# METAMIND

*Keith Lehrer*

CLARENDON PRESS · OXFORD
1990

# METAMIND

*Keith Lehrer*

CLARENDON PRESS · OXFORD
1990

# METAMIND

*Oxford University Press, Walton Street, Oxford OX2 6DP*
*Oxford New York Toronto*
*Delhi Bombay Calcutta Madras Karachi*
*Petaling Jaya Singapore Hong Kong Tokyo*
*Nairobi Dar es Salaam Cape Town*
*Melbourne Auckland*
*and associated companies in*
*Berlin Ibadan*

*Oxford is a trade mark of Oxford University Press*

*Published in the United States*
*by Oxford University Press, New York*



*British Library Cataloguing in Publication Data*
*Lehrer, Keith*
*Metamind*
*1. Knowledge. Philosophical perspectives*
*I. Title*
*121*
*ISBN 0–19–824850–4*

*Library of Congress Cataloging in Publication Data*
*Lehrer, Keith*
*Metamind/Keith Lehrer*
*Includes bibliographical references*
*1. Philosophy of mind. 2. Metacognition. I. Title.*
*BD418.3.L44 1990 128'.—dc20 89–37494*
*ISBN 0–19–824850–4*

*Set by Pentacor PLC, High Wycombe, Bucks*

*Printed in Great Britain by*
*Bookcraft (Bath) Ltd,*
*Midsomer Norton, Avon*

For my sons, with esteem and affection
Mark Alan Lehrer and David Russell Lehrer
and in loving memory of my mother

# Acknowledgements

THE author expresses his appreciation to the editors and publishers concerned for permission to reprint articles that were originally published elsewhere. These include: Humanities Press for permission to reprint 'An Empirical Disproof of Determinism?' from *Freedom and Determinism* (Atlantic Highlands, NJ, 1976) edited by the author; Kluwer Publishing for permission to reprint '"Can" in Theory and Practice: A Possible Worlds Analysis' from *Action Theory* (Dordrecht, 1976), edited by Brand and Walton, which appears as 'A Possible Worlds Analysis of Freedom' in this volume; 'Preferences, Conditionals and Freedom' from *Time and Cause* (Dordrecht, 1980), edited by P. van Inwagen; 'Induction and Conceptual Change' from *Synthese* (1971); 'Reason and Consistency' from *Analysis and Metaphysics: Essays in Honor of R. M. Chisholm* (Dordrecht, 1975) edited by the author; 'Metaknowledge: Undefeated Justification' from *Synthese* (1988); Clarendon Press for permission to reprint 'Metamind: Belief, Consciousness, and Intentionality' from *Belief* (Oxford, 1986), edited by R. Bogdan; University of Minnesota Press for permission to reprint 'Induction, Rational Acceptance, and Minimally Inconsistent Sets' in Minnesota Studies in the Philosophy of Science, 6; from *Induction, Probability and Confirmation* (Minneapolis, 1975), edited by G. Maxwell and R. M. Anderson, Jr.; John Hospers, editor of the *Monist*, for permission to reprint 'Social Information' from the *Monist* (1977), which appears as 'Consensual Rationality and Scientific Revolution' in this volume; Comes Verlag for permission to reprint 'Coherence and the Hierarchy of Method' from *Philosophie als Wissenschaft / Essays in Scientific Philosophy* (Bad Reichenhall, 1981) edited by E. Morscher; Hector-Neri Castañeda, editor of *Nous*, for permission to reprint 'The Knowledge Cycle' from *Nous* (1977); and Christopher Hill, editor of *Philosophical Topics* for permission to reprint 'The Coherence Theory of Knowledge' from *Philosophical Topics*, vol. 14, no. 1, *Papers on Epistemology* (1986). I wish to thank Jonelle DePetro for her dedicated editing, proof reading of the manuscript, and work on composing the index; Lois Day for typing the

manuscript and obtaining the permissions referred to above; and Barbara Hannan and Scott Sturgeon for their helpful comments.

K.L.

# Contents

# Introduction

THESE essays, written over a quarter of a century, are unified in ways that might go unnoticed. They are unified in method, and exhibit the methods of analytic philosophy as I understand them. Many of them represent an attempt to analyse some philosophical notion in terms of necessary and sufficient conditions. Such analysis has been under attack for as long as I have been writing these articles. I do not claim, nor have I ever thought, that such a method would inevitably lead to incontrovertible equivalences any more than the methods of science lead to incontrovertible equations. The virtues of the method are that the results are precise and testable. R. M. Chisholm once presented a paper full of very precisely articulated definitions to which a commentator raised some very astute counterexamples. Chisholm remarked that the decisiveness of the counterexamples might show that his paper had a virtue that some of the others lacked, namely, that it was refutable. That was wit, but I found the retort telling. What is the point of philosophical writing if we cannot decide whether what is written is truth or fantasy? We have no way of ensuring that our theories are correct, but we can, at least, express them in such a way that if false, they are refutable. The analytic method effects that end, and I do not know any other method yielding equally testable results. The essays offer the reader the opportunity to judge for herself or himself whether the method provides a useful method for exposition of testable philosophical theories.

These essays are also unified by an underlying idea, one which became clearer to me over the years, and which motivated me to select these articles rather than others. The articles concern freedom, rational acceptance, social consensus, the analysis of knowledge, and, finally, Thomas Reid's philosophy of mind. What could possibly unify such a diverse collection of intellectual reflections? An idea about the human mind. The idea is contained in the writings of Thomas Reid, and so the book might best be read from end to beginning. The

human mind is a metamind. Human freedom, rationality, consensus, knowledge, and conception depend on metamental operations and would not exist without such operations.

## I. A MODEL OF THE METAMENTAL

What is a metamental operation? It is a thought about a thought, about a feeling, or about an emotion. The intentional object of a metamental operation of the mind is itself a mental operation. Moreover, the intentionality of the metamental operation is crucial. A thought causing a thought is not a metamental operation. A thought about a thought is. Intentionality is not causality. So a being might have thoughts that caused other thoughts without having metamental operations or the capacity for metamental operations. Such a being would lack human freedom, rationality, consensus, knowledge, and even general conceptions. Why are metamental operations so important? They provide for our optionality, plasticity; most of all, for the evaluation of lower-level information. The essays in this volume corroborate and illustrate the thesis of the centrality of the metamental.

It is useful to think of the role of the metamental in terms of a simplified computational model of the human mind. The model is a metaphor. The metaphor is a conjecture. To convert it into an articulated scientific theory of the mind is beyond my current stage of investigation. The model is, however, useful for explaining my reinterpretation of these essays. It is familiar in outline. The mind contains an input system that receives information from the outside world and provides us with a representation of that information as an output. That output is the input for a higher-order system, a central system, that evaluates the lower-level information represented by the input system. The evaluation may result in acceptance of the lower-level information or in rejection of that information. The output of such evaluation—acceptance, for example—is a functional state that plays a special role in memory, inference, and action. It is the sort of state that ordinarily results from reflectively judging the information to be correct; but the same sort of functional state also arises unreflectively from the processing of



information. We arrive at the same sort of functional state by more than one historical route. Representation of information is the function of the input system, while evaluation and application of the information is the function of the central system.

In addition to the input system and central system, there is an output system resulting in action. This system takes representations of actions as input to be executed. These representations are supplied by the central system. Usually representations of the input system are accepted by the central system in a routine manner and the representations for the output system are supplied by the central system in the same manner. Since such operation is routine, and must be so for the sake of efficiency, it may appear automatic. The central system mimics automatic operations of the mind so closely that the operations become almost invisible. They reveal themselves, however, when some untoward circumstances trigger more reflective processing. The central system is virtually ubiquitous in monitoring the operations of the mind. Indeed, I would suggest that the boundary between purely physical processes, secretions in the stomach to digest food, for example, and mental information-processing is marked by the metalevel monitoring of the central system. The operations of this system are metamental operations on the operations of the input system and the central system itself.

This model leaves many questions open. Can the central system ever direct the operations of the input system? Can it, for example, direct how information is represented by the input system, or is the input system encapsulated in the sense that the central system is limited to processing the output of the input system? Similarly, can the central system direct the output system? Can it, for example, direct what action is executed from the representation of an action in the output system, or is the output system encapsulated in the sense that the central system is limited to supplying input for the output system—for example, representations of actions to be executed? Are there multiple independent input systems and output systems? Are all the representations, or even all atomic representations innate? In the present state of enquiry, which is as conjectural as it is fascinating, such questions should be left open.

There are two features of the central system that are important for the reinterpretation of these essays. One is the almost

ubiquitous presence of the metamental operations of the central system, and the other is the functional or computational character of such states. The latter feature enables us to attribute higher-order states (such as accepting that one accepts something or preferring that one prefers something) to human beings without supposing that those states are the objects of conscious reflection. These states are to be understood computationally or functionally in terms of the role they play in the operations of the central system. The model or metaphor of a central system, however limited the articulation, suffices to shed a new light on the subjects dealt with in these essays.

## II. FREEDOM

The chapters in this volume on freedom constitute a defence of compatibilism, that is, the thesis that human freedom is compatible with the claim that everything that occurs is caused or that there are antecedent sufficient conditions for everything that occurs, including human thought and action. Human freedom implies that a person could have done otherwise, and so compatibilism implies that a person could have done otherwise even though the person was causally determined to do what she did. The simple conditional analysis of 'could' is rejected in the first chapter. Since such an analysis has been used to defend compatibilism, my argument against the analysis has been approved of by the opponents of compatibilism, Anscombe, for example, and attacked by compatibilists, Goldman and Davidson, for example.[1]

Nevertheless, the essay contains a defence of compatibilism. The defence is simple. We sometimes have evidence that renders it highly probable that we could have done otherwise, but that evidence does not render it highly probable that determinism is false. If, however, the claim that we could have done otherwise entails the falsity of determinism, which would be the case if freedom and determinism were incompatible, then

[1] G. E. M. Anscombe, 'Soft Determinism', in G. Ryle (ed.), *Contemporary Aspects of Philosophy* (Stocksfield, 1977), 148–60; A. I. Goldman, *A Theory of Human Action* (Englewood Cliffs, NJ, 1970); and D. Davidson, 'Freedom to Act', *Essays on Actions and Events* (Oxford, 1980), 63–81.



any evidence which rendered it highly probable that the thesis of freedom is true would render it highly probable that the thesis of determinism is false. Therefore, the thesis of freedom does not entail the falsity of determinism, and freedom and determinism are compatible.

The argument has had its distinguished detractors, most of whom have claimed that it begs the question. I do not think the argument has the form of a question-begging argument, but I concede that, however formally correct and sound the argument might be, it is unsatisfying to many philosophers for a simple reason. Even if the argument proves the compatibility, it does nothing to explain how the compatibility is possible. How can it be possible that a person could have done otherwise when an action is causally necessitated and, indeed, as one modification of determinism implies, causally necessitated by conditions over which the agent has no control?

To explain this, one needs some analysis or account of freedom, and this I present in the next two chapters. I argue that a person could have done otherwise in a specific instance just in case there is a possible world accessible from the actual world in which the person does otherwise, having no advantages for so doing in the possible world which she lacks in the actual world. I proceed to refine the analysis. The absence of the antecedent conditions causing the action in the actual world need not constitute an advantage. When the question arises as to whether I can do something at a certain time, it is the conditions that exist at that time, and not how those conditions were caused, that make the difference. This claim requires modification and qualification, but it is correct when applied to simple motions of one's body. Whether I can move my hand now depends on the conditions that now exist, and it does not matter at all whether those conditions are ones that were causally determined by something that occurred in the remote past or ones that are undetermined. It is what I am like now and what the surrounding conditions are like now that settles the matter of whether I can now move my hand. It does not matter how I got that way but only what I am now like. This reflection concerns freedom and not moral responsibility. Morality drags the past along like a string.

*A question remains.* Even if we restrict our consideration to

present conditions, it appears as though those conditions bring about the action that they do in a way that is incompatible with freedom. Consider the strongest sort of candidate for a free action, one that involves deliberation and decision. If we suppose that determinism is true, then the action is causally determined by the thoughts, feelings and so forth that produce the decision to act. The train of thought in the mind runs as relentlessly as a locomotive. I may prefer to do one thing rather than another, and I may act on the preference, but the entire sequence is causally determined, and the components are locked together, one with the next, as tightly as the cars in a train. There is no free play, only inevitability. So how is compatibilism possible?

The answer is contained in the doctrine of the metamental. In Chapter 3 I argue, adopting an idea from Jeffrey and Frankfurt,[2] that freedom involves preferences among preferences. My preferences concerning preferences are obviously metamental operations. In addition to having first-order preferences, I have higher-order preferences concerning those first-order preferences. Moreover, higher-order reflection is not merely passive observation but may be effective in altering first-order preferences as well. This is not automatic. It is not by any means always the case that, when I decide that a first-order preference is undesirable, the first-order preference disappears. Everyone driven by habit or obsession knows this. I may lack control over my preferences and not act freely. If I do what I prefer, and if I prefer to prefer what I do, and if I would do otherwise if I preferred otherwise, and if I would prefer otherwise if I preferred to prefer otherwise, and so on up the ladder of preferences, then I will have metamental control over my preferences, and I act freely. The metamental suffices for freedom. Roughly speaking, that is the thesis of the third chapter.

If, moreover, my beliefs and preferences were unaffected by my metamental operations upon them, or if I were incapable of such operations, the beliefs and preferences would not be under my control and my resultant behaviour would not be free. I could not do otherwise. For this reason, a compatibilist

[2] R. C. Jeffrey, 'Preferences Among Preferences', *Journal of Philosophy*, 71 (1974); and H. Frankfurt, 'Freedom of the Will and the Concept of a Person', *Journal of Philosophy*, 68 (1971).



argument assimilating what a simple machine can do to what a person can do, fails. The argument runs as follows. It is perfectly consistent to say that a machine can do something which it does not do, even though we are perfectly convinced that what the machine does is causally determined. Consider an electric clock that is not turned on. It is correct to say that it can run, assuming it to be in perfect working order, even though it is not running at all. Similarly, a compatibilist might argue, it is correct to say that a person can run even though she is not running. Such arguments are as spurious as they are common. The clock has no control over whether it runs or not. The sense in which it can run when it does not is insufficient for saying that it can run. The clock can run in this sense only: it will run if someone starts it. If the runner is shackled, she can run if she is unshackled. But whether she is shackled is beyond the control of the would-be runner. It is not enough that a person would do something if she preferred; the preference must also be within her control. Her preferences will be within her control only if she has cognitive access to her preferences and the power to influence them.

There is an obvious objection to the idea that freedom depends on metamental operations, namely, that this provides too intellectual an account of freedom to apply in many instances in which we wish to claim that a person acts freely or could have done otherwise. I might turn on my computer without having deliberated about whether to do so, for example. My action is a free action, I could have done otherwise, but that is not because of any reflection on my preference to do so. I acted on the preference. I did not think about it. I did what I preferred without the intervention of any metamental operation. So runs the objection. The answer is to admit that I did not reflect on the action. There are free but unreflective actions. That does not mean, however, that no metamental operation is involved. I am conscious of my preference, and that is a metamental operation. I am conscious of acting on the preference as well. Both the preference and acting on the preference are, as it were, metamentally monitored and certified, though not reflectively.

In terms of our model, the output system responds to current beliefs and preferences with action. The output system is

intended to be fast, to imitate a fully automatic system, but to be under the control of the central system. The central system has access to the total background information of the mind and is capable of generating further information; for example, concerning the set of alternatives open to the agent. The output system responds to limited information to the effect that some course of action is preferred. Though the output system is under the control of the higher-order central system, the output system will often operate in a manner that conceals the relationship between the two systems. The reason is that intervention by the central system and application of the information to which it has access is often too slow and inefficient to meet the exigencies of practical life. Hence, simple strategies, rules of thumb, are applied by the central system in most circumstances.

The metamental monitoring of the central system becomes obvious when trouble threatens. That is when our true nature reveals itself. Being thirsty, I will accept and drink a glass of water offered to me, in a manner that appears automatic. The rule is to drink when thirsty, and, in most instances, the output system functions rapidly and unreflectively in conformity to the rule. The action is accepted by the central system in accordance with the rule. The monitoring and acceptance may go on unnoticed, but the role of the central system becomes manifest when it blocks the customary action; for example, when the appearance of the water is unsavoury.

Of course, not all behaviour is under the control of the central system. Some behaviour is reflexive. Other behaviour, such as a nervous tic, though not reflexive, is not under our control. It would result even if we preferred that it did not. Other behaviour, though it would not occur if we were to so prefer, results from preferences that are not in our control. If I am erotically obsessed with a woman I may seek her company, thus responding to my preference, even though I am convinced I would be better off without her and would prefer not to have the preference for her company that drives me to her. Such behaviour entertains the soap-opera fan and pays the psychiatrist, but is not free. In short, my freedom is abrogated when my behaviour is not controlled by my metamental operations. Metamindless brutes lack freedom, as do metamindless people.

There is an objection to the theory of freedom based on



higher-order preferences. It is that the metamental operations are themselves part of a causal order, that they result with the same inevitability as any other part of the causal chain and, therefore, that they are inadequate to account for the difference between those actions that are free and those that are not. This objection seems to me to be fundamentally in error. The error is simple enough to discern. Being a compatibilist, I concede that all human actions are caused, along with everything else. Hence the difference between free actions and those that are not free must be a difference in the way that they are caused. This does not beg the question against the incompatibilist, however, because it is not an argument for compatibilism. The proof of compatibilism is that of the first chapter. My purpose in these remarks is to explain how compatibilism can be true, not to show that it is true. To explain how it can be true, one must explain how a difference in the way in which actions are caused accounts for the difference between those that are free and those that are not free. My explanation is in terms of higher-order preferences and the effectiveness of metamental operations.

There are lower-level processes resulting in action and belief. But there is something peculiar about us. We can evaluate those lower-level processes in some cases and decide whether to accept what they indicate. I find that I have certain desires and indifferences, but these are subject to higher-order evaluation. Sometimes preferences resulting from rational consideration of lower-level indifferences will conflict with those indifferences. Indifference of desire is intransitive, for example, but the necessity of choosing without cycling may require the conversion of such lower-level indifference into transitive preferences. I have argued, with Carl Wagner, that in some cases there are equally reasonable but incompatible strategies for such conversion.[3] Selection of a strategy for conversion is not a matter of the strength of desire but of higher-order rationality concerning first-order preference. Such problems are fairly common. Anyone reflecting on the choice of an item from a menu has, I believe, experienced the inadequacy of desire as the basis for choice. Stephan Körner once remarked to me in a restaurant in Graz that he did not care what he got so long as it was schnitzel.

[3] K. Lehrer and C. Wagner, *Rational Consensus in Science and Society: A Philosophical and Mathematical Study* (Dordrecht, 1981).

It was a joke, no doubt, but it may have represented the triumph of higher-level rationality over lower-level indifference. In need of a principle of preference, Körner preferred schnitzel. Metamental rationality is the antidote to starving from indifference.

## III. RATIONAL ACCEPTANCE

As we progress from the essays about freedom of action to those concerning rational acceptance, we find the same important contrast between a lower-level system and the higher-order system of evaluation. One might take the same hard-deterministic line with respect to belief as was proposed with respect to action. The hard determinist says there is no freedom of action because action is the inevitable outcome of causal antecedents. Similarly, the hard determinist might say that there is no rationality of belief because belief is the inevitable outcome of causal antecedents. Since both action and belief are the inevitable outcome of causal antecedents, we may, it might be concluded, put aside all questions of responsibility, freedom, and rationality, and replace them with a naturalistic account of causal antecedents. We would, therefore, naturalize both morality and epistemology, and substitute causal enquiry for traditional philosophical theory. This proposal is an error because metamental operations are part of the natural order. I do not object to causal enquiry. I object to the oversimplified account of the causal order. Metamental operations take us beyond the simple model of a chain of causation to our intervention in the chain. The naturalizers of morality and epistemology, in a moment of naturalistic absent-mindedness, forgot to include themselves in the causal order.

Metamental operations are as important in the intellectual domain as in the practical. Rational acceptance of statements for the purposes of attaining truth and avoiding error should be viewed as a product of metamental rationality. Information is received by an input system. Described at the level of common sense, the output of the input system is conception and belief, though many aspects of doxastic states actually result from central processing. Described in terms of contemporary science it is representation processed in a default mode. The information received at this level is further processed at a higher level. Additional representations are constructed, those of science, for example, and distinctions in attitude toward representations are introduced. These are the metamental operations of the central system.



How does the central system function? It is a system driven by various goals and ends. Among these is the attainment of truth and the avoidance of error. These objectives contain some stress between them. If we focus on the objective of accepting all that is true, we shall be bold in what we accept and ignore the risk of error. If we focus on the objective of accepting nothing that is false, we shall be cautious in what we accept and ignore the opportunity to accept what is true. We may combine the objectives of attaining truth and avoiding error by striving to accept something if and only if it is true. But there is no obvious way to pursue this goal. As a result, I suggest that the task of acceptance is divided. Some things are accepted as evidence by central processing. We sift the representations of the input system, and reconstructions of those representations at a higher level, to accept some as evidence. What we accept as evidence is the basis for the acceptance of hypotheses. If we err in what we accept as evidence our error may breed further error in what we accept as hypotheses. Hence, consideration of the risk of error, though not the only relevant factor, is the most important, and caution is the reasonable strategy. When we turn to hypothesizing, on the other hand, the quest for truth becomes paramount. The point of accepting hypotheses is to extend our grasp of the truth, and boldness is the reasonable strategy. These subjects are treated in the chapters on induction.

Acceptance is a metamental operation. Belief, for the most part, is a lower-level operation. In the usual case, belief is carried over into acceptance. Indeed, the default mode of the central system is to accept what is believed, especially in the case of perceptual belief. Put another way, a rule-of-thumb principle for the central system is to accept perceptual belief as evidence. But the default mode or the rule of thumb will be overridden when background information indicates that perception is untrustworthy. One function of the central system is to sort lower-level representations, accepting some as evidence, others as hypotheses, and refusing to accept others at all.

The higher-order system also adds representations and, contrary to some theorists, is genuinely constructive in producing new representations that are not mere combinations of some previous conceptions. This constructive activity of the central system is most apparent when some new representation is introduced that is logically inconsistent in terms of the preceding representational system. The representation of a monad in Leibniz is a good example. The idea that stones are composed of minds was, I suggest, a logically inconsistent proposal in terms of preceding representations of stones. Similarly, the representation of an atom that was divisible was a logically inconsistent proposal in terms of antecedent representations of atoms. Thus, an adequate theory of rational acceptance of formerly inconsistent representations must contain an explanation of how formerly inconsistent representations can be altered so that they are subsequently consistent and rationally accepted. This task is undertaken in the chapter on conceptual change. All that I would now add to that account is that such semantic change is the result of metamental operations which are the source of semantic plasticity.

## IV. CONSISTENCY

Consider the rationality of consistency. There is an obvious development of thought in the essays on consistency, but they are unified by consideration of the role of consistency in the rational acceptance of statements. If the set of things accepted as evidence, for example, is logically inconsistent, then it is logically impossible to avoid error or to assign probabilities on the basis of such evidence. The prospects for avoiding error might not be very good in any event, but to adopt a strategy that makes it logically impossible to attain a goal you are striving to achieve seems irrational. Nevertheless, it is quite clear that some of the things one believes, and, indeed, even some things one accepts as evidence, will turn out to be false. This metamental evaluation of what one believes or accepts is something that it seems quite rational to accept. But it becomes logically impossible that everything one accepts is true once one accepts that at least some of the things one has already accepted are false. Moreover, there will be cases in which one must choose



between being arbitrary, inconsistent, and extremely cautious. When a large or important class of equally reasonable hypotheses turns out to be inconsistent, one must choose between arbitrarily rejecting one or more to obtain consistency, accepting all the hypotheses though they are inconsistent, and refusing to accept any of the hypotheses. The argument for the consistency strategy may prove inconclusive, as noted in some of the essays.

Are the essays simply inconsistent with each other, some assuming consistency as a constraint on rationality and some not? Consistency is always a desideratum of rationality because inconsistency logically guarantees error. It is essential that evidence used to assign conditional probabilities be logically consistent. In the case of other truth-directed forms of acceptance, inconsistency, though undesirable, may be warranted. The combination of the objectives of obtaining truth and avoiding error into one directive, to accept something if and only if it is true, might best be served in some instance by accepting an inconsistent set of statements. It may be that one can obtain a more comprehensive acceptance system with a greater balance of truth over error by accepting an inconsistent set of statements. This evaluation is a metamental evaluation. If one considers various acceptance systems, one of which is inconsistent, one may conclude that accepting the inconsistent set provides the best conformity to the directive to accept something if and only if it is true.

Thus, I consider the earlier essays to articulate the results of constraining rationality by a requirement of consistency. Since logical consistency is a desideratum and one that must be fulfilled if one is to avoid error, the essays remain useful as revealing how productive the constraint of consistency is in the domain of rationality. The metamental reflection on the constraint suggests that it will not be a rational constraint on the entire system of what we accept at any one time, though it is a constraint that must be satisfied if we are to be ideally successful in our quest to obtain truth and avoid error.

## V. PROBABILITY

Any evaluation of the rationality of acceptance as well as the rationality of action must be based on probability. The rational

pursuit of any goal, intellectual or practical, rests not only on the merit of the goal but on the probability of attaining it. Rational acceptance aiming at the attainment of truth and the avoidance of error depends on the probability of satisfying those goals by accepting some statement or representation as true. I am a qualified subjectivist or personalist with respect to probability. I consider probabilities to be summaries of the total information of a subject in quantitative form. In that way, I am a subjectivist. I am a qualified subjectivist in that I think of the subjective probabilities as estimates, at least in some cases, of something objective, of frequencies or propensities; and their role as estimates may provide some constraint on how they can be assigned. Assuming probabilities to be a summary of the total information a person possesses at a time, we relieve ourselves of the need to require that a person search through all her information, which would demand an unrealistic search-procedure, to determine what information is relevant to her decision of what to accept or do. The probability assignment of a person represents the total information of a person, including conditions of relevance, in a neatly wrapped mathematical package.

Probability assignment raises a problem which can only be solved by ascent to the metamental. The problem is that our initial probability assignments, like our initial desires, are incoherent. To render them coherent, we must process our initial probabilities and desires at a higher level to convert them into coherent probabilities and preferences for rational acceptance and action. The imposition of a coherence constraint, like the imposition of a consistency constraint, may not always be warranted, even for a rational agent. One might be more effective in seeking truth with incoherent probabilities than with coherent ones, just as one might be more effective in seeking truth by accepting an inconsistent set of statements.

Once we consider probabilities as estimates of frequencies, moreover, the problem of consistency and the problem of coherence become the same. It is, as we know from the theory limits, logically impossible that a set of incoherent estimates should all be correct. We can, therefore, reach beyond the practical objective of avoiding a probability assignment that would lead us to gamble in a 'no win' manner against a mathematically sophisticated opponent, to the objective of



avoiding a probability that would lead us to estimate frequencies in a way that guarantees error. The revision of lower-level probabilities, necessary for ideally fulfilling our objective of accepting and estimating exactly what is true, exhibits once again the role of the central system and metamental operations.

I cannot resist a relevant aside. Philosophers are inclined to make a great deal of the robust experimental results showing that experimental subjects, even learned ones, routinely miscalculate logically and mathematically. The results are, indeed, fascinating. To argue from these results to the conclusion that human rationality does not conform to the logical and mathematical principles that are violated, however, is unwarranted, because correct calculation in accordance with those principles is also a human ability. The experimenters are, after all, human as well, and their representation of the correct calculations and our understanding of them is what makes the results interesting. The moral should be obvious. The miscalculations are the results of the application of simple rules of thumb. The correct calculations are the result of more disciplined computation by the central system. There is, as the experiments demonstrate, a lower-level system that does not conform to the constraints of rationality, but there is also, as the calculations of the experimenters illustrate, a higher-order system conforming to the constraints of rationality. We all contain both systems.

## VI. CONSENSUS

The next essay is concerned with the problem of rational consensus. Here I assume that subjective probabilities are summaries of information and consider the consequences of a person evaluating the probabilities of members of a group, assigning weights reflecting the comparative evaluations, and averaging. The evaluation of the probabilities of various people by one person is an interpersonal metamental operation. It has an exact analogy in the intrapersonal metamental evaluation of the conflicting probabilities of one person by himself. I have gone beyond this essay in a book with Wagner.[4] The convergence theorems appealed to in this essay result from more

[4] Lehrer and Wagner, *Rational Consensus*.

severe constraints than are necessary. The model Wagner and I developed was a model of ideal interpersonal rationality. I include these essays because they illustrate higher-order evaluation of lower-level information and because of the analogy between interpersonal and intrapersonal conflicts and their method of resolution. I suggest weighted averaging of conflicting probabilities as a method of rational conflict resolution in both cases. But averaging will not always be rational.

In the case of interpersonal conflict, an iconoclast may reasonably refuse to average. This is accomplished by refusing to assign others a positive weight with respect to the issue in question. A consensual probability of one group resulting from the positive weight which members of a group have for each other may conflict with the consensual probability of another group, or that of a single iconoclast. There is nothing in the method to preclude this result when people refuse to assign positive weight to others. Moreover, such refusal is perfectly reasonable if one is convinced that the probability assignment of others rests on a discernible error. Sometimes the rational course of action is to refuse to assign positive weight to others and to accept conflict among members of a group and, indeed, within oneself. The purpose of weighted averaging is to exploit additional metamental information in order to resolve conflict when other more objective methods of scientific ratiocination have failed to do so. Even then one must be able to assign positive weight in the interests of truth or the method is inapplicable. Assignment of positive weight aimed at the resolution of conflict subverts this goal and is proscribed.

## VII. KNOWLEDGE

The final essays are concerned with knowledge. In these essays the role of the metamental is brought to the fore and need not be summarized here. I might, however, add a remark about the relationship between what I have written here and my earlier work in epistemology, especially in *Knowledge*.[5] In that book, I assumed that all considerations of epistemic rationality could be

[5] K. Lehrer, *Knowledge* (Oxford, 1974).



packed into the concepts of probability and competition defined in terms of probability. Justified belief was probabilistic superiority to all competitors. I now think this attempt to pack all the factors relevant to epistemic rationality into the concept of probability reflected the optimism of youth. I do not now see any way of making this work, though I still work on it. It would be a simpler account. The problem, however, is that there is more to rationality than probability. Some of the other factors of rationality might be represented by the assignment of high prior probabilities to statements having the desired factors. But already when I wrote *Knowledge* the conflict between comprehensiveness and high probability concerned me.

In these essays, though insisting on the importance of probability and continuing to hold that most factors of epistemic rationality can be represented by our assignment of prior probability, I start with a notion of rationality itself, taking it as basic, rather than attempting to reduce it to probability. One important reason for doing so which is not made explicit in the essays is the importance of metamental evaluation to rationality. The kind of justification required for knowledge involves the evaluation of lower-level information, including probabilities, to ascertain the trustworthiness of the information. Beliefs and probabilities are not by themselves sufficient for justification. My beliefs and probabilities must, in addition, be accepted as trustworthy guides to truth and, finally, I must accept that I am trustworthy in such evaluations, and I must be correct in accepting these things. Correct metamental evaluation and acceptance is essential to human knowledge because without them justification lacks an essential connection to truth.

The final chapter is about Thomas Reid. A more complete account of Reid is contained in my book, *Thomas Reid*.[6] As I interpret Reid, he held that metamental operations of the mind were essential to our general conceptions—our general conception of whiteness, for example—as well as to the attainment of knowledge. Consciousness was, for Reid, the faculty of the mind that gave us immediate knowledge of our mental operations and made metamental operations possible. For him the mind is a metamind, and his theory is superior to

[6] K. Lehrer, *Thomas Reid* (London, 1989).

many modern competitors. The chapter is a metamental conjecture about the nature of the metamind.

## VIII. SUMMARY

his introduction is a unifying reinterpretation of my essays. I d not write these essays with the purpose of exhibiting the portance of metamental operations. The significance of amental operations revealed itself as I tried to explicate the ns of freedom, rationality, and knowledge. My analyses quire further revision, but I am convinced by the course nquiry that appeal to the metamental must be salient in alyses. In casual speech we may attribute freedom, , and knowledge to metamindless beings, but the of these conceptions to such beings, though heur- ful, is not literally correct. The same thing is true of

he clock that it *says* that the time is ten past noon. l metaphor. The clock says nothing. It runs. (That hor.) There is no genuine intentionality in the who interpret the running of the clock. In the n imagine a metamindless animal or robot with plex set of responses to the external world. We free to move in one direction rather than the d be rational for it to do so, and even that it o. That would be a useful metaphor. We are metaphor for understanding other things. eing lacks the metamental power to evalu- as a result, is neither free nor rational. tal power to accept or reject information, it ell. We find it useful to ascribe intention- o the states of metamindless beings— —but this is, as Dennett suggests, only The literal truth is metamental.[7] This ribution to the understanding of the y trusty metamind for making it all

*Intentional Stance* (Cambridge, 1987).

# 1

# An Empirical Disproof of Determinism?

ACCORDING to certain philosophers, the statement that a person could have done what she did not do lacks the proper epistemic credentials. The reason why this statement has been the bone of philosophical contention is its connection with the problem of free will and determinism.

It is usually held that a person acts of her own free will only if she could have acted otherwise. However, both libertarians and determinists have had their doubts about the epistemic qualifications of such statements. For example, Ledger Wood, a determinist, maintains that the statement that a person could have done otherwise is empirically meaningless. He says 'a careful analysis of the import of the retrospective judgement, "I could have acted otherwise than I did," will, I believe, disclose it to be an empirically meaningless statement'.[1] From the other side of the issue, William James, a libertarian, argues that science, and our knowledge of what has actually happened, cannot give us the least grain of information about what it was possible for a person to have done. He says:

> Science professes to draw no conclusions but such as are based on matters of fact, things that have actually happened; but how can any amount of assurance that something actually happened give us the least grain of information as to whether another thing might or might not have happened in its place? Only facts can be proved by other facts. With things that are possibilities and not facts, facts have no concern. If we have no other evidence than the evidence of existing facts, the possibility-question must remain a mystery never to be cleared up.[2]

Thus, both Wood and James, as well as others, think that it is impossible to know empirically that a person could have done other than she did do. I wish to show that this position is

[1] L. Wood, 'The Free Will Controversy', in M. Mandelbaum, F. W. Gramlich, and A. R. Anderson (eds.), *Philosophic Problems* (New York, 1957), 308.
[2] W. James, 'The Dilemma of Determinism', in *Essays in Pragmatism* (New York, 1955), 42.

mistaken—that is, that it is possible to know empirically that a person could have done otherwise. I shall attempt to establish this first by considering in general how we know what a person can do, and then by showing that sceptical doubt concerning our knowledge of what people can do is no better grounded than sceptical doubt concerning our knowledge of the colour properties of unobserved objects. Finally, I wish to consider the implications of the possibility of such empirical knowledge for the problem of free will and determinism. I shall argue that it follows from the possibility of such knowledge that if free will and determinism are not logically consistent, then we can know empirically that the principle of determinism is false. Subsequently, I shall consider the question of the consistency of free will and determinism.

## I

I now wish to argue that we can know empirically that a person could have done otherwise.[3] A person could have done otherwise if she could have done what she did not do. Moreover, if it is true at the present time that a person can now do what she is not now doing, then, later, it will be true that she could have done something at this time which she did not do. This, of course, follows from the fact that 'could' is sometimes merely the past indicative of 'can'.[4] What I now want to argue is that we do sometimes know empirically that a person can do at a certain time what she is not then doing, and, consequently, that she could have done at that time what she did not then do. Moreover, we can obtain empirical evidence in such a way that our methods will satisfy the most rigorous standards of scientific procedure.

[3] For the purpose of this chapter, I shall assume that if a hypothesis is very highly probable with respect to some kind of empirical evidence, then it is possible to know that hypothesis empirically. Thus, I shall attempt to prove that the hypothesis that a person could have done otherwise is very highly probable with respect to some kind of empirical evidence. The line of argument I use was suggested by Richard Taylor, 'I Can', in S. Morgenbesser and J. Walsh (eds.), *Free Will* (Englewood Cliffs, NJ, 1962), 84.

[4] See J. L. Austin, 'Ifs and Cans', in J. O. Urmson and G. J. Warnock (eds.), *Philosophical Papers* (London, 1961), 163.



I shall attempt to show that we can know empirically that a person could have done what she did not do by first considering the more general question of how we ever know what people can do. It is, I suppose, obvious that there is no problem of how we know a person can do something when we see her do it. In this case, the evidence that we have for the hypothesis that a person can do something entails the hypothesis. But all that is entailed by the evidence is that the person can do what we see her do *at the time we see her do it*. It is at least logically possible that she cannot do it at any other time. Thus, when we project the hypothesis that a person can do something at some time when we do not see her do it, the empirical evidence that we have for the hypothesis will not entail the hypothesis.

The problem of our knowledge of what people can do is, therefore, primarily the problem of showing how we know that people can do certain things at those times at which we do not see them do the things in question. The solution to the problem depends upon the recognition of the fact that one fundamental way (there are others) in which we know that a person can do something at some time when we do not see her do it is by seeing her do it at some other time. However, it is not merely a matter of seeing her do something at some other time that would justify our claim to know that she can do it at the time at which we do not see her do it, but of seeing her do it when certain other epistemic conditions are also satisfied. I shall discuss four such conditions, which seem to me to be the most important. I shall call them the conditions of 'temporal propinquity', 'circumstantial variety', 'agent similarity', and 'simple frequency'.

### *Temporal propinquity*

The amount of time that has elapsed between the time at which we see a person perform an action and the time at which it is claimed that the person can perform the action is of considerable importance. For example, if I saw a man perform forty push-ups twenty years ago and have not seen him do it since, that would hardly justify my claim to know that he can do it now. On the other hand, if I saw him do it yesterday, my claim would have much greater merit. The less time that elapses between the time