Introduction

The following is adapted from an example of Haim Gaifman's:[1]

Rowena makes the following offer to Columna: Columna may have either box A (which is empty) or box B (which contains $100), but not both. Rowena also makes the following promise to Columna: if Columna makes an irrational choice in response to the first offer, Rowena will give her a bonus of $1,000. Let us assume that both are ideal reasoners and that Rowena always keeps her promises, and that both of these facts are common knowledge between Rowena and Columna.

How should Columna respond to this situation? If we suppose that taking box A would be irrational, then doing so would yield Columna $900 more than taking box B, which makes taking A the rational thing to do. If, alternatively, we suppose that taking box A would not be irrational, than taking box A would yield at least $100 less than taking box B, so taking box A would be irrational after all. Taking box A is irrational for Columna if and only if it is not irrational.

There is an obvious analogy between this situation and that of the liar paradox. In the liar paradox, we have a sentence that says of itself, 'I am not true.' Such a sentence is true if it is not true (since that is what it says), and it is false, and therefore not true, if it is true (since that is what it denies). Tarski (1956) demonstrated that this ancient puzzle constitutes a genuine antinomy by showing that any theory that implies every instance of an intuitively very plausible schema, convention T, is logically inconsistent. Convention T is simply the requirement that, for every sentence s of the language, our semantic theory should entail the claim that the sentence s is true if and only if φ (where 's' is a name of the sentence 'φ'). For example, where the sentence is 'Snow is white', our semantical theory should imply that 'Snow is white' is true if and only if snow is white.

In order to demonstrate that Gaifman's puzzle also constitutes an antinomy, I must produce intuitively plausible principles concerning

[1] Gaifman (1983), pp. 150–2.

the notion of rationality that force us into inconsistency, just as Tarski produced the intuitively plausible convention T concerning truth. Moreover, these principles should be the ones we are implicitly appealing to the informal reasoning that led to a contradiction earlier. In this book, I will produce such principles, and I will sketch out one way of resolving the antinomy, applying to this case some work on the liar paradox by Charles Parsons[2] and Tyler Burge.[3] Like other antinomies, there is no ordinary, nontechnical solution to this problem. My solution will involve a fairly radical reconstrual of the semantics of the language of justification.

THE SCOPE OF THE PARADOX

Is the Gaifman rationality paradox nothing more than a very artificial and contrived example, of no interest beyond the relatively narrow concerns of the theory of logical antinomies? No. Very close analogues of Gaifman's paradoxical situation recur in a number of heretofore unsolved puzzles in contemporary game theory and game-theoretic economics, as I will demonstrate in Chapter 2. Thus, the Gaifman paradox provides a simplified model by means of which the essential features of these puzzles can be illumined.

I will refer briefly here to three such game-theoretic puzzles: Selten's "chain-store paradox,"[4] the problem of the finite series of "Prisoner's Dilemma" games,[5] and the controversy over the game-theoretic justifiability of deterrent punishment by known act-utilitarians.[6] Selten's chain-store paradox arose from the attempt by game-theoretic economists to analyze and evaluate the rationality of predatory behavior by monopolists. I shall discuss the chain-store paradox in detail in Chapter 2.

The rationality of a strategy of punishment and reward has been discussed in the context of a finite series of Prisoner's Dilemma games by Luce and Raiffa and by Russell Hardin, among others.[7] In the Prisoner's Dilemma game, two guilty prisoners are being separately interrogated. Each faces a choice: either to confess or to hold out. Each

[2] C. Parsons (1974a).
[3] Burge (1979).
[4] Selten (1978).
[5] Luce and Raiffa (1957), pp. 100–2; Hardin (1982), pp. 145–50.
[6] Hodgson (1967), pp. 38–50, 86–8; Regan (1980), pp. 69–80.
[7] Luce and Raiffa (1957); Hardin (1982).

has the following preferences: the best outcome is when one confesses and the other holds out; the second best is when both hold out; the next best is when both confess, and the worst outcome is when one holds out and the other confesses. In a single, isolated Prisoner's Dilemma, it is in one's own interest to confess, whatever the other does. In a long series of games between two players, the issue arises of whether it is rational to try to cooperate by holding out as long as the other holds out (a policy of tit-for-tat). This policy is simply the inverse of deterrence: instead of trying to deter hurtful behavior by punishing hurtful behavior with hurtful behavior, one tries to induce helpful but costly acts by rewarding those acts with helpful but costly acts toward the other. Once again, it turns out that it is rational to be helpful in the first game of such a series if and only if it is not rational to do so.[8]

Finally, the same issue arises in the controversy between Hodgson and Donald Regan concerning whether it is justifiable for an act-utilitarian to punish criminals, given that it is common knowledge in the community that he is a rational act-utilitarian. For a utilitarian, each act of punishment is costly, since even the pain of the guilty subtracts from total utility. Thus, each act of punishment, considered in isolation, is irrational. It is justifiable if and only if it deters potential criminals from commiting future crimes. Assuming again that it is common knowledge that there is some specific, finite number of opportunities for crime, it turns out that such punishment deters crime if and only if it is not rational to think that it does, and so it is rational for an act-utilitarian to punish if and only if it is not rational for him to do so.

In order to understand the essential features of all these examples, it is expedient to examine the simplest one, Gaifman's thought experiment, which I have adapted as a story about *Row*ena and *Column*a. In order to discover the plausible but inconsistent axioms and axiom schemata that underlie our intuitive reasoning about the situation, we must first become quite clear about the meanings of the crucial expressions that appear in the story. When we say that "Columna's taking box A would be rational," we mean that it is justifiable for Columna to think that taking box A is optimal (has maximal expected

[8] Empirical psychological research (e.g., Rapoport and Chammah (1965) indicates that rational players do in fact play tit-for-tat in the beginning of long series of Prisoner's Dilemma games. I am concerned primarily, however, not with the fact of the matter concerning whether tit-for-tat is rational, but rather with the justification of the tit-for-tat policy.

utility), given Columna's total epistemic situation, that is, given the total evidence or data available to Columna in the actual situation.

What, then, do we mean by its being "justifiable" for Columna to think something, given her epistemic situation? Roughly, we mean that the evidence implying the thought in question is stronger than any evidence inconsistent with the thought. In order to make this rough idea precise, we need to develop a theory of how the rational thinker copes with a set of data that may contain unreliable information and may, therefore, be internally inconsistent.[9] Logical deduction alone is not enough, since deduction reveals the implications of a set of assumptions and informs us when that set is inconsistent: it does not tell us what to do after we have discovered that the data set we have been using is inconsistent. (Assuming the logic is classical, it "tells" us to deduce everything from such an inconsistent set, but that is not in practice the reasonable response.)

The notion of 'rational justifiability' dealt with in this book is a rather special one and must be distinguished from a number of other concepts that may be expressed by the same form of words. I am interested here in a notion of 'rationality' that is a generalization of the model of *rational economic man* (or rational political man, etc.) as it occurs in economics and related social sciences. The primary use of such a theory or model of rationality is that of predicting the choices and behavior of agents, given information about the agents' available data and values, goals, desires, and so on.

A certain degree of idealization is essential to such a theory, the assumption being that the effects of mistakes and biases can be dealt with by simply adding the relevant supplementary theories. At the same time, theoretical progress in this area consists in eliminating the unnecessary idealization of agents. A natural progression can be seen here from Ricardo's assumption of unqualified omniscience to the merely logical and mathematical omniscience assumed by the rational expectationists and finally to the resource-bounded rationality of Herbert Simon's theory. The development of a theory of rational belief in Part I of this book parallels this progression, culminating in a resource-bounded account in Chapter 4.[10]

[9] Compare the recent work of Rescher (1976) on "plausibilistic reasoning."

[10] This sort of rationality should be clearly distinguished from the juridical notion of rational justification discussed by such epistemologists as Gettier and Chisholm. Such a juridical

As ideal thinkers, we must assign to the various sources of purported information on which we are relying some degree of apparent reliability, that is, a degree of cognitive tenacity in the face of conflicting data. This degree of reliability cannot be identified with degree of probability, since it does not in general satisfy anything like the axioms of the probability calculus, nor does it have anything much to do with betting ratios. Application of the probability calculus to an individual's judgments presupposes that the individual is "logically omniscient," that is, that the sum of the probabilities of two inconsistent propositions never exceeds 1. Degrees of reliability of data have to do with an earlier, predeductive aspect of ratiocination. We want to consider cases in which two inconsistent sentences both have a very high initial plausibility or apparent reliability, which is possible if their mutual inconsistency is not immediately apparent. When a data set is revealed through logical analysis to be inconsistent or otherwise dissonant, the rational reasoner rejects the elements of the set with the lowest degree of reliability until consistency and coherency are restored.

A reasoner's epistemic situation can simply be identified with the set of sentences that are found by the reasoner to be initially plausible, together with an assignment of a degree of apparent reliability or cognitive tenacity to each such sentence. Ideally, one should accept everything that follows logically from the epistemically strongest, logically consistent subset of one's data. (The "epistemically strongest" such subset is, roughly, the one that preserves the most sentences with the greatest degree of apparent reliability.)

We are finally in a position to explicate the principles underlying Gaifman's paradox. In order to simplify this problem, I will assume that the objects of justifiable acceptance or belief can be identified, for our purposes, with sentences of some formal language that includes the language of arithmetic and a primitive predicate of sentences '$J(x)$' representing the justifiability of accepting sentence (whose code is) x. With such machinery, we can dispense with the details of the Rowena–

notion of justification may be needed in giving an account of when belief (even the belief of a cognitively idealized agent) counts as *knowledge*. It may also be needed by a theory of the ethics of belief, e.g. giving an analysis of the process of defending one's cognitive performances as having satisfied various epistemic duties. I do not wish to denigrate the importance of such research: in fact, I think that a complete theory of rational belief will need to borrow from such research when it gives an account of forming rational beliefs *about* one's own or another's knowledge. Nonetheless, these are two quite distinct sorts of rational justification.

Columna story, since we can, using diagonalization, construct a sentence σ that is provably equivalent (in arithmetic) with the sentence stating that σ is not ultimately justifiable in Columna's epistemic situation (identified with a set of weighted data sentences). Such a sentence will, in effect, say of itself that it is not justifiable in that situation.

As the first principle of justifiability, it is clear that the set of ultimately justifiable sentences, relative to any epistemic situation, is closed under deductive consequence: if a sentence is ultimately justifiable and logically implies a second sentence, then the second sentence is also ultimately justifiable (it will be accepted at some stage of the process just sketched). Let us call this the principle of deductive closure.

The so-called lottery paradox, the paradox of the preface, and similar problems have led some to doubt the principle of deductive closure for justified beliefs.[11] In particular, those who think that the black-and-white accept–reject dichotomy should always be replaced by degrees of assent (subjective probabilities) will be suspicious of this principle.

Nonetheless, the paradox of reflexive reasoning is independent of these issues. First of all, the beliefs to which this principle are to apply are theorems of arithmetic and of epistemic logic. Uncertainty about empirical facts is irrelevant. Typically, mathematical axioms are assigned a probability of 1, so there is no problem about requiring deductive closure. Second, even if we assign a subjective probability of less than 1 to the axioms of arithmetic and of epistemic logic, it is still possible to construct a version of the paradox, replacing the concept of justified acceptance with that of justified degree of belief (rational probability) and replacing the principle with an unexceptionable principle concerning the consistency of rational probabilities (see Section 1.3).

Second, we can assume that all theorems of arithmetic are justifiable (the "justifiability of arithmetic"). This principle enables us to claim that the crucial biconditional – 'σ' is not justifiable if and only if σ – is justifiable in Columna's situation. (The point of the original story was to produce such a sentence: 'taking box A is optimal' is not justifiable if and only if taking box A is optimal.)

Third, we implicitly assumed that anything that we can prove, using general epistemological principles such as these, are among the things that it is justifiable for Columna to accept. The third principle, then, is the rule of inference, which permits us to infer that anything that is provable in the system of epistemic logic we are constructing is jus-

[11] Kyburg (1970).

tifiable, relative to any epistemic situation (a rule of necessitation).

The fourth and last principle is the one that is most difficult to extract from our informal reasoning about Rowena and Columna. As a first attempt, we could produce an inconsistent logic by adding the principle of iteration: If something is justifiable in a given epistemic situation, then it is justifiable in that same situation to think that it is justifiable. Unfortunately, this principle is not very plausible in light of the explication of ultimate justifiability constructed earlier (see Chapter 4 for a fuller discussion of this point).

There is, however, an epistemological principle that is, in the presence of the other assumptions, sufficient for deriving a contradiction and for which there is strong intuitive support. I will call it the principle of negative noniteration: if something is justifiable (in a given situation), then it is not justifiable (in that situation) to think that it is not justifiable. The contrapositive of this principle is perhaps more perspicuous: if it is justifiable to think that something is not justifiable, then it really is not justifiable.

This insight can easily be incorporated into the picture of plausibilistic reasoning already sketched. The principle of negative noniteration represents the fact that there is a kind of cognitive dissonance, comparable to but distinct from logical inconsistency, in holding both p and that one is not justifiable in holding p. At each stage of the process of logical analysis, at least one of p and 'p is not justifiable' will not be tentatively believed at that stage. Therefore, it is impossible for both of them to be ultimately accepted by an ideal reasoner, since if they were both ultimately accepted there would be a stage in the process after which both were accepted continuously by the ideal reasoner, which as we have seen is impossible.

The inconsistency of these four principles can be shown as follows. First, assume (for a reductio) that 'σ' is justifiable. By the justifiability of arithmetic, we know that the conditional

'If σ, then 'σ' is not justifiable'

is justifiable, and by deductive closure it follows that ''σ' is not justifiable' is justifiable. From this, by negative noniteration, it follows that 'σ' is not justifiable, contradicting our original assumption. So 'σ' is not justifiable.

This last conclusion was reached on the basis of three general epistemological principles. By necessitation, we know that this conclusion must itself be justifiable in the relevant epistemic situation:

That is, ''σ' is not justifiable' is justifiable. As the argument makes clear, the rule of inference necessitation is stronger than we need. We could use instead an axiom schema to the effect that any instance of the principles of deductive closure, the justifiability of arithmetic, or negative noniteration is justifiable in every epistemic situation. By the justifiability of arithmetic, we know that the conditional

'If 'σ' is not justifiable, then σ'

is justifiable (since it's provable in arithmetic). By deductive closure, it follows that 'σ' itself is justifiable after all. Thus, we are forced into contradicting ourselves. This paradox is closely related to the "paradox of the knower" of Kaplan and Montague (which will be discussed in Chapter 3).[12]

THE SIGNIFICANCE OF THE PARADOX

The immediate significance of this paradox is threefold. First, any attempt to construct a formal logic of justification and belief (a project of current interest among researchers in artificial intelligence and cognitive science, as well as philosophers) must take this (and certain other related paradoxes) into account, just as any set theorist must take into account Russell's paradox and any truth-theoretic semanticist must take into account the paradox of the liar. The discovery of paradoxes is one of the most important tasks of the philosopher, since through paradoxes we become aware of inadequacies in our naive conception of the relevant concept, be it that of sets, truth, or justification.

A genuine paradox, in the sense in which the liar paradox and Russell's demonstration of the inconsistency of naive abstraction are paradoxes, is more than a merely surprising result. A paradox is an inconsistency among nearly unrevisable principles that can be resolved only by recognizing some essential limitation of thought or language. The liar paradox shows that no sufficiently powerful language can be semantically closed and that, if propositions (objects of thought) possess sentence-like structure, then there can be no unitary, nonrelativized concept of truth that applies to all propositions. Similarly, the doxic paradoxes demonstrate that there can be no such concept of rational acceptability that applies to all propositions.

Second, the clear, explicit formulation of the paradox, together with the realization that it is a liar-like logical antinomy, illuminates the study of several heretofore unrelated problems and puzzles in game

[12] Kaplan and Montague (1960).

Cambridge University Press
978-0-521-41269-8 - Paradoxes of Belief and Strategic Rationality
Robert C. Koons
Excerpt
More information

theory, moral and social philosophy, and economics. As we have seen, the structure of the paradox of reflexive reasoning recurs in such problems as the rationality of cooperation in iterated Prisoner's Dilemma games, the effectiveness of deterrence by known act-utilitarians, and the rationality of predatory behavior by monopolists. In the absence of the discovery of the paradox, each of these problems would have been handled separately and in an unavoidably ad hoc fashion. Such isolated treatment of each problem could lead to distortions of the various fields involved, due to generalizations based on too narrow a range of cases. (This implication is discussed further in Chapters 2 and 7.)

Finally, the paradox of reflexive reasoning sheds light on the general phenomenon of paradoxes or logical antinomies. Discovering a new member of the family of vicious-circle paradoxes is significant, because it enables us to test various generalizations about paradoxicality that were made on the basis of Russell's paradox and the liar paradox alone. In fact, the doxic paradox provides strong reasons for preferring some proposed solutions of the liar paradox to others. In Chapters 5 and 6, I show that context-insensitive solutions do not transfer well to doxic paradox, while context-sensitive ones do.

In the book's conclusion, I discuss some more far reaching implications of my results. First, I conclude that a materialist theory of the mind is compatible with a fully adequate resolution of the logical antinomies. Second, I indicate the implications of this model for the selection of the correct solution concept for noncooperative game theory. Finally, I suggest that the existence of rules and rule following, and, therefore, of institutions and practices, is to be explained in terms of the "cognitive blindspots" that these paradoxes generate. This has significance for ethics as well, specifically, by demonstrating the compatibility of deontic (rule-based) ethics with the rational agent model of decision theory.

PART I

Paradoxes