig@unibas.ch.

[1] Interpreting the principal components of SEU theory in "as-if" terms, as is often proposed, skirts the question of what cognitive processes lead people to their decisions.

[2] Gigerenzer et al. (1991) defined the ecological validity of a cue as the proportion of correct inferences that a person using only that cue would make in the subset of paired comparisons where the cue discriminates between alternatives (e.g., where one city is a capital and the other is not).

[3] An extensive literature (e.g., Zacks & Hasher, 2002) indeed suggests that people are well calibrated to environmental frequencies.

[4] Creating an exhaustive set of paired comparisons of the 20 largest Australian cities results in 190 comparisons. In 171 of the 190 pairs, the capital cue does not discriminate (because neither of the cities is a capital). In such cases, let us assume that the person guesses and estimates her confidence to be 50%. In 19 cases, the capital cue discriminates. Let us assume that the person estimates her confidence to be the cue's ecological validity, which is 75%. Averaged across all cases, her mean confidence should therefore be 53%, as should be the percentage of comparisons to which she provides the correct answer.

[5] Since then, a small but increasing number of economics studies has addressed the issue of representative design. An encouraging development in this vein is field experiments that use nontraditional subject pools, real-life decision situations, and real-life goods and services (Harrison & List, in press).

[6] This argument is consistent with developmental studies indicating the primacy of reasoning about discrete numbers and counts over fractions and with studies of adult humans and animals showing that they can monitor frequency information in their natural environment in fairly accurate and automatic ways (see Gigerenzer & Hoffrage, 1995).

[7] For discussion of these issues, see, for instance, Johnson-Laird, Legrenzi, Girotto, Legrenzi, and Caverni (1999), and Hoffrage, Gigerenzer, Kraus, and Martignon (2002).

[8] The possible reasons for why representation matters, however, are controversially discussed (e.g., Tversky & Kahneman, 1983; Hertwig & Gigerenzer, 1999; Mellers, Hertwig, & Kahneman, 2001).

REFERENCES

Banks, J., Camerer, C. F., & Porter, D. (1994). An experimental analysis of Nash refinements in signaling games. *Games and Economic Behavior*, 6, 1–31.

Barber, B. M., & Odean T. (2001). Boys will we boys: Gender, overconfidence, and common stock investment. *Quarterly Journal of Economics*, *116*, 261–292.

Binmore, K. (1994). *Game theory and the social contract: Playing fair* (Vol. 1). Cambridge, MA: MIT Press.

Binmore, K. (1997). *Game theory and the social contract: Just playing* (Vol. 2). Cambridge, MA: MIT Press.

Brenner, L. A., Koehler, D. J., Liberman, V., & Tversky, A. (1996). Overconfidence in probability and frequency judgments: A critical examination. *Organizational Behavior and Human Decision Processes*, *65*, 212–219.

Brunswik, E. (1944). Distal focussing of perception: Size constancy in a representative sample of situations. *Psychological Monographs*, *56*, 1–49.

Brunswik, E. (1955). Representative design and probabilistic theory in a functional psychology. *Psychological Review*, *62*, 193–217.

Bukszar, E., & Connolly, T. (1988). Hindsight bias and strategic choice: Some problems in learning from experience. *Academy of Management Journal*, *31*, 628–641.

Camerer, C. F. (2003). *Behavioral game theory. Experiments in strategic interaction*. Princeton, NJ: Princeton University Press.

Camerer, C. F. (1995). Individual decision making. In J. H. Kagel & A. E. Roth (eds.). *Handbook of experimental economics* (pp. 587–703). Princeton, NJ: Princeton University Press.

Colman, A. M. (2003). Cooperation, psychological game theory, and the limitations of rationality in social interaction. *Behavioral and Brain Sciences*, *26*, 139–198.

Colman, A. M. (1995). *Game theory and its applications in the social and biological sciences* (2nd ed.). Amsterdam: Butterworth-Heinemann.

Conlisk, J. (1989). Three variants on the Allais paradox. *American Economic Review*, *79*, 392–407.

Cooper, D., & Kagel, J. (2003). The impact of meaningful context on strategic play in signaling games. *Journal of Economic Behavior and Organization*, *50*, 331–337.

Cosmides, L., & Tooby, J. (1996). Are humans good intuitive statisticians after all? Rethinking some conclusions from the literature on judgment under uncertainty. *Cognition*, *58*, 1–73.

Cox, J. C. (2004). How to identify trust and reciprocity. *Games and Economic Behavior*, *46*, 260–281.

Dhami, M. K., Hertwig, R., & Hoffrage, U. (in press). The role of representative design in an ecological approach to cognition. *Psychological Bulletin*.

Daston, L. (1988). *Classical probability in the Enlightenment*. Princeton, NJ: Princeton University Press.

Dyer, D., & Kagel, J. H. (1996). Bidding in common value auctions: How the commercial construction industry corrects for the winner's curse. *Management Science*, *42*, 1463–1475.

Eddy, D. M. (1982). Probabilistic reasoning in clinical medicine: Problems and opportunities. In D. Kahneman, P. Slovic, & A. Tversky (eds.), *Judgment under uncertainty: Heuristics and biases* (pp. 249–267). Cambridge, UK: Cambridge University Press.

Edwards, W., & von Winterfeldt, D. (1986). On cognitive illusions and their implications. In H. R. Arkes & K. R. Hammond (eds.), *Judgment and decision making* (pp. 642–679). Cambridge, UK: Cambridge University Press.

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

Fehr, E., & Gächter, S. (2004). Fairness and retaliation: The economics of reciprocity. In C. F. Camerer, G. Loewenstein, & M. Rabin (eds.), *Advances in behavioral economics* (pp. 510–532). Princeton, NJ: Princeton University Press.

Feyman, R. (1967). *The character of physical law*. Cambridge, MA: MIT Press.

Fishburn, P. C. (1979). On the nature of expected utility. In M. Allais & O. Hagen (eds.), *Expected utility hypotheses and the Allais paradox* (pp. 243–257). Dordrecht, Netherlands: Reidel.

Gigerenzer, G. (1991). How to make cognitive illusions disappear: Beyond heuristics and biases. In W. Stroebe & M. Hewstone (eds.), *European Review of Social Psychology* (Vol. 2, pp. 83–115). Chichester, UK: Wiley.

Gigerenzer, G. (1996). On narrow norms and vague heuristics: A reply to Kahneman and Tversky. *Psychological Review*, *103*, 592–596.

Gigerenzer, G., & Hoffrage, U. (1995). How to improve Bayesian reasoning without instruction: Frequency formats. *Psychological Review*, *102*, 684–704.

Gigerenzer, G., Hoffrage, U., & Kleinbölting, H. (1991). Probabilistic mental models: A Brunswikian theory of confidence. *Psychological Review*, *98*, 506–528.

Gigerenzer, G., Todd, P. M., & the ABC Research Group. (1999). *Simple heuristics that make us smart*. Oxford: Oxford University Press.

Gilovich, T., Griffin, D., & Kahneman, D. (eds.). (2002). *Heuristics and biases: The psychology of intuitive judgment*. Cambridge, England: Cambridge University Press.

Griffin, D., & Tversky, A. (1992). The weighing of evidence and the determinants of confidence. *Cognitive Psychology*, *24*, 411–435.

Harrison, G. W., & List, J. (in press). *Field experiments.*

Hertwig, R., & Gigerenzer, G. (1999). The "conjunction fallacy" revisited: How intelligent inferences look like reasoning errors. *Journal of Behavioral Decision Making*, *12*, 275–305.

Hertwig, R., & Ortmann, A. (2001). Experimental practices in economics: A challenge for psychologists? *Behavioral and Brain Sciences*, *24*, 383–451.

Hirshleifer, D. (2001). Investor psychology and asset pricing. *The Journal of Finance*, *61*, 1533–1597.

Hoffman, E., McCabe, K., & Smith, V. L. (1996). On expectations and the monetary stakes in ultimatum games. *International Journal of Game Theory*, *25*, 289–301.

Hoffrage, U., Gigerenzer, G., Kraus, S., & Martignon, L. (2002). Representation facilitates reasoning: What natural frequencies are and what they are not. *Cognition*, *84*, 343–352.

Hoffrage, U., Hertwig, R., & Gigerenzer, G. (2000). Hindsight bias: A by-product of knowledge updating? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *26*, 566–581.

Hoffrage, U., Lindsey, S., Hertwig, R., & Gigerenzer, G. (2000). Communicating statistical information. *Science*, *290*, 2261–2262.

Johnson-Laird, P. N., Legrenzi, P., Girotto, V., Legrenzi, M. S., & Caverni, J.-P. (1999). Naive probability: A mental model theory of extensional reasoning. *Psychological Review*, *106*, 62–88.

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*, 384–396.

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

Kahneman, D., & Tversky, A. (1996). On the reality of cognitive illusions: A reply to Gigerenzer's critique. *Psychological Review*, *103*, 582–591.

Kahneman, D., & Tversky, A. (1972). Subjective probability: A judgment of representativeness. *Cognitive Psychology*, *3*, 430–454.

Koehler, J. J. (1996). The base rate fallacy reconsidered: Descriptive, normative, and methodological challenges. *Behavioral and Brain Sciences*, *19*, 1–53.

Kreps, D. M. (1990). *A course in microeconomic theory*. Princeton, NJ: Princeton University Press.

Krueger, J. I., & Funder, I. (2004). Towards a balanced social psychology: Causes, consequences and cures for the problem-seeking approach to social behavior and cognition. *Behavioral and Brain Sciences*.

Ledyard, J. (1995). Public good experiments. In J. H. Kagel & A. E. Roth (eds.), *Handbook of experimental economics* (pp. 111–194). Princeton, NJ: Princeton University Press.

Marr, D. (1982). *Vision*. New York: Freeman and Company.

Mas-Colell, A., Whinston, M. D., & Green, J. R. (1995). *Microeconomic theory*. New York: Oxford University Press.

MacCrimmon, K. R. (1968). Descriptive and normative implications of the decision-theory postulate. In K. H. Borch & J. Mossin (eds.), *Risk and Uncertainty* (pp. 3–23). New York: St. Martin's Press.

Mellers, B., Hertwig, R., & Kahneman, D. (2001). Do frequency representations eliminate conjunction effects? An exercise in adversarial collaboration. *Psychological Science*, *12*, 269–275.
Odean, T. (1999). Do investors trade too much? *American Economic Review*, *89*, 1279–1298.
Ortmann, A., & Gigerenzer, G. (1997). Reasoning in economics and psychology: Why social context matters. *Journal of Institutional and Theoretical Economics*, *153*, 700–710.
Ortmann, A., & Hertwig, R. (2002). The costs of deception: Evidence from psychology. *Experimental Economics*, *5*, 111–131.
Payne, J. W., Bettman, J. R., & Johnson, E. J. (1993). *The adaptive decision maker*. New York: Cambridge University Press.
Plott, C. R. (1987). Dimensions of parallelism. Some policy applications of experimental economics. In A. E. Roth (ed.), *Laboratory experimentation in economics: Six points of view* (pp. 193–219). New York: Cambridge University Press.
Savage, L. J. (1954). *The foundations of statistics* (2nd ed.). New York: Dover.
Schoemaker, P. J. H. (1982). The expected utility model: Its variants, purposes, evidence and limitations. *Journal of Economic Literature*, *20*, 529–563.
Sedlmeier, P., & Gigerenzer, G. (2001). Teaching Bayesian reasoning in less than two hours. *Journal of Experimental Psychology: General*, *130*, 380–400.
Shiller, R. J. (2000). *Irrational exuberance*. Princeton, NJ: Princeton University Press.
Simon, H. A. (1956). Rational choice and the structure of environments. *Psychological Review*, *63*, 129–138.
Simon, H. A. (1990a). Alternative visions of rationality. In P. K. Moser (ed.), *Rationality in action: Contemporary approaches* (pp. 189–204). New York: Cambridge University Press.
Simon, H. A. (1990b). Invariants of human behavior. *Annual Review of Psychology*, *41*, 1–19.
Smith, A. (1759/1982). *The theory of moral sentiments*. Indianapolis, IN: Liberty Classics.
Smith, V. L. (2002). Method in experiment: Rhetoric and reality. *Experimental Economics*, *5*, 91–110.
Spence, M. (1974). *Market signaling: Informational transfer in hiring and related processes*. Cambridge, MA: Harvard University Press.
Sunstein, C. R. (2000). *Behavioral law and economics*. Cambridge, UK: Cambridge University Press.
Thaler, R. (1980). Toward a positive theory of consumer choice. *Journal of Economic Behavior and Organization*, *1*, 39–60.
Thaler, R. H. (1993). *Advances in behavioral finance*. New York: Russell Sage Foundation.
Tversky, A., & Kahneman, D. (1983). Extensional versus intuitive reasoning: The conjunction fallacy in probability judgment. *Psychological Review*, *90*, 293–315.
Winman, A. (1997). The importance of item selection in "knew-it-all-long" studies of general knowledge. *Scandinavian Journal of Psychology*, *38*, 63–72.
Zacks, R. T., & Hasher, L. (2002). Frequency processing: A twenty-fie year perspective. In Sedlmeier, P. & Betsch, T. (eds.), *Etc. Frequency processing and cognition* (pp. 21–36). Oxford: Oxford University Press.