THE ECONOMIC JOURNAL

Review 2. Rationality and its Discontents*

Ken Binmore needs no introduction to readers of this Journal. In *Rational Decisions*, this mathematician turned economist turned philosopher combines brief introductions to Bayesian decision theory and game theory with a far-reaching and synthetic assessment of the limits of Bayesian decision theory and offer new directions in extending decision theory to situations where the traditional approach does not apply. Binmore's arguments are generally sketchy and impressionistic, seeking to convey ideas rather than rigorously justifying them. His references to the literature are broad and deep, hence likely to keep the interested reader busy for a good period of time. This book is thus not for the economist who is blissfully contented with the usual version of the theory learned in graduate school (Varian, 1992; Mas-Colell *et al.*, 1995). Nor is this book for the beginner, who would do better reading Savage (1954) directly, or a more recent didactic exposition (Kreps, 1988; Gintis, 2009*b*). For instance, Binmore follows Anscombe and Aumann (1963) in deriving the expected utility theorem axiomatically, thus side-stepping the behaviourally important issue of how consistent preferences lead to probability distributions when the choices involved do not include lotteries with objective probabilities. More generally, Binmore passes seamlessly from elementary exposition to deep and complex issues that are bound to leave the novice in the dust.

1. Small Worlds, Large Worlds

Binmore begins by defining the enemy as *Bayesianism*, which he describes as the doctrine that 'Bayesian decision theory is always rational'. The doctrine entails, for example, that David Hume was wrong to argue that scientific induction cannot be justified on rational grounds. Lindley (1988) is one of many scholars who are convinced that Bayesian inference has been shown to be the only coherent form of inference' (p. 1). By contrast, Binmore holds that there are broad areas that, following Savage (1954), he calls 'Large Worlds', over which rational choice does not conform to Bayesian principles. 'I hope to distinguish Bayesian decision theory,' he asserts in the opening chapter, 'from Bayesianism. We can hold on to the virtues of the former without falling prey to the excesses of the latter.'

Binmore's claim of course requires that we define 'rational' in a manner independent from its representation in standard decision theory, which is the theory taught in the textbooks and shared by most economists. I have always been comfortable with identifying rationality with the Savage axioms, which may be described in shorthand as 'preference consistency over lotteries with subjective probabilities'. Binmore reminds us that Savage himself had a much broader concept of rationality, one that allowed him to assert that in the 'Small World' described by his axioms, rationality equals consistent preferences over lotteries but, in other 'Large Worlds', rationality must be captured in some other (unspecified) manner.

Whereas most decision-theorists simply ignore Savage's ruminations concerning Small and Large Worlds, Binmore elevates the distinction to the central theme of

* I thank the European Science Foundation and the Hungarian Scientific Research Fund for financial support.

Rational Decisions. Like Savage, Binmore considers the idea that one can use the Bayesian model in any setting whatever as implausible or, in Binmore's more decisive terms, 'utterly ridiculous' and 'preposterous' (p. 117). Like Savage before him, Binmore does not explicitly specify the nature of Small Worlds, although he does comment that 'the Worlds of macroeconomics and high finance most certainly don't fall into this category' (p. 2). Neither Savage nor Binmore defines rationality, so their quest for a Large World is ineluctably impressionistic. Binmore writes, 'No formal definition of rationality will be offered.... To insist on an a priori definition would be to make the Pythagorean mistake of prematurely closing our minds to possible future inventions' (p. 2). I submit, however, that a tentative definition that fits the bill until some 'future invention' renders it obsolete is: a rational decision is one that assesses the implications of alternative choices as accurately as possible given the evidence available and the cost of obtaining new evidence, and chooses a course of action that is best fit to achieve the decision-maker's goals.

Binmore's idea of a Small World is one in which we know that events are governed by one of a fixed number of possible models. Considering an event E, we start out without a subjective prior $P(E)$ but rather, for every possible piece of information F, we use our gut reaction to form a conditional probability $P(E \mid F)$. If the resulting array of conditional probabilities cannot be deduced from a subjective prior over events, we 'massage' our conditional probabilities until they can be so deduced. The resulting probability distribution $P(E)$ will satisfy the Savage axioms, Binmore asserts, and in addition $P(F) > 0$ for any possible new information, so Bayesian updating is always possible.

In a Small World, then, there is no real learning in the face of new information. Rather, new information merely leads us to favour one of our pre-existing models over another. In Binmore's words, 'After the massaging is over, Pandora would then be invulnerable to surprise, because she would have already taken account of the impact that any future information might have on the internal model that she uses in determining her beliefs.... Bayes' rule is therefore reduced to nothing more than a bookkeeping tool that saves Pandora from having to remember all her massaged posterior probabilities' (p. 132).

The attractiveness of Binmore's depiction of the Small World in which Bayesian decision theory is valid lies in its conformance with the Savage axioms, especially the independence of irrelevant alternatives, which is key to the establishment of conditional probabilities and the assumption that preference consistency is assumed only on 'non-null' events, which is equivalent to the assumption that all events have positive probability. In this situation, as Binmore asserts, nothing essentially new or unexpected can happen and, should true novelty arise, the poor Bayesian would have no means of adjusting.

According to this theory, which certainly could be tested in the psychologist's laboratory, a 'Small World' is one in which

(*a*) all events that occur with positive probability are non-null and
(*b*) decision-makers believe that (*a*) is the case.

Note that it is not at all a tautology to say that a null event could occur with positive probability. Indeed, my subjective prior for event E may be zero but E could still occur and I might even entertain that my subjective prior is incorrect. When I make plans to go to the hardware store, I attach probability zero to the event that I there might encounter something looking like a rotten tree-stump eating green garden hose, and when I get a take-away from the local shop, I attach probability zero to, and hence do not make provisions for, the server slipping her hand in my shirt and blowing in my ear while handing me my package. Nevertheless, the objective probability of both events is surely strictly positive. The important point is that I have no pre-given way of updating should either of these events occur.

I am deeply impressed with Binmore's elevating the Small vs. Large World conceptual dichotomy to the centrepiece in evaluating rational decision theory. We live our daily lives in a Large World. Many times in my life I have encountered events that I could never have conceived of occurring and that did not fit into any of the alternative world-views that were represented in my mind as among the possible. The fact is that human beings update following zero probability events and, indeed, do so very ably, while Bayesian decision theory gives no hint as to how that might be done, or how decisions are made when people realise that they are in a situation for which they have no appropriate model.

2. Deliberative Choice: The Psychologist's Large World Model

Most psychologists working in Small World contexts accept the rational actor model as appropriate (Luce, 2000; Baron, 2007). For instance, Newell *et al.* (2007) assert, 'We view judgment and decision making as often exquisitely subtle and well-tuned to the world, especially in situations where we have the opportunity to respond repeatedly under similar conditions where we can learn from feedback' (p. 2). Yet Small World psychologists recognise that there is no obvious way to extend the model to the more complex, Large World, situations they study.

Psychologists who study complex decision-making, by contrast, appear in recent years to have rejected the rational actor model altogether. Indeed, many have interpreted the brilliant work of Daniel Kahneman, Adam Tversky and their colleagues as a refutation thereof. 'People are not logical', the saying goes, 'but rather are psychological'. I recently went through several of the leading introductory graduate textbooks in cognitive psychology, and found a striking uniformity: decision-making is not dealt with until very late in the book (doubtless the teacher rarely gets to this material), and the message is always that humans are poor decision-makers, they cannot apply Bayes's rule and the rational actor model is simply a pipe dream of armchair economists. Thus, for most psychologists, there simply is no Small World in the real world, so we must start from scratch to understand human choice behaviour.

The problem with this view is not only that economists have done quite strikingly decent analysis of real-world problems using the Bayesian rational actor but also the psychological literature on decision-making, while rich, multifaceted and having developed neural net theory and neuroscientific data on brain functioning (Kahneman *et al.*, 1982; Baron, 2007; Oaksford and Chater, 2007; Hinton and Sejnowski, 1999; Newell *et al.*, 2007; Juslin and Montgomery, 1999; Bush and Mosteller, 1955; Gigerenzer and Todd, 1999; Betch and Haberstroh, 2005; Koehler and Harvey, 2004), has not even come close to developing a unitary model of the psychology of judgment and decision-making.

The sorts of decision-making studied by psychologists include the formation of long-term goals, which are evaluated according to the value if attained, the range of probable costs and the probability of goal attainment. All three dimensions of goal formation have inherent uncertainties, so among the strategies of goal choice is the formation of subgoals with the aim of reducing these uncertainties. The most complex of human decisions tend to involve goals that arise infrequently in the course of a life, such as choosing a career, whether to marry and to whom, how many children to have, and how to deal with a health threat, where the scope for learning from mistakes is narrow. Psychologists also study how people make decisions based on noisy single or multi-dimensional data under conditions of trial-and-error learning.

The difficulty in modelling such deliberative choice in Large Worlds is exacerbated by the fact that, because of the complexity of such decisions, much human decision making has a distinctly group dynamic, in which some individuals experiment and other imitate the more successful of the experimenters (Bandura, 1977). This dynamic cannot be successfully modelled on the individual level. I return to this theme below.

If we recognise the power of the Bayesian model in the Small World context and admit that we need additional concepts to deal with Large Worlds, there is then no conceptual divide between the psychological approach to decision-making and the economic approach. While in some important areas, human decision-makers appear to violate the consistency condition for rational choice, in virtually all such cases, as I suggest in Gintis (2009*a*, ch. 12) consistency can be restored by assuming that the current state of the agent is an argument of the preference structure. Another possible challenge to preference consistency is preference reversal in the choice of lotteries. Lichtenstein and Slovic (1971) were the first to find that in many cases, individuals who prefer lottery *A* to lottery *B* are nevertheless willing to take less money for *A* than for *B*. Reporting this to economists several years later, Grether and Plott (1979) asserted, ‘A body of data and theory has been developed… [that] are simply inconsistent with preference theory… ’(p. 623). These preference reversals were explained several years later by Tversky *et al.* (1990) as a bias toward the higher probability of winning in a lottery choice and toward the higher maximum amount of winnings in monetary valuation. However, the phenomenon has been documented only when the lottery pairs *A* and *B* are so close in expected value that one needs a calculator (or a quick mind) to determine which would be preferred by an expected value maximiser. For instance, in Grether and Plott (1979) the average difference between expected values of comparison pairs was 2.51% (calculated from Table 2, p. 629). The corresponding figure for Tversky *et al.* (1990) was 13.01%. When the choices are so close to indifference, it is not surprising that inappropriate cues are relied upon to determine choice, as would be suggested by the heuristics and biases model (Kahneman *et al.*, 1982) favoured by behavioural economists and psychologists.

The expected utility model is close to the concerns of psychologists because it deals with uncertainty in a fundamental way, and applying Bayes’ rule certainly may involve complex deliberations. The Ellsberg paradox is an especially clear example of the failure of the probability reasoning behind the expected utility model. Nevertheless the model has a considerable body of empirical support, so the basic modelling issue is to be able to say clearly when the expected utility theorem is likely to be violated and to supply an alternative model outside this range (Newell *et al.*, 2007; Oaksford and Chater, 2007).

3. Beyond Bayesian Updating: The Role of Imitation

Binmore's proposal to develop a version of the rational actor model that applies to Large Worlds begins with the suggestion by Luce and Raiffa (1957, ch. 13) that in Large Worlds decisions are made in complete ignorance, and are based on the principle of insufficient reason and the maximin criterion. Binmore presents the complete ignorance axiomatic system of Milnor (1954), followed by his own model of choice when subjective probabilities are closed intervals rather than points. While this material is of interest, it certainly is not the case that Large Worlds mean either complete ignorance or simple uncertainty concerning subjective priors. Indeed, one of Binmore's first references is to an approach to Large Worlds, based on the phenomenon of *imitation*, by economists Gilboa and Schmeidler (2001), of which I was unfamiliar. The book is a major step forward, although I am afraid its law-schoolish title is likely to lead many economists to ignore it,

There have been several important theoretical and empirical contributions to the study of imitation by economists, including the seminal studies of Conlisk (1988) and Bikhchandani *et al.* (1992), the learning models of Bannerjee (1992), Ellison and Fudenberg (1995), Vega-Redondo (1997) and Schlag (1998,1999), as well as experimental work by Offerman and Schotter (2008), Abbink and Brandts (2008) and others. However, before Gilboa and Schmeidler imitation was not considered a fundamental part of rational decision theory.

Offerman and Schotter (2008), for instance, open their paper with the sentence 'imitation may be called the poor man's rationality'. By contrast, animal behaviourists have shown that the capacity to learn by imitation is extremely rare in the animal world (Tomasello, 1999; Meltzoff and Prinz, 2002) and generally requires sophisticated understanding of intersubjective epistemology when more than one sensory modality is involved. Moreover, imitiation is the driving mechanism in most models of the dynamics of cultural evolution (Gintis, 2009*b*, ch. 12).

Gilboa and Schmeidler start with decision problem p and they assume the decision-maker has a repertoire M of 'cases' $[q, a, r]$, where q is another decision problem, a is the action taken in the case of decision problem q, and r is the result of the action. The decision-maker then forms a subjective 'similarity' $s[p, q] \in \mathbf{R}^+$ of the current problem p with the problem q, and set the value

$$U[a] = \sum_{[q,a,r] \in M} s[p, q]u[r],$$

where $u[r]$ is the utility of the result. Finally, the decision-maker chooses the action a that maximises $U[a]$. Note that the repertoire M can include both the decision-maker's own past experience and the experience of others with whom the decision-maker is sufficiently familiar that a plausible similarity rating can be found.

Gilboa and Schmeidler do not stress the interpersonal aspect of their theory, but it appears to me to represent the most critical way their approach goes beyond the standard Savage model of rational choice. It should be clear from a number of studies of human behaviour that a central weakness in the Bayesian decision model lies in its failure to use the choice experience of others in updating one's own knowledge base in Large Worlds. In general, the social dimension in Bayesian decision-making should extend to the assessment of utilities as well as probabilities. This is easily incorporated into Gilboa and Schmeidler's case-based decision theory. Gilboa and Schmeidler stress that their approach is not contradictory to Bayesian theory. Indeed, if we interpret 'similarity' as the 'probability' that the act will lead to the result *r*, then case-based reduces to Bayesian decision theory, where the choice set is limited to what the decision-maker actually knows.

One of the more intriguing possibilities is that the repeated application of case-based decision-making, under appropriate conditions, might lead to standard Bayesian choices in the long run. This is especially important because, although humans are excellent Bayesian decision-makers on the level of language acquisition, word recognition, and the like (see below), they are notoriously poor formal decision-makers, as has been repeatedly shown by Daniel Kahneman, Amos Tversky and their colleagues. Because most decisions in real life depend on assessing how others have fared making 'similar' decisions rather than on a purely subjective expected utility maximisation, case-based reasoning may lead to quasi-Bayesian outcomes in the long run.

4. Rational Decisions, Imitation and Large World Macroeconomics

The world of macroeconomics 'most certainly' does not fall into the Small World category, asserts Binmore. The standard Walrasian and rational expectations models of the macroeconomy, however, most certainly *do* fall into the Small World category. In the Walrasian model, for instance, prices are public knowledge given by the auctioneer, firms know their production functions, consumers know their utility function and budget constraints and nothing else is needed to determine production and consumption decisions. Could this Small World framework be why there is no plausible Walrasian macrodynamics?

While the equilibrium properties of the Walrasian model have been well known since Arrow and Debreu (1954), progress in understanding its dynamical properties has been meagre. In Walras's original description of general equilibrium (Walras 1954 [1874]), market clearing was effected by a central authority. This authority, which has come to be known as the 'auctioneer', remains today because no one has succeeded in producing a plausible decentralised dynamic model of producers and consumers engaged in market interaction in which prices and quantities move towards market-clearing levels. Only under implausible assumptions can the continuous 'auctioneer' dynamic be shown to be stable (Fisher, 1983), and in a discrete model, even these assumptions (gross substitutability, for instance) do not preclude instability and chaos in price movements (Saari, 1985; Bala and Majumdar, 1992).

Suppose we move to a Large World in which there is no auctioneer, so expected future prices are private information, consumers form price expectations through search in each period, producers adjust pricing and production parameters by imitating more successful firms, and individual workers' formulate wage offers by imitating other, more successful workers. In Gintis (2007), I presented an agent-based model of such an economy and found that the resulting dynamic had a globally stable stationary state using plausible parameters for economies that are unstable in the traditional tâtonnment process.

Two characteristics of this model are relevant for rational decision theory. First, moving from a Small World with complete information and straightforward profit and utility maximisation to a Large World in which rational decision-making takes the form of search and imitation, we move from instability to stability. Second, even without any aggregate shocks to the system, there is considerable short-term volatility in all markets, due to the fact that the imitation process gives rise to correlated distributions of individual behaviour and, hence, of 'fat tails' that lead to periodic significant excursions from equilibrium. It need hardly be mentioned that such excursions are well-known aspects of competitive markets that cannot be explained either in the Walrasian Small World (where prices are chaotic) or the rational expectations Small World (where markets always clear).

Figure 1, taken from Gintis (2007), shows the time series of the standard deviation of prices in an agent-based model of a ten-sector Walrasian macroeconomy, where quantities are normalised so that equilibrium prices are all unity. Note that there is considerable short-term price volatility but relative prices closely approximate their equilibrium values in the long run. The reason for the volatility is the same as the reason there is long-run (approximate) equilibrium: we are in a Large World where rational decisions involve search and imitation.

5. The Logical Impossibility of Large World Bayesianism

Binmore offers an ingenious second attack on the feasibility of Large World Bayesianism. He suggests in Chapter 8 that Bayesian decision is based on an epistemology that becomes self-contradictory in Large Worlds. Binmore identifies Bayesian epistemology with standard interactive epistemology in the modal logic of knowledge

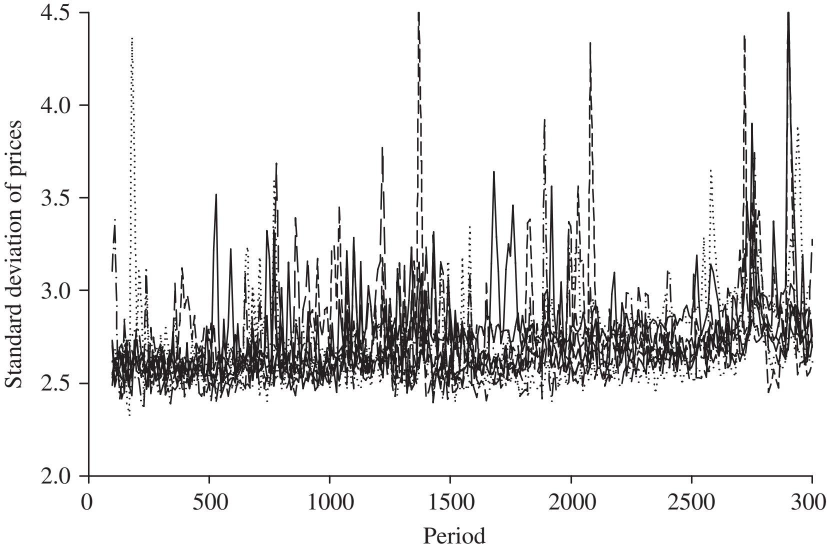

Fig. 1. *Deviation of Sectoral Prices from Equilibrium Values*

Notes. Prices are volatile, but the standard deviation of prices is rarely more than 10% of average prices and the mean standard deviation of prices is 5.8% of averages prices. Taken from Figure 5 in Gintis (2007).

(Kripke, 1963; Aumann, 1999; Gintis, 2009*a*) and argues that the completeness assumption $k(\Omega) = \Omega$ and the consistency assumption $k(E) \supseteq E$, where Ω is the knowledge universe $E \supseteq \Omega$ is any event and $k(\cdot)$ is the knowledge operator, are in fact mutually contradictory. He does this by constructing a Turing machine using Gödel numbering that defines an event E such that a state $\omega \in E$ if and only if $\omega \in k(\neg E)$.

Binmore's argument, however, suffers from ambiguity, as the precise logical system he is using is not defined. Certainly standard interactive epistemology, predicated upon the propositional calculus, is sound and consistent. A system must include axioms as strong as those of the Peano axioms for arithmetic, minus the axiom of mathematical induction, in order to sustain Gödel numbering. We have to thank the philosopher Binmore for bringing, perhaps for the first time, a rigorous logical analysis of knowledge and self-reference to the general economic public. I add the following material, which may help the reader to make sense of Binmore's argument.

Binmore cites logicians Kaplan and Montague(1960) as having proved 'a similar result'. Indeed this famous paper deals with one old paradox, The Hangman, or the Surprise Examination, and one totally new one, known as the Knower's Paradox, that spawned a large literature and is still going strong. The first will be more accessible to economists and is related to the common knowledge of rationality assumption in game theory (Gintis, 2009*a*, ch. 9), while the second is more closely related to the Liar's Paradox (Tarski, 1956[1933]; Kripke, 1975). I do not believe, however, that either paradox is related to the issue of Small vs. Large Worlds in any obvious way.

For the Surprise Exam, consider a class of game theory students taking a thirteen week logic course that meets each week from Monday to Friday. The school subscribes to a learning theory that stresses surprise exams and dismisses any instructor who fails to warn the students on the first day of class that there will be a surprise exam on some subsequent day, or who fails to give at least one surprise exam during the course of the term. Moreover, a student who can prove to the Dean's satisfaction that no surprise exam was announced on the first day, or was not given on some subsequent day, would be given a valuable reward.

The professor duly announces the surprise exam on the first day of class. One student thought to himself, 'The exam cannot be given on the last day of the semester because then it would not be a surprise'. He then noted that a similar argument shows that the exam could not be given on the next-to-last day. Reasoning similarly, he concludes that a surprise exam cannot be given. He then excitedly explains this reasoning to all of the other students, who of course agree with the ironclad reasoning. On Wednesday of the eighth week, the professor gives an exam, and all the students are surprised.

For an overview of the many proposed solutions to the problem by philosophers and logicians (I have read at least twenty papers on the subject and there are some that I have not read); see Margalit and Bar-Hillel (1983) and Chow (1998) for insightful reviews. Interpretations vary widely, and there is no single accepted solution. There are a number of cogent analyses using standard logic and modal logic to show that the professor's statement is impermissively self-referential or self-contradictory and because a false statement can validly imply anything, there is no paradox in the professor's prediction being correct.

I will follow Binkley (1968) in presenting a solution that is both elegant and close to Binmore's epistemic concerns. Let us assume there are only two days, Monday and

Tuesday. The full term argument is similar, but (much) longer. We take the case of a single student with knowledge operator k. We assume for any knowledge operator that

A1 $kf \Longrightarrow \neg k \neg f$
A2 $kf \,\&\, k(f \Longrightarrow g) \Longrightarrow kg$
A3 $kf \Longrightarrow kkf$.

Note that A1 is weaker than the usual assumption $kf \Longrightarrow f$; i.e., 'what is known is possible' is weaker than 'what is known is true'. We also assume the student knows all tautologies of the propositional calculus and all axioms.

Let $k_m f$ mean 'the student knows f on Monday' and let $k_t f$ mean 'the student knows f on Tuesday'. Let E_m be the event 'the exam is given on Monday', and let E_t be the event 'the exam is given on Tuesday'. We assume

A4 $\neg E_m \Longrightarrow k_t \neg E_m$
A5 $k_m f \Longrightarrow k_m k_t f$.

A4 says that if the exam is not given on Monday, then on Tuesday the student knows this fact, and A5 says that if the student knows something on Monday, he knows on Monday that he will continue to know it on Tuesday. The professor's assertion can be written as

$$E = (\neg E_m \Longleftrightarrow E_t) \,\&\, (E_m \Longrightarrow \neg k_m E_m) \,\&\, (E_t \Longrightarrow \neg k_t E_t). \tag{1}$$

Let us assume $k_m E$. From A4 we have

$$k_m(\neg E_m \Longrightarrow k_t \neg E_m). \tag{2}$$

From $k_m E$ we have $k_m(E_t \Longrightarrow \neg E_m)$, which with (2) gives

$$k_m(E_t \Longrightarrow k_t \neg E_m). \tag{3}$$

Now from $k_m E$ and A5, we have $k_m k_t(\neg E_m \Longrightarrow E_t)$, so

$$k_m(k_t \neg E_m \Longrightarrow k_t E_t). \tag{4}$$

From (3) and (4), we have

$$k_m(E_t \Longrightarrow k_t E_t). \tag{5}$$

Now $k_m E$ implies $k_m(E_t \Longrightarrow k_t \neg E_t)$, which, together with (5), implies $k_m(\neg E_t)$ and, hence, $k_m E_m$. This, together with $k_m E$ gives

$$k_m \neg k_m E_m. \tag{6}$$

However, $k_m E_m$ and A3 imply $k_m k_m E_m$, so by A1, we have $\neg k_m \neg k_m E_m$, which contradicts (6). Therefore the original assumption $k_m E$ is false. E is thus true but it is inadmissible to assume therefore that the student knows that E is true.

It may seem peculiar that something can be true but it is impossible for an individual to know that this is the case. However, there are well-known examples of this, including the so-called Moore paradox, an example of which is 'It is raining outside but I don't know it' (Green and Williams, 2007).

This example of epistemic blindspot (Sorensen, 1988) should be a cautionary tale for the assumption of the common knowledge of rationality (CKR), to which I turn in the next Section. Economists tend to think of CKR as just a strong form of rationality. In fact, we can show in the current case that CKR is self-contradictory and, hence, impossible. CKR is then certainly not in general an acceptable assumption. In the current case, let us assume CKR. The student knows the professor is rational; and hence will not give the exam on the final day of the term, because the professor knows that the student is rational and will tell the Dean that there was no surprise exam to get the reward. But the professor knows that the student knows that he is rational and that he knows the student is rational, so the professor knows that he cannot give the exam on the next-to-last day of the term. But the student knows that the professor knows that he knows that the professor is rational and that he the professor knows that he is rational, so the student knows that the professor knows that he cannot give the exam on the next-to-last day of the term. And so on. Therefore, CKR implies that no surprise can be given. However, the professor still can give a surprise exam on Wednesday of the eighth week and the student can follow the logic that indicates that is possible. Since CKR implies a proposition and its negation, CKR must be false. Since CKR implies CKR is false, CKR is self-contradictory.

The second paradox is much deeper, and is related to the famous proof by Tarski (1956[1933]) that the truth predicate is not definable in any system that includes basic arithmetic. Kurt Gödel proved that in any such system, for any well-formed predicate $P(n)$ that takes the values true or false for all natural numbers $n = 0,1, \ldots$, there is a statement z concerning natural number such that $z \equiv \neg P(\#z)$, where $\#z$ is the Gödel number associated with z (this is Gödel's famous diagonalisation argument). Suppose we let the predicate in question be the knowledge operator k, as in Binmore's argument above. The following properties of k are then inconsistent:

(A) $k(\#q) \implies q$;
(B) (A) is known; i.e. $k(\#A)$;
(C) if ϕ is a logical truth, then $k(\#\phi)$
(D) modus ponens: if $k(\#(\phi \implies \psi))\&\ k(\#\phi)$ then $k(\#\psi)$.

In the above, we assume the knowledge operator k applies to the Gödel number of a sentence rather than the sentence itself, so that all semantic notions are replaced by simple arithmetic notions. If a is any sentence, we write its Gödel number as $\#a$. To prove the above four axioms are inconsistent, define the sentence s by $s \equiv \neg k(\#s)$, which is possible by Gödel's diagonalisation argument. Now suppose $k(\#s)$. By the definition of s we have $k(\#s) \implies \neg s$, so by (D), we have $\neg s$. But by (A) and $k(\#s)$, we have s. Thus the assumption $k(\#s)$ was false, and hence $\neg k(\#s)$ is true, which by modus ponens implies s. This whole argument is logically true, so s is logically true and hence by (C), $k(\#s)$. This is a contradiction, proving that axioms (A) to (D) are inconsistent.

The Surprise Examination and Knower's paradoxes are important and interesting in their own right, and they serve as cautionary tales to decision and game theorists who would naively wander into the modal logic of knowledge unaware of its pitfalls. The Surprise Examination paradox warns us of the pitfalls involved in backward induction arguments based on self-references and the Knower's paradox cautions us that seemingly innocuous epistemological assumptions may have pitfalls for the unwary.

It might be thought that the Knower's paradox is not relevant for the Small World vs. Large World question because we can get away with an arithmetic system using only a portion of the Peano axioms, such that the Gödel diagonalisation algorithm cannot be applied. However, Robinson (1950) showed that as long as both addition and multiplication are defined and satisfy the most basic laws of arithmetic, Gödel's argument goes through. It would be bizarre to banish arithmetic from either Small or Large Worlds, so clearly Binmore is correct to raise this as an important question for decision theory.

6. Common Priors and Common Knowledge of Priors

Binmore uses his account of the formation of subjective priors in Small Worlds to show the implausibility of the assumption of common priors. 'In complicated games', he comments, 'one can expect the massaging process to converge on the same common prior for all players only if their gut feelings are similar. But we can only expect the player to have similar gut feelings if they all share a common culture and so have a similar history of experience' (p. 136). This insightful comment, like many others in this book rich with lightly explored insights, is left without further development for the reader to ponder.

Actually, there is much more to be said on this point, as well as the general issue of common priors. If Binmore's suggested process of prior formation is correct, we could expect, at best only *approximately* common priors. Moreover, what is really needed in game theory is *common knowledge* of common priors. If it is common knowledge among rational agents that their priors are formed in a Binmorean manner, and it is common knowledge that they all have the same gut feelings, then we could construct an epistemological justification for the common knowledge of approximately common priors assumption. But, it is hardly plausible that these epistemological preconditions generally hold.

For an example of a case where approximately equal common priors gives a completely different result than common priors, however accurate the approximation, consider a 100-round repeated Prisoner's Dilemma where the first time either player defects, the game is terminated. There is a unique Nash equilibrium in which both players defect on round one and, if there is common knowledge of common priors and common knowledge of rationality, this is what rational players will in fact do. To see this, note that each player, by assumption, has a prior distribution over when his opponent will defect for the first time and chooses when to defect to maximise his payoff subject to this subjective prior. If both players have the same prior, they will choose to defect on the same round. But, if this prior is common knowledge, the defection round would be mutually known, so each would gain from defecting on the previous round, unless the defection round is the first. In fact, it is well known that real players cooperate for many rounds in this game. Assuming they are rational (cooperating, of course, does not violate rationality, although it does violate common knowledge of rationality), there cannot be common knowledge of common priors.

Binmore's interpretation of Small World Bayesianism thus calls into question the capacity of using Bayesian rationality as the basis for game theory, because common knowledge of common priors is assumed as a condition for rational agents playing a Nash equilibrium (Aumann and Brandenburger, 1995), or more generally, a correlated equilibrium (Aumann, 1987).

In Gintis (2009*a*), I come to the conclusion, as Binmore, that the common prior assumption is the product not of 'common culture', but rather of common social norms, in the context of a 'social epistemology' predicated upon evolved characteristics of the human brain. Even then common priors are generally operative only in the case of clear social norms and conventions. Thus, if I encounter a green light while driving, I may place 100% probability on the fact that drivers on the cross street see a red light, that red means 'stop', and that cross street drivers have the same prior.

7. Inference and Updating in Large Worlds

I wish I had read this book five years ago, when I was struggling with the problem of explaining why whole disciplines, including psychology and sociology, so decisively and almost uniformly rejected the rational actor model (each discipline has a small stable of believers but they are generally not even mentioned in the graduate textbooks), despite the fact that this rejection leaves the field devoid of an analytical core. Extensive discussions with anti-rational actor behavioural scientists led me to the realisation that when they said 'decision-making', they instinctively thought of big, fuzzy, ill-specified decisions, such as how to escape from a building on fire, how to choose a mate or a career, or whether to have a life-threatening surgical procedure. By contrast, when economists think of decision-making, they envision the Small World choice of how to fill one's basket at the supermarket. Of course, economists generally show no compunction at extending the model to apply to mate and career choice but it is clearly pushing assumptions beyond their plausible limits to hold that people have complete and consistent preferences over choices they make only once or twice in a lifetime. Sociologists and psychologists reject the rational actor model then, because they do not care about Small Worlds and it does not apply to Large Worlds. This, of course, is exactly Binmore's point about the preposterousness of applying Bayesian models to Large World problems.

While these concerns may explain the rejection of the rational actor model outside of economics, they certainly do not justify this rejection. Indeed, the rational actor model is an important part of a correct theory of human choice behaviour. It is reasonable to argue that the rational actor model must be modified and extended to Large World contexts, not rejected outright. The following is one proposed extension to what might be termed a 'Medium World'.

8. Medium Worlds

Consider the belief revision approach of Alchourron *et al.* (1985) (AGM) inspired by intelligent system computer design. Let $\mathcal{L}$ be a knowledge system, by which I mean the set formulas of the propositional calculus, augmented by a set of atoms P and the classical logical connectives. Let $\mathcal{K} \subseteq \mathcal{L}$ be the set of formulas that the decision-maker knows. I assume $\mathcal{K}$ includes the tautologies of the propositional calculus. For any $\Sigma \subseteq \mathcal{L}$, let $\mathcal{C}n(\Sigma)$ be the logical closure of Σ; i.e., the set of formulas in $\mathcal{L}$ that can be proved using statements from Σ. If $\mathcal{K} = \mathcal{C}n(\mathcal{K})$, so $\mathcal{K}$ is logically closed, I say that $\mathcal{K}$ is a *belief set* (van Ditmarsch *et al.*, 2007). I assume that all formulas in the agent's belief set are considered true with probability one.

Suppose an agent with belief set $\mathcal{K}$ learns a new fact $\phi \in \mathcal{L}$. Then it is plausible to define 'expanding' $\mathcal{K}$ to include ϕ to be the smallest belief set $\mathcal{K} \oplus \phi$ that includes both $\mathcal{K}$ and ϕ. I can then say that the agent *rationally updates* his subjective prior, moving from belief set $\mathcal{K}$ to belief set $\mathcal{K} \oplus \phi$, provided $\mathcal{K} \oplus \phi \neq \mathcal{L}$; i.e., provided ϕ is not contradictory to any belief in $\mathcal{K}$.

Suppose, however, the agent learns ϕ and this is contradictory to some belief in $\mathcal{K}$. Then I assume the individual updates by first dropping a subset of $\mathcal{K}$ such that his new beliefs are a belief set $\mathcal{K} \ominus \phi \subset \mathcal{K}$ and $\neg\phi \notin \mathcal{K} \ominus \phi$. In general there will be more than one way to accomplish this, so I must add additional specifications to indicate exactly how this will be done. For instance, suppose there are several beliefs $\mathcal{B} \subset \mathcal{K}$ that jointly imply $\neg\phi$, and suppose each $\psi \in \mathcal{B}$ asserts that a certain event will occur with positive probability. I may decide to drop from $\mathcal{B}$ the formula with the lowest probability and any others that imply this formula. Having determined $\mathcal{K} \ominus \phi$, the rational agent will update his beliefs by adding ϕ, getting $(\mathcal{K} \ominus \phi) \oplus \phi$.

In other Medium Worlds, I may suppose that the agent has a model of the choice set that is used to define his preferences over the choice set. When an event occurs that violates this model, the agent may make minimal changes in the model, giving rise to a new model that generates consistent preferences, albeit preferences that violate the pre-update preferences. Moreover, if the violation is sufficiently severe, the individual may search for a complete alternative to his current model, with even more severe violations of one or more preferences from the previous model.

Contemporary research in developmental psychology suggests that such a model correctly characterises the way in which babies learn the nature of their environments as well as the rules of language (see Section 10). Indeed, the learning of stereoscopic vision in two to three month old infants can be modelled as updating a genetically determined prior $P(S)$ for the nature of a property S from input D from the visual field, constructing a likelihood function $P(D \mid S)$, and using a neural algorithm to find the S that maximises $P(D \mid S)P(S)$ (Yuille *et al.*, 1999).

9. Rationality vs. Common Knowledge of Rationality

It is very common for economists to confuse the game-theoretic implications of Bayesian rationality with the orders of magnitude stronger common knowledge of rationality (CKR), a confusion that is unlikely to be corrected by reviewers, because they too share in this confusion (Gintis 2009*a*). Binmore of course knows this very well, but he often writes in a manner bound to mislead the reader. Consider for instance the following analysis of updating.

> For example, rational players in game theory are assumed to know that their opponents are also rational. As long as everybody behaves rationally and so play follows an equilibrium path, no inconsistency in what the players regard as known can occur. But rational players stay on the equilibrium path because of what would happen if they were to deviate. In the counterfactual world that would be created by such a deviation, the players would have to live with the fact their knowledge that nobody will play irrationally has proved fallible. (p. 149)

In fact, rationality does *not* imply that 'play follows an equilibrium path' and deviation from an equilibrium path in an extensive form game certainly need *not* imply irrationality. For instance, cooperating on the first round of a ten-round Centipede game (Rosenthal, 1981) does *not* indicate that the player is irrational. Indeed, it takes ten rounds of '*A* knows that *B* knows that *A* knows that… is rational' to ensure that the first player defects on the first round.

I have found this elision of rationality and common knowledge of rationality in several of Binmore's writings. In Binmore (1996), he objects to Robert Aumann's (1995) proof that CKR necessarily implies backward induction. First, Binmore writes 'According to Aumann (1995), common knowledge of rationality in the Centipede makes it irrational for player I to choose across at his opening move.' In fact, of course, it does not make choosing across *irrational*; rather choosing across violates CKR. Therefore, choosing across given CKR is *self-contradictory*, not *irrational*. Second, Binmore writes 'If down is the only Bayesian-rational action at the opening, then $p < 1/2$'. In fact, down is *not* the only Bayesian-rational action; rather it is the only action compatible with CKR. Finally, Binmore states 'if nothing can be said about what would happen off the backward-induction path, then it seems obvious that nothing can be said about the rationality of remaining on the backward-induction path. How else do we assess the cleverness of taking an action than by considering what would have happened if one of the alternative actions had been taken? But this is precisely what Aumann's (1995) definition of rationality fails to do.' In fact, Aumann's proof says *nothing* about what happens off the backward induction path, and certainly does not deny that agents are rational off the backward induction path. He denies that there are nodes off the backward induction path at which CKR holds. This is, of course, correct.

10. The Tiny World of Neuronal Bayesianism

If the passage from Small World to Large World requires considerable reworking of the rational actor model, the passage from the Small World of Savage (1954) to the Tiny World of neural architecture may require broader interpretation but little reworking of the rules of Bayesian updating.

The idea that rational decision involves choice among a personal library of *small scale mental models* can be traced back to Craik (1943). Mental models, according to Craik, have a neural topology that corresponds to a proposed structure of the phenomena they are candidates to represent (Conte and Castelfranchi, 1995). Craik's small scale mental models are accordingly akin to an architect's drawings, to an electronic engineer's schematics, to a molecular biologists' stick-and-ball representation of real molecules, and to a computer scientist's block diagram of a multiprocessor. Cognitive scientists in the Bayesian tradition argue that infants come equipped with a rudimentary repertoire of models for social relationships, for the nature of mind, for language, for causality in the physical world and for other Large World spheres of life with which they must come to terms in the course of maturation. The mind then modifies and chooses among mental models through experience.

I want to review the recent scientific evidence on this 'Tiny World' of neuronal Bayesianism because of its central importance to decision theory and because it reasserts the centrality of the rational actor model in an era that is dominated by popular critiques of this model and its rejection by large numbers of behavioural scientists, who tend to offer nothing in its place except perhaps ad hoceries that work for particular cases but are incapable of generalisation (Gintis 2009*a*, ch. 12).

Bayesian models of cognitive inference are increasingly prominent is several areas of cognitive psychology, including animal and human learning (Courville *et al.*, 2006; Tenenbaum *et al.*, 2006; Steyvers *et al.*, 2003; Griffiths and Tenenbaum, 2008), visual perception and motor control (Yuille and Kersten, 2006; Kording and Wolpert, 2006), semantic memory and language processing (Steyvers *et al.*, 2006; Chater and Manning, 2006; Xu and Tenenbaum, in press) and social cognition (Baker *et al.*, 2007). For a recent overview of Bayesian models of cognition, see Griffiths *et al.* (2008). These models are especially satisfying because they bridge the gap between traditional cognitive models that stress symbolic representations and their equally traditional adversaries that stress statistical testing. Bayesian models are symbolic in that they are predicated upon a repertoire of pre-existing models that can be tested, as well as statistical techniques that carry out the testing and provide the feedback through which the underlying models can be chosen and modified.

While Binmore may well be correct that Bayesian information processing models may not solve the ancient problem of induction (Hume 1975[1777]), they may solve the problem of how humans acquire complex understandings of the world given severely underdetermining data. For instance, the spectrum of light waves received in the eye depends both on the colour spectrum of the object being observed and the way the object is illuminated. Therefore inferring the object's colour is severely underdetermined, yet we manage to consider most objects to have constant colour even as the background illumination changes. Brainard and Freeman (1997) show that a Bayesian models solves this problem fairly well, given reasonable subjective priors as to the object's colour and the effects of the illuminating spectra on the object's surface.

Several students of developmental learning have stressed that children's learning is similar to scientific hypothesis testing (Carey, 1985; Gopnik and Meltzoff, 1997) but without offering specific suggestions as to the calculation mechanisms involved. Recent studies suggest that these mechanisms include causal Bayesian networks (Glymour, 2001; Gopnik and Schultz, 2007; Gopnik and Tenenbaum, 2007). One schema, known as constraint-based learning, uses observed patterns of independence and dependence among a set of observational variables experienced under different conditions to work backwards in determining the set of causal structures compatible with the set of observations (Pearl, 2000; Spirtes *et al.*, 2001). Eight-month-old babies can calculate elementary conditional independence relations well enough to make accurate predictions (Sobel and Kirkham, 2007). Two-year-olds can combine conditional independence and hands-on information to isolate causes of an effect and four-year-olds can design purposive interventions to gain relevant information (Glymour *et al.*, 2001; Schultz and Gopnik, 2004). 'By age four', observe Gopnik and Tenenbaum (2007), 'children appear able to combine prior knowledge about hypotheses and new evidence in a Bayesian fashion' (p. 284). Moreover, neuroscientists have begun studying how Bayesian updating is implemented in neural circuitry (Knill and Pouget, 2004).

For instance, suppose an individual wishes to evaluate an hypothesis h about the natural world given observed data x and under the constraints of a background repertoire T. The value of h may be measured by the Bayesian formula

$$P(h \mid x, T) = \frac{P(x \mid h, T)P(h \mid T)}{\sum_{h' \in T} P(x \mid h', T)P(h' \mid T)}. \qquad (7)$$

Here, $P(x \mid h,T)$ is the likelihood of the observed data x, given h and the background theory T, and $P(h \mid T)$ gives the likelihood of h in the agent's repertoire T. The constitution of T is an area of active research. In language acquisition, it will include predispositions to recognise certain forms as grammatical and not others. In other cases, T might include different models of folk-physics, folk-biology, or natural theology.

11. Conclusion

In this book, Ken Binmore is like the trail guide who is constantly pointing out curiosities of the forest that one would miss when walking alone but urges us to move on before we have fully appreciated them. To get the full benefit of the book, one must return to the trail on one's own and inspect each curiosity in depth at a leisurely pace. I have spent the most of a summer following up a half dozen of Binmore's suggestions, the subjects discussed in this review being the result. I close by reiterating several points:

- We must have a substantive definition of 'rational' going beyond the Savage axioms, without which it is meaningless to talk about rational but not Bayesian decisions. I suggest that a rational decision is one that assesses the implications of alternative choices as accurately as possible given the evidence available and the cost of obtaining new evidence, and choose a course of action that is best fit to achieve the decision-maker's goals.
- There is a wide range of Small Worlds in which the Bayesian rational actor provides a good description of decision-making.
- There is a range of Tiny Worlds in which decision-making is controlled by mostly unconscious neural activity but is well described by Bayesian models working on a pre-given repertoire of potential models of the world.
- There is a range of Middle Worlds in which subjectively zero probability events occur with strictly positive probability and cause macro-level belief revision.
- In Large Worlds, where the Savage Axioms fail, rational decision-makers assess the experience of other rational actors and choose a course of action accordingly.

Herbert Gintis

Santa Fe Institute and Central European University

References

Abbink, K. and Brandts, J. (2008). 'Pricing in Bertrand competition with increasing marginal costs', *Games and Economic Behavior*, vol. 63, pp. 1–31.

Alchourron, C. E., Gärdenfors, P. and Makinson, D. (1985). 'On the logic of theory chage: partial meet contraction and revision functions', *Journal of Symbolic Logic*, vol. 50, pp. 510–30.

Anscombe, F. and Aumann, R. J. (1963). 'A definition of subjective probability', *Annals of Mathematical Statistics*, vol. 34, pp. 199–205.

Arrow, K. J. and Debreu, G. (1954). 'Existence of an equilibrium for a competitive economy', *Econometrica*, vol. 22(3), pp. 265–90.

Aumann, R. J. (1987). 'Correlated equilibrium and an expression of Bayesian rationality', *Econometrica*, vol. 55, pp. 1–18.

Aumann, R. J. (1995). 'Backward induction and common knowledge of rationality', *Games and Economic Behavior*, vol. 8, pp. 6–19.

Aumann, R. J. (1999). 'Interactive epistemology I: knowledge', *International Journal of Game Theory*, vol. 28, pp. 264–300.

Aumann, R. J. and Brandenburger, A. (1995). 'Epistemic conditions for Nash equilibrium', *Econometrica*, vol. 65(5), pp. 1161–80.

Baker, C. L., Tenenbaum, J. B. and Saxe, R. R. (2007) 'Goal inference as inverse planning', Proceedings of the 29th annual meeting of the cognitive science society.

Bala, V. and Majumdar, M. (1992). 'Chaotic tatonnement', *Economic Theory*, vol. 2, pp. 437–45.

Bandura, A. (1977). *Social Learning Theory, Englewood Cliffs*, NJ: Prentice Hall.

Bannerjee, A. (1992). 'A simple model of herd behavior', *Quarterly Journal of Economics*, vol. 107(3), pp. 797–817.

Baron, J. (2007). *Thinking and Deciding*. Cambridge: Cambridge University Press.

Betch, T. and Haberstroh, H. (2005). *The Routines of Decision Making*, Mahwah, NJ: Lawrence Erlbaum Associates.

Bikhchandani, S., Hirshleifer, D. and Welch, I. (1992). 'A theory of fads, fashion, custom, and cultural change as informational cascades', *Journal of Political Economcy*, vol. 100(5), pp. 992–1026.

Binkley, R. (1968). 'The surprise examination in modal logic', *Journal of Philosophy*, vol. 65, pp. 127–35.

Binmore, K. G. (1996). 'A note on backward induction', *Games and Economic Behavior*, vol. 18, pp. 135–7.

Brainard, D. H. and Freeman, W. T. (1997). 'Bayesian color constancy', *Journal of the Optical Society of America A*, vol. 14, pp. 1393–411.

Bush, R. R. and Mosteller, F. (1955). *Stochastic Models for Learning*, New York: John Wiley & Sons.

Carey, S. (1985). *Conceptual Change in Childhood*, Cambridge: The MIT Press.

Chater, N. and Manning, C. (2006). 'Probabilistic models of language processing and acquisition', *Trends in Cognitive Sciences*, vol. 10, pp. 335–44.

Chow, T. Y. (1998). 'The surprise examination or unexpected hanging paradox', *American Mathematical Monthly*, vol. 105, pp. 41–51.

Conlisk, J. (1988). 'Optimization cost', *Journal of Economic Behavior and Organization*, vol. 9, pp. 213–28.

Conte, R. and Castelfranchi, C. (1995). *Cognitive and Social Action*, London: UCL Press.

Courville, A. C., Daw, N. D. and Touretzky, D. S. (2006). 'Bayesian theories of conditioning in a changing world', *Trends in Cognitive Sciences*, vol. 10, pp. 294–300.

Craik, K. (1943). *The Nature of Explanation*, London: London University Press.

Ellison, G. and Fudenberg, D. (1995). 'Word-of-mouth communication and social learning', *Quarterly Journal of Economics* vol. 110, pp. 93–125.

Fisher, F. M. (1983). *Disequilibrium Foundations of Equilibrium Economics*, Cambridge: Cambridge University Press.

Gigerenzer, G. and Todd, P. M. (1999). *Simple Heuristics That Make Us Smart*, New York: Oxford University Press.

Gilboa, I. and Schmeidler, D. (2001). *A Theory of Case-based Decisions*. Cambridge: Cambridge University Press.

Gintis, H. (2007). 'The dynamics of general equilibrium', Economic Journal, vol. 117 (October), pp. 1289–309.

Gintis, H. (2009*a*). *The Bounds of Reason: Game Theory and the Unification of the Behavioral Sciences*, Princeton, NJ: Princeton University Press.

Gintis, H. (2009*b*). *Game Theory Evolving*, 2nd edn, Princeton NJ: Princeton University Press.

Glymour, A., Sobel, D. M., Schultz, L. and Glymour, C. (2001). 'Causal learning mechanism in very young children: two- three- and four-year-olds infer causal relations from patterns of variation and covariation', *Developmental Psychology*, vol. 37(50), pp. 620–9.

Glymour, C. (2001). *The Mind's Arrows: Bayes Nets and Graphical Causal Models in Psychology*, Cambridge: The MIT Press.

Gopnik, A. and Meltzoff, A. (1997). *Words, Thoughts, and Theories*, Cambridge: The MIT Press.

Gopnik, A. and Schultz, L. (2007). *Causal Learning, Psychology, Philosophy, and Computation*, Oxford: Oxford University Press.

Gopnik, A. and Tenenbaum, J. B. (2007). 'Bayesian networks, Bayesian learning and cognitive development', *Developmental Studies*, vol. 10(3), pp. 281–7.

Green, M. S. and Williams, J. N. (2007). *Moore's Paradox: New Essays on Belief, Rationality and the First-Person*, Oxford: Oxford University Press.

Grether, D. and Plott, C. (1979). 'Economic theory of choice and the preference reversal phenomenon', *American Economic Review*, vol. 69(4), pp. 623–38.

Griffiths, T. L. and Tenenbaum, J. B. (2008). 'From mere coincidences to meaningful discoveries', *Cognition*, vol. 103, pp. 180–226.

Griffiths, T. L., Kemp, C. and Tenenbaum, J. B. (2008). 'Bayesian models of cognition', in (R. Sun, ed.), *The Cambridge Handbook of Computational Cognitive Modeling*, Cambridge MA: Cambridge University Press.

Hinton, G. and Sejnowski, T. J. (1999). *Unsupervised Learning: Fundation of Neural Computation*, Cambridge, MA: MIT Press.

Hume, D. (1975[1777]). *Enquiries Concerning Human Understanding and Concerning the Principles of Morals*, Oxford: Clarendon Press.

Juslin, P. and Montgomery, H. (1999). *Judgment and Decision Making: New-Burswikian and Process-Tracing Approaches*, Hillsdale, NJ: Lawrence Erlbaum Associates.

Kahneman, D., Slovic, P. and Tversky, A. (1982). *Judgment under Uncertainty: Heuristics and Biases*, Cambridge: Cambridge University Press.

Kaplan, D. and Montague, R. (1960). 'A paradox regained', *Notre Dame Journal of Formal Logic*, vol. 1(3), pp. 79–90.

Knill, D. and Pouget, A. (2004). 'The Bayesian brain: the role of uncertainty in neural coding and computation', *Trends in Cognitive Psychology*, vol. 27(12) pp. 712–9.

Koehler, D. and Harvey, N. (2004). *Blackwell Handbook of Judgment and Decision Making*, New York: Blackwell.

Kording, K. P. and Wolpert, D. M. (2006). 'Bayesian decision theory in sensorimotor control', *Trends in Cognitive Sciences*, vol. 10, pp. 319–26.

Kreps, D. M. (1988). *Notes on the Theory of Choice*, London: Westview.

Kripke, S. (1963). 'Semantical considerations on modal logic', *Acta Philosophica Fennica*, vol. 16, pp. 83–94.

Kripke, S. (1975). 'Outline of a theory of truth', *Journal of Philosophy*, vol. 72, pp. 690–716.

Lichtenstein, S. and Slovic, P. (1971). 'Reversals of preferences between bids and choices in gambling decisions', *Journal of Experimental Psychology*, vol. 89, pp. 46–55.

Lindley, D. V. (1988). *Making Decisions*, Chichester: John Wiley & Sons.

Luce, D. (2000). *Utility of Gains and Losses: Measurement-Theoretical and Experimental Approaches*, Mahwah, NJ: Lawrence Erlbaum Association.

Luce, R. D. and Raiffa, H. (1957). *Games and Decisions*, New York: John Wiley.

Margalit, A. and Bar-Hillel, M. (1983). 'Expecting the unexpected', *Philosophia*, vol. 13, pp. 263–88.

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

Meltzoff, A. N. and Prinz, W. (2002). *The Imitative Mind: Development, Evollution, and Brain Bases*, Cambridge: Cambridge University Press.

Milnor, J. (1954) 'Games Against Nature', in (R. M. Thrall, C. H. Coombes and R. L. Davies, eds), *Decision Processes*, pp. 49–60, New York: Wiley.

Newell, B. R., Lagnado, D. A. and Shanks, D. R. (2007). *Straight Choices: The Psychology of Decision Making*, New York: Psychology Press.

Oaksford, M. and Chater, N. (2007). *Bayesian Rationality: The Probabilistic Approach to Human Reasoning*, Oxford: Oxford University Press.

Offerman, T. and Schotter, A. (2008). 'Imitation and luck: an experimental study on social sampling', *Games and Economic Behavior*, vol. 65, pp. 461–502.

Pearl, J. (2000). *Causality*, New York: Oxford University Press.

Robinson, R. M. (1950). 'An essentially undecidable axiom system', *Proceedings of the International Congress of Mathematics*, vol. 1, pp. 729–30.

Rosenthal, R. W. (1981). 'Games of perfect information, predatory pricing and the chain-store paradox', *Journal of Economic Theory*, vol. 25, pp. 92–100.

Saari, D. G. (1985). 'Iterative price mechanisms', *Econometrica*, vol. 53, pp. 1117–31.

Savage, L. J. (1954). *The Foundations of Statistics*, New York: John Wiley & Sons.

Schlag, K. (1998). 'Why imitate, and if so, how? A boundedly rational approach to multi-arm bandits', *Journal of Economic Theory*, vol. 78, pp. 130–56.

Schlag, K. (1999). 'Which one should I imitate?', *Journal of Mathematical Economics*, vol. 31, pp. 493–522.

Schultz, L. and Gopnik, A. (2004). 'Causal learning across domains', *Developmental Psychology*, vol. 40, pp. 162–76.

Sobel, D. M. and Kirkham, N. Z. (2007). 'Bayes nets and babies: infants' developing statistical reasoning abilities and their representations of causal knowledge', *Developmental Science*, vol. 10(3), pp. 298–306.

Sorensen, R. A. (1988). *Blindspots*, Oxford: Oxford University Press.

Spirtes, P., Glymour, C. and Scheines, R. (2001). *Causation, Prediction, and Search*, Cambridge: The MIT Press.

Steyvers, M., Tenenbaum, J. B., Wagenmakers, E. J. and Blum, B. (2003). 'Inferring causal networks from observations and interventions', *Cognitive Science*, vol. 27, pp. 453–89.

Steyvers, M., Griffiths, T. L. and Dennis, S. (2006). 'Probabilistic inference in human semantic memory', *Trends in Cognitive Sciences*, vol. 10, pp. 327–34.

Tarski, A. (1956[1933]) 'The concept of truth in formalized languages', in (Tarski, A. ed.) *Logic, Semantics, Metamathematics*, pp. 152–278, Indianapolis: Hackett.

Tenenbaum, J. B., Griffiths, T. L. and Kemp, C. (2006). 'Bayesian models of inductive learning and reasoning', *Trends in Cognitive Science*, vol. 10, pp. 309–18.

Tomasello, M. (1999). *The Cultural Origins of Human Cognition*, Cambridge, MA: Harvard University Press.

Tversky, A., Slovic, P. and Kahneman, D. (1990). 'The causes of preference reversal', *American Economic Review*, vol. 80(1), pp. 204–17.

van Ditmarsch, H., van der Hoek W. and Kooi, B. (2007). *Dynamic Epistemic Logic*, Dordrecht: Springer.

Varian, H. R. (1992). *Microeconomic Analysis*, New York: W.W. Norton.

Vega-Redondo, F. (1997). 'The evolution of Walrasian behavior', *Econometrica*, vol. 61, pp. 57–84.

Walras, L. (1954 [1874]). *Elements of Pure Economics*, London: George Allen and Unwin.

Xu, F. and Tenenbaum, J. B. (in press). 'Word learning as Bayesian inference', *Cognitive Sciences*, vol. 10, pp. 301–8.

Yuille, A. and Kersten, D. (2006). 'Vision as Bayesian inference: analysis by synthesis?', *Cognitive Sciences*, vol. 10, pp. 301–8.

Yuille, A. L., Smimakis, S. M. and Xu, L. (1999) 'Bayesian self-organization driven by prior probability distributions', in (G. Hinton and T. J. Sejnowski, eds), *Unsupervised Learning: Foundations of Neural Computation*, pp. 235–48, Cambridge: MIT Press.

doi: 10.1111/j.1468-0297.2009.02342.x