Comp. by: DRajalakshmi Stage: Proof Chapter No.: 27 Title Name: HbRT
Page Number: 0 Date:12/7/11 Time:01:54:08

27 Paradoxes of Rational Choice Theory

Till Gruene-Yanoff
University of Helsinki, Helsinki, Finland

S. Roeser (ed.), *Handbook of Risk Theory*, DOI 10.1007/978-94-007-1433-5_27,

Abstract: Rational choice theory (RCT) is beset with paradoxes. Why, then, bother with a theory that raises numerous counterexamples, contradictions, and a seemingly endless stream of mutually conflicting remedies? In contrast to this impression, I argue in this chapter that RCT paradoxes play much more productive roles. Eight paradoxes are described in detail, some of their proposed solutions are sketched out, and they are classified according to the kind of paradox they pose. At their example I argue that RCT paradoxes, rather than providing evidence for straightforward rejections of the theory, play important roles in education and in normative research.

Introduction

Rational choice theory (RCT) is beset with paradoxes. Why, then, bother with a theory that raises numerous counterexamples, contradictions and a seemingly endless stream of mutually conflicting remedies? That, at least, may be the impression of a novice student of RCT. In contrast, I argue in this chapter, RCT paradoxes play a much more productive role. Rather than suggesting straightforward rejections of the theory, or repellents to any newcomer, paradoxes play important roles in education and in normative research.

RCT has a clear normative function: it offers tools for judging how people *ought* to form their preferences – and by extension, how they ought to choose. A major problem is that there is no hard basis against which to test normative theoretical claims – one cannot seek to falsify such a theory with controlled experiments. Instead, researchers have to rely on normative intuitions about assumptions and conclusions, and use theory to check whether these intuitions can be held consistently. This is where RCT paradoxes play a crucial role: they elicit normative intuitions that pitch RCT assumptions and conclusions against each other. If a paradox leads to a revision of the theory, it serves a research purpose. If it leads to a better understanding of the assumptions and their conclusions, it serves an educational purpose. Thus, many RCT paradoxes have proved, and continue, to be productive.

The chapter continues with a brief overview of RCT in Sect. 2, recalling the normative claims it really makes. Section 3 discusses how its normative validity can be examined, and the roles paradoxes play in that. Section 4 offers a classification of different kinds of paradoxes. Eight selected paradoxes are surveyed in Sect. 5, sub-sectioned into paradoxes of preference, belief, expected utility, and strategic interaction. Each one is explained, some of its proposed solutions are sketched out, and it is classified according to the scheme proposed in Sect. 4. Section 6 concludes the chapter.

Rational Choice Theory

RCT is the dominant theoretical approach in microeconomics (although economists rarely use the term "rational choice theory"). It is also widely used in other social-science disciplines, in particular political science. In this context, the term rational choice theory is often associated with the notion of economic "imperialism," implying that its use extends economics methodology into their fields.

Explicit theories of rational economic choice were first developed in the late nineteenth century, and commonly linked the choice of an object to the increase in happiness an additional increment of this object would bring. Early neoclassical economists (e.g., William Stanley Jevons) held that agents make consumption choices so as to maximize their own happiness. In contrast, twentieth-century economists disassociated RCT and the notion of happiness: they presented rationality merely as maintaining a consistent ranking of alternatives. Such a ranking is commonly interpreted as agents' desires or values.

Having no foundation in an ultimate end, the notion of rationality is reduced to the consistent ranking of choice alternatives, the consistent derivation of this ranking from evaluations of possible outcomes, and a consistency of beliefs employed in this derivation. Thus, "rationality" explicated in RCT is considerably narrower and possibly sometimes at odds with colloquial or philosophical notions. In philosophical contexts it often includes judgments about ends, the prudent weighting of long-term versus short-term results, and insights into purportedly fundamental moral principles. Nothing of this sort is invoked in RCT, which simply claims that a rational person chooses actions in a manner consistent with his or her beliefs and evaluations. Accordingly, a person considered "rational" in this sense may believe that the moon is made of green cheese, may desire to waste his or her life, or may intend to bring widespread destruction.

At the core of RCT is a formal framework that (1) makes the notion of preference consistency precise and (2) offers formal proof that "maximizing one's utility" is identical to "choosing according to a consistent preference ranking." A brief sketch of this framework follows. (The framework presented here is based on von Neumann and Morgenstern (1947). Alternative formal frameworks are to be found in Savage (1954) and Jeffrey (1990).)

Let $A = \{X_1, \ldots, X_n\}$ be a set of alternatives. Alternatives are either pure prospects or lotteries. A pure prospect is a future event or state of the world that occurs with certainty. For example, when purchasing a sandwich from a well-known international restaurant chain I may expect certain taste experiences with near certainty. Lotteries, also called prospects under risk, are probability distributions over events or states. For example, when consuming "pick-your-own" mushrooms an agent faces the lottery $(X_1,p;\ X_2,1\text{-}p)$, where X_1 denotes the compound outcome (which has probability p) of falling ill due to poisoning and X_2 (with probability $1\text{-}p$) the compound outcome of not doing so. More generally, a lottery X consists of a set of prospects $X_1,\ldots,X_n$ and assigned probabilities $p_1,\ldots,p_n$, such that $X = (X_1,p_1;\ldots X_n,p_n)$. Obviously, the prospects $X_1,\ldots,X_n$ can be lotteries in themselves.

RCT takes preferences over actions to be evaluations of lotteries over action outcomes. Its main contribution is to specify the relationship between preferences over actions, and preferences as well as beliefs over the compound outcomes of the respective lottery. It does so by proving *representation theorems*. Such theorems show that, under certain conditions, all of an agent's preferences can be represented by a numerical function, the so-called utility function. Furthermore, the utility numbers of an action (i.e., lottery) $X = (X_1,p_1;\ldots X_n,p_n)$ and its compound outcomes $X_1, \ldots, X_n$ are related to each other through the following principle:

$$u(X) = \sum\nolimits_{\mathrm{i}} p_{\mathrm{i}} \times u(X_i) \tag{1}$$

In other words, the utility of a lottery is equal to the sum of the utilities of its compound outcomes, weighted by the probability with which each outcome comes about. This is an important result that significantly constrains the kind of preferences an agent can have. Of course, because the representation result is a formal proof, all the constraining information must already be present in the theorem's assumptions. I will sketch the main features of these assumptions here. (For a detailed discussion, see the references in footnote 2. For more in-depth overviews, see textbooks such as Luce and Raiffa (1957), Mas-Colell et al. (1995, Chaps. 1 and 6) and Resnik (1987). Hargreaves Heap et al. (1992, pp. 3–26) give an introductory treatment.)

RCT assumes that, at any time, there is a fixed set of alternatives $A = \{X_1, \ldots, X_n\}$ for any agent. With respect to the agent's evaluation of these prospects, it assumes that agents can always say that they prefer one prospect to another or are indifferent between them. More specifically, it assumes that the agent has a preference ordering $\succeq$ over A, which satisfies the following conditions. First, the ordering is assumed to be *complete*, that is,

$$\text{either } X_i \succeq X_j \text{ or } X_j \succeq X_i \text{ for all } X_i, X_j \in \mathrm{A}. \tag{2}$$

Second, the ordering is assumed to be *transitive*, that is,

$$\text{if } X_i \succeq X_j \text{ and } X_j \succeq X_k, \text{ then also } X_i \succeq X_k \text{ for all } X_i, X_j, X_k \in \mathrm{A}. \tag{3}$$

Completeness and transitivity together ensure that the agent has a so-called weak ordering over all prospects.

The second domain in which RCT makes consistency assumptions concerns beliefs. In particular, it assumes that each rational agent has a *coherent set of probabilistic beliefs*. Coherence here means that beliefs can be represented as probability distributions that satisfy certain properties. In particular, it is assumed that there is a probability function p over all elements of A, and that this function satisfies the following assumptions: first, for any X, $1 \geq p(X) \geq 0$; second, if X is certain, then $p(X) = 1$; third, if two alternatives X and Y are mutually exclusive, then $p(X \text{ or } Y) = p(X) + p(Y)$; finally, for any two alternatives X and Y, $p(X \text{ and } Y) = p(X) \times P(Y|X)$ – in other words the probability of the alternative "X and Y" is identical to the probability of X multiplied by the probability of Y *given* that X is true.

The third domain in which rational choice theory makes consistency assumptions concerns preferences over lotteries. In particular, it assumes the *independence condition*. If a prospect X is preferred to a prospect Y, then a prospect that has X as one compound outcome with a probability p is preferred to a prospect that has Y as one compound with a probability p and is identical otherwise: that is, for all X,Y,Z: if $X \succeq Y$ then $(X,p;Z,1\text{-}p) \succeq (Y,p;Z,1\text{-}p)$.

These assumptions (together with a few others that are not relevant here) imply that preferences over lottery prospects $X = (X_1,p_1;\ldots X_n,p_n)$ are represented by a utility function such that for all X,Y:

$$X \succeq Y \Leftrightarrow \sum\nolimits_i [p_i \times u(X_i)] \geq \sum\nolimits_i [p_i \times u(Y_i)]. \tag{4}$$

This formal result has been given different interpretations. My focus in the following is on the *normative* interpretation of RCT.

Normative Validity and the Role of Paradox in RCT

RCT is often interpreted as a theory of how people *ought* to form their preferences – and by extension, how they ought to choose (for a history of this approach, see Guala 2000). Although the normative content of the theory is limited to the norms of a consistent ranking of choice alternatives, I showed in the previous section that this notion of consistency depends on a number of substantial axioms. This raises the question of the *normative validity* of these axioms: why ought people to choose in accordance with them?

Various attempts have been made to defend the normative validity of RCT and its axioms. The most prominent justifications are pragmatic: they seek to show that agents who fail to retain RCT-consistency will incur certain losses. Two well-known examples are the *money pump* and the *Dutch book* arguments (for more on this and other normative justifications, see Hansson and Grüne-Yanoff 2009, Sect. 1).

Interpreted literally, neither the money pump nor the Dutch book is very convincing. An agent could simply refuse to accept money-pumping trades or Dutch-booking bets. Thus, rationality does not literally require one to be willing to wager in accordance with RCT. Defenders of pragmatic justifications may argue that money pumps or Dutch books reveal possible vulnerability from RCT-violations: a RCT violator might have an incentive to accept a money pump trade or a Dutch book bet, while a RCT-abider does not. However, even such a hypothetical interpretation is problematic. For example, one could deny that the situations considered are normatively relevant to actual preferences. Consequently, it could be argued that norms of preference consistency are primitive in the sense that they are not derived from anything, and in particular not from pragmatic considerations.

Instead, some argue that normative judgments arise directly through human intuition, guided by reflection. These judgments are grounded in characteristic human responses of an emotional or motivating kind. (Such a view does not presuppose a non-cognitivist account of normative judgment. At least on the epistemological level, even cognitivist theories of normativity are likely to appeal to something like natural human responses – no doubt refined by education and reason – to explain how we identify moral facts and evaluate moral claims.) While considerations such as the money pump or the Dutch book may elicit such intuitions, it would be misguided to assume that pragmatic considerations form their basis. Rather, normative intuitions themselves are basic, and form the basis of normative validity judgments of RCT, in this view.

So much the worse for the normative validity of RCT axioms, one might be tempted to reply. To be sure, our emotional or motivating responses to questions of preference consistency often differ and are contradictory. Hence, it seems to follow that any proposed set of axioms is nothing more than the expression of a subjective intuition, fuelled at best by positional or rhetorical power.

Defenders of a stronger validity claim may respond in at least two ways to this challenge. First, they may point out that normative intuitions are not merely claimed to be valid individually, but rather that RCT makes a claim about the normative validity of the whole set of assumptions *and all the results deduced from it*. For example, the maximization of expected utility is a consequence of standard RCT axioms, not an axiom itself. If one has doubts about the normative validity of this conclusion, one has to trace it back to these axioms, re-check their validity, and weight one's doubts in the conclusion against one's confidence in them. This view of normativity thus rests on the idea of a reflective equilibrium: we "test" various parts of our system of normative intuitions against the other intuitions we have made, looking for ways in which some of them support others, seeking coherence among the widest set, and revising and refining them at all levels when challenges to some arise from others (for more on the method of reflective equilibrium, see Daniels 2008).

Second, the defender may point out that normative intuitions are widely accepted only if they withstand being tested in a communally shared effort of "normative falsification" (Guala 2000). Savage described this effort as follows:

> In general, a person who has tentatively accepted a normative theory must conscientiously study situations in which the theory seems to lead him astray; he must decide for each by reflection – deduction will typically be of little relevance – whether to retain his initial impression of the situation or to accept the implications of the theory for it. (Savage 1954, p. 102)

Theorists have to engage in thought experiments in order to elicit these normative intuitions – or "initial impressions," as Savage calls them. Thereby they investigate their normative intuitions in as wide a scope of hypothetical situations as possible, either challenging or confirming particular normative judgments. At the end of this process they only use the intuitions that hold up against normative falsification to challenge the theory:

> If, after thorough deliberation, anyone maintains a pair of distinct preferences that are in conflict with the sure-thing principle, he must abandon, or modify, the principle; for that kind of discrepancy seems intolerable in a normative theory. (Savage 1954, p. 102)

Normative falsification and reflective equilibrium thus go hand in hand: the former generates "corroborated" normative intuitions, and the latter weighs the importance of these intuitions against conflicting intuitions in the theory under scrutiny.

How, then, does one go about normative falsification? How are "situations" constructed in which one obtains "initial impressions" that conflict with the theory? This is where RCT paradoxes come into play. These paradoxes are exemplar narratives of situations that have posed problems for RCT, many of which have been discussed amongst experts for decades. Sometimes agreed-upon solutions exist, and the paradox is used only for pedagogical purposes – to increase understanding of the theory or to illustrate the process of thought experimentation. At other times, competing solutions are offered, some of which may threaten the current theory. In that case, RCT paradoxes constitute the laboratory equipment of ongoing decision-theoretical research.

The Notion of Paradox

Philosophers have distinguished between two accounts of paradoxes. The *argumentative model*, proposed by Quine (1966) and Sainsbury (1988), defines a paradox as an argument that appears to lead from a seemingly true statement or group of statements to an apparent or real contradiction, or to a conclusion that defies intuition. To resolve a paradox, on this account, is to show either (1) that the conclusion, despite appearances, is true, that the argument is fallacious, or that some of the premises are false, or (2) to explain away the deceptive appearances. The *non-argumentative model*, proposed by Lycan (2010), defines a paradox as an inconsistent set of propositions, each of which is very plausible. To resolve a paradox under this account is to decide on some principled grounds which of the propositions to abandon.

Consequently, the argumentative model allows distinguishing different kinds of paradoxes. Quine divides them into three groups. A *veridical paradox* produces a conclusion that is valid, although it appears absurd. (Quine thought of paradoxes pertaining to the truth of deductive statements. In contrast, the validity of decision-theoretic assumptions and conclusions concerns normative validity, which may or may not be reducible to truth. I will therefore use "validity" where Quine spoke of "truth.") For example, the paradox of Frederic's birthday in *The Pirates of Penzance* establishes the surprising fact that a 21-year-old would have had only five birthdays had he been born in a leap year on February 29.

A *falsidical paradox* establishes a result that is actually invalid due to a fallacy in the demonstration. DeMorgan's invalid mathematical proof that 1 = 2 is a classic example, relying on a hidden division by zero.

Quine's distinction here is not fine enough for the current purposes. A falsidical paradox in his terminology, so it seems, can be the result of two very different processes. A genuine falsidical paradox, I suggest, identifies the root of the invalidity of the conclusion in the invalidity of one or more of the assumptions. In contrast, what I call an *apparent paradox* establishes the root of the invalidity of the conclusion in the unsoundness of the argument.

A paradox that is in neither class may be an *antinomy*, which reaches a self-contradictory result by properly applying accepted ways of reasoning. Antinomies resist resolutions: the appearances cannot be explained away, nor can the conclusion be shown to be valid, some premises shown to be invalid, or the argument shown to be unsound. Antinomies, Quine says, "bring on the crisis in thought" (1966, p. 5). They show the need for drastic revision in our customary way of looking at things.

The non-argumentative account rejects Quine's classification, pointing out his assumption of an intrinsic direction in the relationship between "assumptions" and "conclusions." This, so Lycan argues, may give the wrong impression that certain kinds of paradoxes are to be solved in particular ways. Conversely, he points out that two theorists may disagree on whether a paradox is veridical or not: "one theorist may find the argument veridical while the other finds the 'conclusion's' denial more plausible than one of the 'premises'" (Lycan 2010, p. 3). In what follows, I will make use of Quine's classification. Nevertheless, I stress – in agreement with Lycan – that it is only to be understood as an indicator of how the majority of theorists have sought resolution, not as a claim about the intrinsic nature of the paradox itself.

The Paradoxes

Below I survey a selection of paradoxes that are currently relevant to RCT. By relevant here I mean that they challenge one or more of the RCT axioms that are currently in wide use. For more comprehensive literature on paradoxes, also in RCT, see Richmond and Sowden (1985), Diekmann and Mitter (1986), Sainsbury (1988), and Koons (1992).

The survey is structured according to the aspect of RCT under challenge. As it turns out, it is not always clear which axiom is being challenged. I therefore divide the subsections into paradoxes of preferences, belief, expected utility maximization, and strategic choice.

Preferences

Of the many paradoxes challenging assumptions about preferences I will survey two: the Sorites Paradox applied to preference transitivity and Allais' paradox.

Sorites paradoxes are arguments that arise from the indeterminacy surrounding the limits of application of the predicates involved (for a general overview, see Hyde 2008). The Sorites scheme has been applied to RCT in order to cast doubt on the rationality of the transitivity of preference. Quinn's (1990) version goes as follows:

A person (call him *the self-torturer*) is strapped to a conveniently portable machine, which administers a continuous electric current. The device has 1,001 settings: 0 (off) and 1 . . . 1,000, of increasing current. The increments in current are so tiny that he cannot feel them. The self-torturer has time to experiment with the device so that he knows what each of the settings feels like. Then, at any time, he has two options: to stay put or to advance the dial one setting. However, he may advance only one step each week, and he may never retreat. At each advance he gets $10,000.

Since the self-torturer cannot feel any difference in comfort between adjacent settings, he appears to have a clear and repeatable reason to increase the current each week. The trouble is that there are noticeable differences in comfort between settings that are sufficiently far apart. Eventually, he will reach settings that will be so painful that he would gladly relinquish his monetary rewards and return to zero.

The paradox lies in the conclusion that the self-torturer's preferences are intransitive. All things considered, he prefers 1–0, 2–1, 3–2, and so on, but certainly not 1,000–1. Furthermore, there seems to be nothing irrational about these preferences.

► The self-torturer's intransitive preferences seem perfectly natural and appropriate given his circumstances. (Quinn 1990, p. 80)

If this were correct, the normative validity of the transitivity axiom would be in doubt. Defenders of transitivity argue that there is a mistake either in the conception of the decision situation or in the process of evaluation that leads to the intransitive preferences. Both Arntzenius and McCarthy (1997) and Voorhoeve and Binmore (2006) follow the first option in rejecting the implicit assumption that there is a "least-noticeable difference": a magnitude of physical change so small that human beings always fail to detect a difference between situations in which a change smaller than this magnitude has or has not occurred.

Instead, they argue that it is rational for the self-torturer to take differences in long-run frequencies of pain reports into account. In other words, when repeatedly experimenting with the machine he may well experience different amounts of pain at the same notch. He will represent this information about how a notch feels by means of a distribution over different levels of pain: two notches "feel the same" only if they have the same distribution. Then it is implausible that all adjacent notches feel the same when the self-torturer runs through them in ascending order, and the intransitivity disappears.

Thus, Quinn's paradox has been treated as an apparent paradox: an implicit, illegitimate assumption of the derivation – of a "least-noticeable difference" – is exposed, and the dependency of the deductive conclusion on this assumption is shown.

Allais' Paradox (Allais 1953) sets up two specific choices between lotteries in order to challenge the sure-thing principle (an axiom in Savage's decision theory, related to the axiom of independence). This choice experiment is described in ➔ *Table 27.1*.

RCT prescribes that agents choose *C* if they have chosen *A* (and vice versa), and that they choose *D* if they have chosen *B* (and vice versa). To see this, simply re-partition the prizes of the two problems as follows: Instead of "2,400 with certainty" in *B*, partition the outcome such that it reads "2,400 with probability 0.66" and "2,400 with probability 0.34." Instead of "0 with

Table 27.1
Allais' two pairs of choices

Choice problem 1 – choose between:					
A:	$2,500	With probability 0.33	*B*:	$2,400	With certainty
	$2,400	With probability 0.66			
	$0	With probability 0.01			
Choice problem 2 – choose between:					
C:	$2,500	With probability 0.33	*D*:	$2,400	With probability 0.34
	$0	With probability 0.67		$0	With probability 0.66

Table 27.2
The redescribed choice pairs

Choice problem 1* – choose between:					
A:	$2,500	With probability 0.33	*B**:	$2,400	With probability 0.34
	$0	With probability 0.01			
Choice problem 2* – choose between:					
*C**:	$2,500	With probability 0.33	*D*:	$2,400	With probability 0.34
	$0	With probability 0.01			

probability 0.67" in *C*, partition the outcome such that it reads "0 with probability 0.66" and "0 with probability 0.01." Of course, these are just redescriptions that do not change the nature of the choice problem. They are shown in *Table 27.2*.

Through this redescription, we now have an outcome "2,400 with probability 0.66" in both *A* and B^*, and an outcome "0 with probability 0.66" in both C^* and *D*. According to the RCT independence condition, these identical outcomes can be disregarded in the deliberation, but once they are disregarded it becomes clear that option *A* is identical to option C^* and option B^* is identical to option *D*. Hence, anyone choosing *A* should also choose *C* and anyone choosing *B* should also choose *D*.

This result has been found both empirically and normatively challenging for RCT. (In sharp contrast to the RCT result, in an experiment involving 72 people, 82% of the sample chose *B* and 83% chose *C* (Kahneman and Tversky 1979).) On the normative level, many people seem to have intuitions contradicting the above conclusions:

> When the two situations were first presented, I immediately expressed preference for Gamble 1 [*A*] as opposed to Gamble 2 [*B*] and for Gamble 4 [*D*] as opposed to Gamble 3 [*C*], and I still feel an intuitive attraction to these preferences. (Savage 1954, p. 103)

This empirical and normative challenge has given rise to a number of alternatives, including prospect theory (Kahneman and Tversky 1979), weighted utility (Chew 1983) and rank-dependent expected utility (Quiggin 1982). However, the normative validity of these theories is often controversial. Many decision theorists have rather followed Savage, who despite his initial intuitions decided that the sure-thing axiom was, after all, correct. He arrived at this by redescribing Allais' choice situation in yet another form, observing a change of preference from *C* to *D* in this case, and concluded that "in revising my preferences between Gambles 3 [*C*] and 4 [*D*] I have corrected an error" (Savage 1954, p. 103). Decision theorists following Savage have thus treated Allais' paradox as a veridical paradox: the initial impression that the conclusion is absurd is explained away, and the theoretical conclusion is confirmed to be correct.

Belief

I discuss only one paradox of belief here, namely the *Monty Hall problem*. It is posed as follows:

- Suppose you're on a game show, and you're given the choice of three doors: Behind one door is a car; behind the others, goats. You pick a door, say No. 1, and the host, who knows what's behind the doors, opens another door, say No. 3, which has a goat. He then says to you, "Do you want to pick door No. 2?" Is it to your advantage to switch your choice? (vos Savant 1990).

After picking a door at random, it may seem that it is rational to believe that the remaining door holds the car with probability $\frac{1}{2}$. After all, either the chosen door or the other one conceals the prize – so should both doors rather be assigned an equal probability of holding the car?

They should not. At first, before the contestant picks a door, it is rational for him or her to believe that the car is behind any of them with probability $\frac{1}{3}$, knowing that the host will be able to open a door not holding the prize since at least one of the other doors must conceal a goat. Therefore, when the host opens a door the contestant does not learn anything relevant to his or her belief in having chosen the winning door – it remains at $\frac{1}{3}$. Now the offered swap is equivalent to the opportunity of opening both other doors – and he or she should rationally believe that this offers a $\frac{2}{3}$ probability of winning the car. Hence, it is advantageous to swap.

The Monty Hall problem clearly falls into the class of veridical paradoxes: the argument is correct and does not rely on implicit illegitimate assumptions, and the conclusion, despite appearances, is valid. Furthermore (unlike in most other paradoxes discussed here), the result can be experimentally confirmed: the frequency of the prize being behind the other door is observed indeed to converge to $\frac{2}{3}$ (hence there is even a pragmatic confirmation of the normative intuition). The striking thing about this paradox is that the presentation of the correct answer initially created a huge outcry, not least from academically trained mathematicians and logicians (see http://www.marilynvossavant.com/articles/gameshow.html for a selection). The correct solution appeared to be false, but this appearance was explained away with the help of standard probability theory.

Expected Utility

I will now discuss three paradoxes that challenge some fundamental assumptions about the rationality of expected utility maximization: the Ellsberg paradox, Newcomb's problem and the Envelope paradox.

The *Ellsberg paradox* (Ellsberg 1961) goes as follows. An urn contains 30 red balls and 60 other balls that are either black or yellow. You do not know how many black or yellow balls there are, but you do know that the total number of black balls plus the total number of yellow balls equals 60. The balls are well mixed so that each individual one is as likely to be drawn as any other. You are now given a choice between two gambles (*Table 27.3*).

You are also given the choice between these two gambles with regard to a different draw from the same urn (*Table 27.4*).

Standard (Bayesian) decision theory postulates that when choosing between these gambles, people assume a probability that the non-red balls are yellow versus black, and then compute the expected utility of the two gambles. This leads to the following line of reasoning. The prizes are exactly the same. Hence, according to expected utility theory, a rational agent (weakly) prefers *A* to *B* if and only if he or she believes that drawing a red ball is at least as likely as drawing a black ball. Similarly, a rational agent (weakly) prefers *C* to *D* if and only if he or she believes that drawing a red or yellow ball is at least as likely as drawing a black or yellow ball.

Now, if drawing a red ball is at least as likely as drawing a black ball, then drawing a red or yellow ball is also at least as likely as drawing a black or yellow ball. Thus, supposing that a rational agent (weakly) prefers *A* to *B*, it follows that he or she will also (weakly) prefer *C* to *D*, whereas supposing instead a weak preference for *D* over *C*, it follows that he or she will also weakly prefer *B* to *A*. When surveyed, however, most people (strictly) prefer Gamble A to Gamble B and Gamble D to Gamble C. Furthermore, they often insist on this choice, even if the theory is explained to them. Therefore, the normative validity of some assumptions of RCT seems in question.

Ellsberg's paradox poses an interesting challenge to RCT. On the one hand, some scholars have insisted that the standard solution is correct, making it a veridical paradox whose paradoxical impression is explained away. Fox and Tversky (1995), for example, offer an empirical explanation of why people may be biased in their decision-making through an impression of comparative ignorance. Such bias, of course, has no normative validity: it only explains why people have wrong intuitions about what should be chosen. On the other hand, others have argued that this ambiguity aversion is part of a rational decision, in a similar way as a risk aversion is (Schmeidler 1989). Such a position would suggest that the Ellsberg paradox is falsidical, brought about by assuming away the (rationally relevant) impact of ambiguity aversion.

Table 27.3
Ellsberg's first pair of choices

Gamble *A*	Gamble *B*
You receive $100 if you draw a red ball	You receive $100 if you draw a black ball

Table 27.4
Ellsberg's pair of choices

Gamble *C*	Gamble *D*
You receive $100 if you draw a red or yellow ball	You receive $100 if you draw a black or yellow ball

Newcomb's Problem (Nozick 1969) involves an agent's choosing either an opaque box or the opaque and a transparent box. The transparent box contains one thousand dollars ($\$T$) that the agent plainly sees. The opaque box contains either nothing ($0) or one million dollars ($\$M$), depending on a prediction already made concerning the agent's choice. If the prediction was that the agent would take both boxes, then the opaque box will be empty, and if it was that the agent would take just the opaque box then the opaque box would contain a million dollars. The prediction is reliable. The agent knows all these features of the decision problem.

Table 27.5 displays the agent's choices and their outcomes. A row represents an option, a column a state of the world, and a cell an option outcome in a state of the world.

Standard RCT posits that a rational agent should choose the option that maximizes expected utility. This approach recommends taking only one box, for the following reasons. First, the prediction is reliable. In other words, if the agent chooses only one box, then the probability that "take one box" was predicted is high. Similarly, if the agent chooses two boxes, then the probability that "take two boxes" was predicted is high. Hence the probability of outcome $\$M$ given that the agent chose only one box will be high, and the probability of outcome $\$M + \$T$ given that the agent chose two boxes will be low – sufficiently low in most plausible cases for the expected utility of "taking one box" to be higher that that of "taking two boxes." Hence one-boxing is the rational choice according to the principle of expected-utility maximization.

Yet this recommendation violates two deeply entrenched intuitions. First, it violates the principle of dominance, according to which an agent prefers one action to another if he or she prefers every outcome of the first action to the corresponding outcomes of the second. The normative validity of dominance is widely agreed upon. Yet, two-boxing clearly dominates one-boxing in this sense. Consequently, RCT violates dominance.

Second, it violates the intuition that actions should be chosen on the basis of their causal effects rather than their probabilistic correlations to benefits or drawbacks. Because the prediction is made before the agent chooses, the choice has no causal impact on it, and the probabilistic correlation should not matter.

This analysis has motivated a reformulation of decision theory on causal rather than evidential grounds. In various accounts, causal decision theorists seek to represent causal influence with their probability functions rather than with mere probabilistic correlation (see e.g., Gibbard and Harper 1981; Skyrms 1980; Joyce 1999). They clearly see Newcomb's problem as a falsidical paradox based on the misspecification of a decision maker's relevant beliefs.

Some authors opposing the causal approach hold that it yields the wrong choice in Newcomb's problem, in other words two-boxing rather than one-boxing. Horgan (1985)

Table 27.5
Newcomb's problem

	Prediction of one-boxing	Prediction of two-boxing
Take one box	$\$M$	$0
Take two boxes	$\$M + \$T$	$\$T$

and Horwich (1987), for example, argue that one-boxers fare better than two-boxers, and that one-boxing is therefore the rational choice of action. They both see Newcomb's problem as a veridical paradox, and propose explaining away the conflicting intuition about dominance and causal influence with reference to pragmatic success. Causal decision theorists, in turn, reject the relevance of these pragmatic considerations for the validity of the above intuitions. They insist that Newcomb's problem is an unusual case, which rewards irrationality: one-boxing is irrational even if one-boxers prosper.

The *two-envelope paradox* goes as follows. You are asked to make a choice between two envelopes. You know that one of them contains twice the amount of money as the other, but you do not know which one. You arbitrarily choose one envelope – call it Envelope A – but do not open it. Call the amount of money in that envelope X. Since your choice was arbitrary, there is a 50–50 chance that the other envelope (Envelope B) will contain more money, and a 50–50 chance that it will contain less. Would you now wish to switch envelopes?

Calculating the apparent expected value of switching proceeds as follows. Switching to B will give you a 50% chance of doubling your money and a 50% chance of halving it. Thus it seems that the expected value of switching to B is $E(Y\text{-}X) = 0.5^*1/2X + 0.5^*2X - X = 0.25X$. Hence, switching to B will give you a 25% higher expected return than sticking with A. This seems absurd. First, many people have an intuition that one should be indifferent between A and B as long as the envelope remains unopened. Second, once you have switched to B in line with the above argument, a symmetrical calculation could persuade you to switch back to A. Therein lies the paradox.

It is now widely agreed that the expected gain from switching, $E(Y\text{-}X)$, is mathematically undefined because the value of the infinite sum of all probability-weighted values of $Y\text{-}X$ depends on the order of summation (Meacham and Weisberg 2003). However, the conclusions from this observation differ widely. Clark and Shackel (2000) argue that there is a "correct" order of summation, which results in a zero infinite sum, and that this result justifies indifference before opening the envelope. In contrast, Meacham and Weisberg (2003) express reservations about selecting the "correct" order of summation: because the expected gain from switching is undefined, standard decision theory does not rank switching against keeping.

Dietrich and List (2005) go along a different route and offer an axiomatic justification for indifference before opening without appeal to infinite expectations. They supplement standard decision theory with an additional axiom, the "indifference principle," according to which if two lotteries have identical distributions, a rational agent is indifferent between them. From this they are able to deduce a justification for indifference before opening. All three of these responses, although formulated against each other, consider the two-envelope paradox falsidical: Clark and Shackel and Dietrich and List introduce additional assumptions, which yield the intuitive conclusion, whereas Mechaem and Weisberg insist that the argument is fallacious, and no conclusion is warranted from the given assumptions.

Strategic Interaction

Game theory is closely related to RCT. Although it requires certain assumptions beyond those of the standard RCT axioms, its models give additional significance to those standard RCT axioms that also play a role in game theory. For this reason, I include two game-theoretic paradoxes here: the Prisoners' dilemma and the paradox of common knowledge.

The One-Shot *Prisoners' Dilemma* has attracted much attention because standard game-theoretic solution concepts unanimously advise each player to choose a strategy that will result in a Pareto-dominated outcome. It goes as follows.

Two gangsters who have been arrested for robbery are placed in separate cells. Both care much more about their personal freedom than about the welfare of their accomplice. A prosecutor offers each the following deal: choose to confess or remain silent. If one prisoner confesses and the accomplice remains silent all charges against the former are dropped (resulting in a utility of 3 in *Table 27.6* – the first number in each cell is the utility of this outcome for the player who chooses between rows, and the second number, the utility for the player who chooses between columns). If the accomplice confesses and the first prisoner remains silent however, the latter will do time (utility 0 in *Table 27.6*). If both confess, each will get reduced sentences (utility 1), and if both remain silent the prosecutor has to settle for token sentences for firearms possession (utility 2) (for extensive discussion and a literature review, see Kuhn 2009).

The choice situation is solved by appeal to a simple dominance argument. For each player, if the other player stays silent it is better to confess than to stay silent. If the other player confesses, it is also better to confess than to stay silent. Hence, no matter what the other player does, it is always better to confess.

This result is often described as paradoxical in the following sense. The outcome obtained when both confess, although it is rational for each to do so, is worse for each than the outcome they would have obtained had both remained silent. Both would prefer to reach the outcome "stay silent, stay silent," but their individually rational actions led them to the inferior result "confess, confess." (To add some more urgency to this example, consider the structurally similar problem of the "tragedy of the commons," according to which multiple individuals acting independently and rationally will ultimately deplete a shared limited resource even when it is clear that it is not in anyone's long-term interest for this to happen (Hardin 1968).) How can such an inferior outcome be the result of rational decisions?

Some authors argue that the Prisoners' Dilemma indeed exposes a limitation of RCT rationality. Gauthier (1986), for example, suggests that, instead of always confessing, it would be rational for players to commit to playing cooperatively when faced with other cooperators who are equally committed to not exploiting one another's good will. This argument crucially depends on player confidence in that most players are clearly identifiable as being committed to cooperating or not. Whether such a belief can be rationally justified is questionable, and with it the whole solution to the dilemma.

The majority of authors see no conceptual problem in the Prisoners' Dilemma. The assumptions say nothing about the necessary selfishness of the players (charity organizations may also find themselves in such situations!), and no other illegitimate assumptions are

Table 27.6
The Prisoners' dilemma

	Stay silent	Confess
Stay silent	2,2	0,3
Confess	3,0	1,1

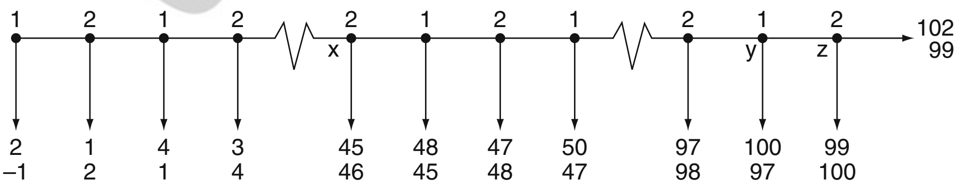
evident. The argument itself is valid, and the conclusion is not contradicted by normative intuitions. The only problem is terminological: some people chafe against the idea that the conclusion is supposed to be the result of rational decision-making. However, they simply subscribe to a different concept of rationality that RCT does not support. Hence the Prisoners' Dilemma is a veridical paradox, whose paradoxical nature relies on terminological ambiguity.

Backward induction is the process of reasoning backward in time, from the end of a problem, to determine a sequence of optimal actions. It proceeds by first considering the latest time a decision can be made and choosing what to do in any situation at that time. One can then use this information to determine what to do at the second-to-last time for the decision. This process continues backward until one has determined the best action for every possible situation at any point in time.

Let us take a concrete example, the "centipede game." This game progresses from left to right in *Fig. 27.1*. Player 1 (female) starts at the extreme left node, choosing to end the game by playing *down*, or to continue (giving player 2, male, the choice) by playing *right*. The payoffs are such that at each node it is best for the player whose move it is to stop the game if and only if he or she expects it to end at the next stage if he or she continues (if the other player stops the game or if it is terminated). The two zigzags stand for the continuation of the payoffs along those lines. Now backward induction advises resolving the game by starting at the last node *z*, asking what player 2 would have done had he ended up there. A comparison of player 2's payoffs for the two choices implies that he should have rationally chosen *down*. Given common knowledge of rationality, the payoffs that result from this choice of *down* can be substituted for node *z*. Let us now move backwards to player 1's decision node. What would she have done had she ended up at node *y*? Given player 2's choice of *down*, she would have chosen *down*, too. This line of argument then continues all the way back to the first node. Backward induction thus recommends player 1 to play *down* at the first node.

What, then, should player 2 do if he actually found himself at node *x*? Backward induction tells him to play "down," but backward induction also tells him that if player 1 were rational he would not be facing the choice at node *x* in the first place. This is not a problem in that Backward induction predicts that player 2 will never find himself at *x* unless both players are irrational.

Yet what does this imply for the Backward-induction reasoning process itself? Backward induction requires the players to *counterfactually* consider out-of-equilibrium play. For example, player 1, according to Backward induction, should choose *down* at node 1, because she knows that player 2 would have chosen *down* at node 2, which in turn she knows because she would have chosen *down* at node 3, and so on, because ultimately she knows that player 2 would have chosen *down* at *z*. She knows this because she knows that player 2 is rational,



Au3

Fig. 25.1

and that player 2 knows that she is rational, and so on. In the language of game theory, because rationality is *common knowledge* amongst the players, backward induction applies (for more on common knowledge, including a formal treatment of this paradox, see Vanderschraaf and Sillari 2009).

Given common knowledge of rationality, each player can affirm the counterfactual "A rational player finding himself or herself at any node in the centipede would choose *down*." Yet we also concluded that if a player *finds* himself or herself at any node with an index number larger than two then both player and opponent know that they *are* not rational. What if that conclusion also made true the counterfactual "if a player *found* himself or herself at any node with an index larger than two then both player and opponent *would* know that they are not rational"? If it did, Backward induction would break down: it requires (1) common knowledge and (2) counterfactual consideration of how players would choose if they found themselves at nodes with indices larger than two. However, (2) implies that both player and opponent would know that they are not rational, contradicting (1). Herein lies the paradox (Pettit and Sugden 1989; Bicchieri 1989).

This is an intensely discussed problem in game theory and philosophy. There is space here only to sketch two possible solutions. According to the first, common knowledge of rationality implies backward induction in games of perfect information (Aumann 1995). This position is correct in that it denies the connection between the indicative and the counterfactual conditional. Players have common knowledge of rationality, and they are not going to lose it regardless of the counterfactual considerations they engage in. Only if common knowledge were not immune to evidence, and would be revised in the light of the opponents' moves, might this sufficient condition for backward induction run into the *conceptual problem* sketched above. However, common knowledge, by definition, is not revisable, and thus the argument has to assume a *common belief* in rationality instead. If one looks more closely at the versions of the above argument (e.g., Pettit and Sugden 1989) it becomes clear that they employ the notion of common belief rather than common knowledge. Hence, the backward-induction paradox is only apparent: the argument that led to the seemingly contradictory conclusion is unsound.

The second potential solution obtains when one shows, as Bicchieri (1993, Chap. 4) does, that limited knowledge (and *not* common knowledge) of rationality and of the structure of the game suffice for backward induction. All that is needed is that a player at each information set knows what the next player to move knows. This condition does not get entangled in internal inconsistency, and backward induction is justifiable without conceptual problems. In that case, the backward-induction paradox is falsidical.

Further Research

I have surveyed eight paradoxes of RCT. There is considerable divergence among them, under a rather rough classificatory scheme, even in this small selection. First, there are the veridical paradoxes, like the Prisoner's dilemma, the paradoxical nature of which rests merely on terminological ambiguity. Ways of explaining away the paradoxical appearance of other veridical paradoxes such as the Monty Hall problem are obvious, but are baffling to the novice. They can still serve an educational purpose, however, in that studying them clarifies the meaning of the assumptions and the derivation of the conclusion.

Second, there are (relatively) clear cases of falsidical paradoxes, such as the two-envelope paradox. Here, there is a clear research result: RCT needs revision.

Third, there are some clear cases of apparent paradoxes, such as the self-torturer, in which the whole bluster is caused by a fallaciously set up argument.

Finally, there are cases on which researchers cannot agree. These include Newcomb's problem, Allais' and Ellsberg's paradox, which vacillates between veridical and falsidical assessment, and the Backward-induction paradox, which vacillates between falsidical and apparent assessment. In all these cases the verdict is still open as to whether they necessitate RCT revision or not. Hence their continuing examination is part of active research in this area.

References

Au4

Allais M (1953) Le comportement de l'homme rationnel devant le risque: critique des postulats et axiomes de l'école Américaine. Econometrica 21:503–546

Arntzenius F, McCarthy D (1997) Self torture and group beneficence. Erkenntnis 47(1):129–144

Aumann R (1995) Backward induction and common knowledge of rationality. Game Econ Behav 8:6–19

Bicchieri C (1989) Self refuting theories of strategic interaction: a paradox of common knowledge. Erkenntnis 30:69–85

Bicchieri C (1993) Rationality and coordination. Cambridge University Press, Cambridge

Campbell R, Sowden L (eds) (1985) Paradoxes of rationality and cooperation: prisoner's dilemma and Newcomb's problem. University of British Columbia Press, Vancouver

Chew SH (1983) A generalization of the quasilinear mean with application to the measurement of income inequality and decision theory resolving the Allais paradox. Econometrica 51:1065–1092

Clark M, Shackel N (2000) The two-envelope paradox. Mind 109(435):415–442

Daniels N (2008) Reflective equilibrium. In: Zalta EN (ed) The Stanford encyclopedia of philosophy, fallth edn. The Metaphysics Research Lab, Stanford, Available online: http://plato.stanford.edu/archives/fall2008/entries/reflective-equilibrium/

Diekmann A, Mitter P (1986) Paradoxical effects of social behavior: essays in honor of Anatol Rapoport. Physica-Verlag, Heidelberg and Vienna, Available online: http://www.socio.ethz.ch/vlib/pesb/index

Dietrich F, List C (2005) The two-envelope paradox: an axiomatic approach. Mind 114:239–248

Ellsberg D (1961) Risk, ambiguity, and the savage axioms. Q J Econ 75(4):643–669

Fox CR, Tversky A (1995) Ambiguity aversion and comparative ignorance. Q J Econ 110(3):585–603

Gauthier D (1986) Morals by agreement. Oxford University Press, Oxford

Gibbard A, Harper W (1981) Counterfactuals and two kinds of expected utility. In: Harper W, Stalnaker R, Pearce G (eds) Ifs: conditionals, belief, decision, chance, and time. Reidel, Dordrecht, pp 153–190

Guala F (2000) The logic of normative falsification: rationality and experiments in decision theory. J Econ Methodol 7(1):59–93

Hansson SO, Grüne-Yanoff T (2009) Preferences. In: Zalta EN (ed) The stanford encyclopedia of philosophy, springth edn. The Metaphysics Research Lab, Stanford, Available online: http://plato.stanford.edu/entries/preferences/

Hardin G (1968) The tragedy of the commons. Science 162(3859):1243–1248

Hargreaves-Heap S, Hollis M, Lyons B, Sugden R, Weale A (1992) The theory of choice: a critical guide. Blackwell, Oxford

Horgan T (1985) Counterfactuals and Newcomb's problem. In: Campbell R, Sowden L (eds) Paradoxes of rationality and cooperation: Prisoner's dilemma and Newcomb's problem. University of British Columbia Press, Vancouver, pp 159–182

Horwich P (1987) Asymmetries in time. MIT Press, Cambridge, MA

Hyde D (2008) Sorites paradox. In: Zalta EN (ed) The stanford encyclopedia of philosophy, fallth edn. The Metaphysics Research Lab, Stanford, Available online: http://plato.stanford.edu/archives/fall2008/entries/sorites-paradox/

Jeffrey R (1990) The logic of decision, 2nd edn. University of Chicago Press, Chicago

Joyce J (1999) The foundations of causal decision theory. Cambridge University Press, Cambridge

Kahneman D, Tversky A (1979) Prospect theory: an analysis of decision under risk. Econometrica 47:263–291

Koons R (1992) Paradoxes of belief and strategic rationality. Cambridge University Press, Cambridge

Kuhn S (2009) Prisoner's dilemma. In: Zalta EN (ed) The Stanford encyclopedia of philosophy, springth edn. The Metaphysics Research Lab, Stanford, Available online: http://plato.stanford.edu/archives/spr2009/entries/prisoner-dilemma/

Lewis D (1979) Prisoner's dilemma is a Newcomb problem. Philos Public Aff 8:235–240

Luce RD, Raiffa H (1957) Games and decisions: introduction and critical survey. Wiley, New York

Lycan WG (2010) What, exactly, is a paradox? Analysis 70(4):615–622

Mas-Colell A, Whinston MD, Green JR (1995) Microeconomic theory. Oxford University Press, New York

Meacham CJG, Weisberg J (2003) Clark and Shackel on the two-envelope paradox. Mind 112(448):685–689

Nozick R (1969) Newcomb's problem and two principles of choice. In: Rescher N (ed) Essays in honor of Carl G. Hempel. Reidel, Dordrecht, pp 114–146

Pettit P, Sugden R (1989) The backward induction paradox. J Philos 86:169–182

Quiggin J (1982) A theory of anticipated utility. J Econ Behav Organ 3(4):323–343

Quine WVO (1966) The ways of paradox. In: The ways of paradox and other essays. Random House, New York, pp 3–20, Original published as Paradox (1962) in Scientific American 206(4):84–95

Quinn WS (1990) The puzzle of the self-torturer. Philos Stud 59(1):79–90

Resnik MD (1987) Choices: an introduction to decision theory. University of Minnesota Press, Minneapolis

Richmond C, Sowden L (eds) (1985) Paradoxes of rationality and cooperation: Prisoners' dilemma and Newcomb's problem. University of British Columbia Press, Vancouver

Sainsbury RM (1988) Paradoxes. Cambridge University Press, Cambridge

Savage LJ (1954) The foundations of statistics. Wiley, New York

Schmeidler D (1989) Subjective probability and expected utility without additivity. Econometrica 57:571–587

Skyrms B (1980) Causal necessity: a pragmatic investigation of the necessity of laws. Yale University Press, New Haven

Vanderschraaf P, Sillari G (2009) Common knowledge. In: Zalta EN (ed) The stanford encyclopedia of philosophy, springth edn. The Metaphysics Research Lab, Stanford, Available online: http://plato.stanford.edu/archives/spr2009/entries/common-knowledge/

von Neumann J, Morgenstern O (1947) The theory of games and economic behavior, 2nd edn. Princeton University Press, Princeton

Voorhoeve A, Binmore K (2006) Transitivity, the Sorites paradox, and similarity-based decision-making. Erkenntnis 64(1):101–114

vos Savant M (1990) Ask marilyn column. Parade magazine, 9 Sept 1990, p 16

Author Query Form

Handbook of Risk Theory
Chapter No.: 27

Query Refs.	Details Required	Author's response
AU1	Please check if insertion of closing quote is okay in extracted text.	
AU2	Please confirm if the inserted in-text citations for Tables 27.3 and 27.4 are correct.	
AU3	Please provide a caption for Figure 1.	
AU4	Kindly cite the following references in text: Campbell and Sowden (eds) 1985; Lewis 1979.	