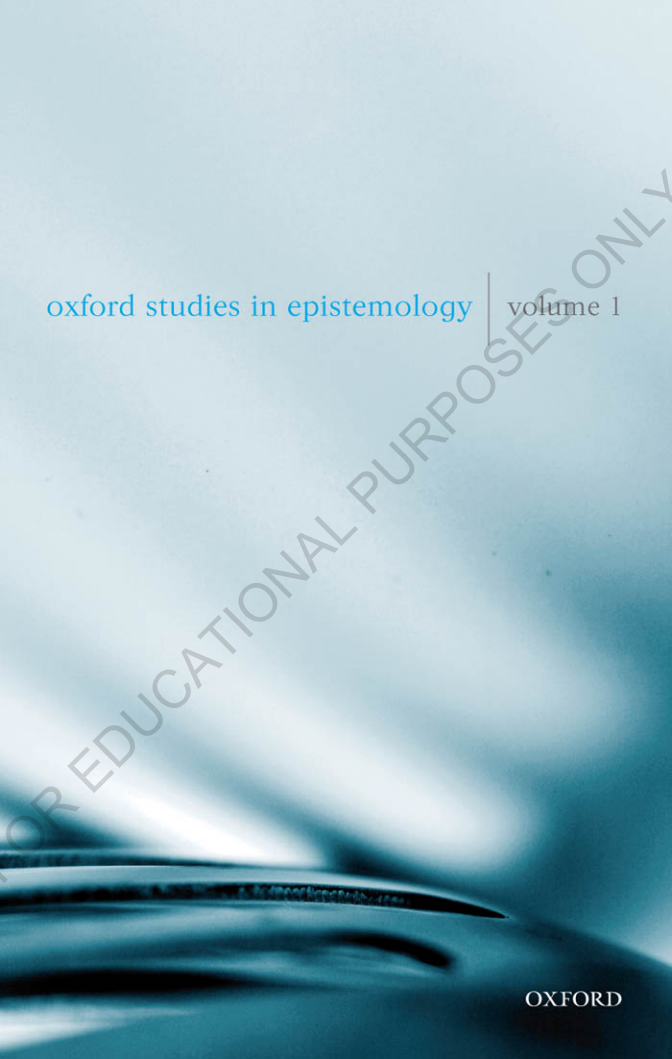

OXFORD STUDIES IN EPISTEMOLOGY

OXFORD STUDIES IN EPISTEMOLOGY

OXFORD STUDIES IN EPISTEMOLOGY

Volume 1

Edited by

Tamar Szabó Gendler and John Hawthorne

CLARENDON PRESS · OXFORD

OXFORD
UNIVERSITY PRESS
Great Clarendon Street, Oxford OX2 6DP

Oxford University Press is a department of the University of Oxford.
It furthers the University's objective of excellence in research, scholarship,
and education by publishing worldwide in

Oxford New York

Auckland Cape Town Dar es Salaam Hong Kong Karachi
Kuala Lumpur Madrid Melbourne Mexico City Nairobi
New Delhi Shanghai Taipei Toronto

With offices in

Argentina Austria Brazil Chile Czech Republic France Greece
Guatemala Hungary Italy Japan Poland Portugal Singapore
South Korea Switzerland Thailand Turkey Ukraine Vietnam

Oxford is a registered trade mark of Oxford University Press
in the UK and in certain other countries

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

First published 2005

British Library Cataloguing in Publication Data
Data available

Library of Congress Cataloging in Publication Data
Data available.

Typeset by SPI Publisher Services, Pondicherry, India
Printed in Great Britain
on acid-free paper by
Biddles Ltd, King's Lynn, Norfolk

ISBN 0-19-928589-6 978-0-19-928589-1
ISBN 0-19-928590-X (Pbk.) 978-0-19-928590-7 (Pbk.)

1 3 5 7 9 10 8 6 4 2

EDITORS' PREFACE

With this inaugural issue, *Oxford Studies in Epistemology* joins *Oxford Studies in Ancient Philosophy*, *Oxford Studies in Metaphysics* and *Oxford Studies in Early Modern Philosophy* as a regular showcase for leading work in a central area of philosophy.

Published biennially under the guidance of a distinguished editorial board, each issue will include an assortment of exemplary papers in epistemology, broadly construed. *OSE*'s mandate is far-reaching: it seeks to present not only traditional works in epistemology—essays on topics such as the nature of belief, justification, and knowledge, the status of skepticism, the nature of the a priori etc.—but also to display work that brings new perspectives to traditional epistemological questions. Among these will be essays addressing new developments in epistemology—discussions of novel approaches (such as contextualism) and recent movements (such as naturalized feminist, social, virtue and experimental epistemology)—as well as essays addressing topics in related philosophical areas. This will include work on foundational questions in decision-theory, work in confirmation theory and other branches of philosophy of science, discussions of perception, and work that examines connections between epistemology and social philosophy, including work on testimony, the ethics of belief, and the distribution of knowledge and information. Finally, the journal is committed to publishing works by thinkers in related fields whose writings bear on epistemological questions, including figures in cognitive science, computer science, and developmental, cognitive, and social psychology.

Many of these commitments are evident in the inaugural issue, which includes eleven new papers by a distinguished range of philosophers, as well as two non-philosophers—one a computer scientist, the other a cognitive and developmental psychologist. Together, the papers provide a state-of-the-art snapshot of some of the best work in epistemology going on today.

Three of the papers—Hartry Field's 'Recent Debates about the A Priori', Kit Fine's 'Our Knowledge of Mathematical Objects', and Stephen Schiffer's 'Paradox and the A Priori'—address general or

specific questions about the nature of a priori knowledge. Two other papers explore contextualism and its alternatives: John MacFarlane's 'The Assessment Sensitivity of Knowledge Attributions', and Jonathan Schaffer's 'Contrastive Knowledge.' Two other papers explore the interplay between particular philosophical theses and traditional skeptical worries: Alexander Bird's 'Abductive Knowledge and Holmesian Inference', and Brian Weatherson's 'Skepticism, Rationalism and Externalism'. The epistemological ramifications of some perplexing puzzles serve as the jumping-off point for two additional papers: James Cargile's 'The Fallacy of Epistemicism', and Joseph Halpern's 'Sleeping Beauty Reconsidered: Conditioning and Reflection in Asynchronous Systems'. Finally, two papers explore broadly social questions in epistemology: Frank Keil's 'Doubt, Deference and Deliberation: Understanding and Using the Division of Cognitive Labor', and Tom Kelly's 'The Epistemic Significance of Disagreement'.

Together, this broad-ranging set of papers—some brought to our attention by members of the editorial board, others solicited directly from authors—reveal the breadth and depth of work going on in epistemology today. It is a testament to the vibrancy of the field that assembling such an outstanding collection was a straightforward, easy and pleasant task. We have every reason to expect that future issues will provide an equally rich and diverse array of offerings.

ACKNOWLEDGEMENTS

We are grateful to the members of our editorial board for bringing to our attention a number of the papers included in this volume, and for serving as referees for all of them. Special thanks are due to Richard Fumerton, Alan Hájek, Gil Harman, and Jim Joyce, who prepared reports on particularly short notice. We are also indebted to Roald Nashi, for his excellent work as managing editor and for his preparation of the outstanding index, and to peter Momtchiloff, for his continuing support of this project.

This page intentionally left blank

CONTENTS

LIST OF FIGURES

LIST OF CONTRIBUTORS

Alexander Bird
DEPARTMENT OF PHILOSOPHY, UNIVERSITY OF BRISTOL

James Cargile
DEPARTMENT OF PHILOSOPHY, UNIVERSITY OF VIRGINIA, CHARLOTTESVILLE

Hartry Field
DEPARTMENT OF PHILOSOPHY, NEW YORK UNIVERSITY

Kit Fine
DEPARTMENT OF PHILOSOPHY, NEW YORK UNIVERSITY

Joseph Halpern
DEPARTMENT OF COMPUTER SCIENCE, CORNELL UNIVERSITY

Frank Keil
DEPARTMENT OF PSYCHOLOGY, YALE UNIVERSITY

Thomas Kelly
DEPARTMENT OF PHILOSOPHY, PRINCETON UNIVERSITY

John MacFarlane
DEPARTMENT OF PHILOSOPHY, UNIVERSITY OF CALIFORNIA, BERKELEY

This page intentionally left blank

1. Abductive Knowledge and Holmesian Inference

Alexander Bird

1. INTRODUCTION

The usual, comparative, conception of Inference to the Best Explanation (IBE) takes it to be ampliative. In this paper I propose a conception of IBE ('Holmesian inference') that takes it to be a species of eliminative induction and hence not ampliative. This avoids several problems for comparative IBE (e.g. how could it be reliable enough to generate knowledge?). My account of Holmesian inference raises the suspicion that it could never be applied, on the grounds that scientific hypotheses are inevitably underdetermined by the evidence (i.e. are inevitably ampliative). I argue that this concern may be resisted by acknowledging, as Timothy Williamson has shown, that all knowledge is evidence. This suggests an approach to resisting scepticism different from those (e.g. the reliabilist approach) that embrace fallibilism.

2. SCEPTICISM AND EVIDENCE

There is a sceptical argument that goes like this. We like to think that we are in a world not only such that we seem to see an environment of physical objects of certain sorts, but also where such objects do indeed exist and are in many respects as they seem to us to be. Such a world we may call the 'good situation'. However, says the sceptic, our evidence for thinking that we are in the good situation is just the way things seem to us to be—our subjective sense-impressions. And, says the sceptic, our evidence would be just the same if we are in the bad situation, namely, one where an evil demon or some other deceiving device causes us to have the same set of sense-impressions. Since our total evidence in the good situation is identical to what it is in the bad situation, we cannot

know that we are in the one situation rather than in the other. Hence we cannot know that there really are the physical objects there appear to be.

As I have outlined it, the sceptic's argument rests on two premises:

(EVEQ) S's evidence in the good situation is the same as in the bad situation;

(DIFF) If S is to know p then S's evidence must be different from what it would have been in any situation where $\neg p$;

from which the sceptical conclusion:

(SCEP) If S is in the good situation, S does not know that S is in the good situation;

follows immediately.

There are many ways of attempting to deal with scepticism. In the context of the current sceptical argument two strategies are apparent, corresponding to the denial of the two premises, (DIFF) and (EVEQ).

(DIFF) might be denied by arguing that differences in states of knowledge are not dependent on differences in evidence alone. They may also depend on differences in facts external to the subject's evidence. In particular, knowledge is sensitive to the nature of causal connections between the subject and the environment or the reliability of the processes by which the subject acquires his or her beliefs. These facts need not be included in the content of the subject's evidence. So two individuals may have identical sets of evidence, and identical beliefs, but the one knows something the other does not, since the two are in different environments, where one is propitious for knowing and the other is not. So, S can know that S is in the good situation since in the good situation the methods that link S's evidence and S's beliefs will be reliable, which they won't be in the bad situation. Although reliabilism is not the only instance of the approach to scepticism that denies (DIFF), I shall for convenience call this approach the 'reliabilist strategy'. Since (DIFF) is incompatible with fallibilism, as usually conceived, the term 'reliabilist strategy' may do duty for any strategy that rejects scepticism while also embracing fallibilism.[1]

To deny the sceptical conclusion, we may instead reject premise (EVEQ), as has been suggested by Timothy Williamson.[2] Quite

[1] For a discussion of what fallibilism amounts to, see Reed (2002: 143–57).

[2] Williamson (2000*b*: 625 n. 13). Note first that Williamson does not explicitly portray the sceptic's argument as proceeding from (DIFF) as well as (EVEQ). And secondly that

independent of any discussion of scepticism, there is reason to think that our evidence is just what we know (Williamson 1997):

(E = K) S's evidence is precisely what S knows.

So, if we take that on board, (EVEQ) becomes:

(EVEQ*) S's knowledge in the good situation is the same as in the bad situation.

But clearly (EVEQ*) is just what is in dispute. The sceptic's conclusion is that we can know no more in the good situation than in the bad situation, which is very little. So (EVEQ*) is no good as a premise in that argument, and likewise (EVEQ) is no good too. The sceptic just begs the question. Of course a bad argument can have two false premises and so the two diagnoses of the error in this version of scepticism are not in direct conflict. One aim of this paper is to explore the consequences and limitations of this argument as a response to scepticism that does not rely on the rejection of (DIFF). My particular focus will be upon abductive inferences. Williamson himself (2000*a*: 174) is inclined towards acceptance of (DIFF) when discussing whether the difference between relevant and irrelevant alternatives makes trouble for the sceptic: 'it is difficult not to feel sympathy for the sceptic here. If one's evidence is insufficient for the truth of one's belief, in the sense that one could falsely believe *p* with the very same evidence, then one seems to know *p* in at best a stretched and weakened sense of ''know''.' An immediate implication of Williamson's comment is that where *p* is the conclusion of a normal ampliative argument, one whose conclusion is not entailed by its premises, then that argument cannot yield knowledge of the conclusion in the proper sense of 'knowledge'. (Why I say a 'normal' ampliative argument, rather than 'any' I shall explain shortly.)

Are there then reasons for wanting to retain some version or other of (DIFF)? First, it seems that the power of scepticism would be less easily explained if we think that all its premises are at fault than if we identify just one subtle error. Secondly, (DIFF) is related to Williamson's safety condition on knowledge, that, when cases α and β are close to one another, if S knows *p* in case α then in case β S does not falsely believe

Williamson does not intend his case against scepticism to depend upon this argument. (Rather, it seems, this argument forces the sceptic to articulate a different, phenomenal, conception of evidence, that is ultimately indefensible.)

p (Williamson 2000*a*: 128). If cases where the subjects have identical evidence are classed as close to one another, then the safety principle entails (DIFF). Thirdly, given the equation (E = K) (DIFF) comes out as trivially true as it stands (although this is not the case for a diachronic version of (DIFF), as we shall see). Fourthly, (DIFF) rules out any account of knowledge of the following form, knowledge is justified true belief plus X, where the truth condition is non-redundant.[3] This is because such an account contemplates situations just like knowing with regard to justification but without truth. But (DIFF) requires situations that are like knowing but without truth to differ with respect to evidence also. Such situations will thus differ with regard to justification also (if justification is a relation to the evidence). So there are no situations of the kind JTB + X accounts envisage. Since we want to exclude such accounts anyway, thanks to Gettier's examples, a principled condition on knowledge such as (DIFF) that does so is thereby at an advantage. Finally, many people find (DIFF) compelling for the following reason. Knowing should imply epistemic responsibility, and responsibility is reasonably taken to mean appropriate sensitivity to the evidence. But if we reject the idea that there is some evidential difference between knowing and not knowing, the difference between S in the good situation who does know and S* in the bad situation who does not, has nothing to do with epistemic responsibility and has everything to do with epistemic luck. The reliabilist strategy allows that, for all else S knew, S might have been in the bad situation, but nonetheless gets to know that she is not. Many find this counter-intuitive, and it is a major part of the force behind epistemological internalism.

By retaining some difference principle we may respect these intuitions; at the same time sceptical conclusions may be resisted by rejecting the sceptic's appeal to premises akin to (EVEQ). In accordance with the Williamson strategy, (E = K) will be assumed throughout this paper.

3. RELIABILISM AND INFERENCE TO THE BEST EXPLANATION

A further reason for wanting to rest resistance to scepticism at least partly on the Williamson strategy rather than the reliabilist strategy is

[3] I am grateful to Richard Fumerton for this point.

the concern that reliabilism alone does not have sufficient resources to account for all knowledge gained by abductive inference. It is reasonable to hope that reliabilism might be able to account for knowledge gained from enumerative (or 'Humean') induction. Such is Mellor's approach, for example.[4] The difficulty is that Humean induction describes very little of the inference that takes place in science. The larger part of scientific inference is abductive. By 'abductive' inference I shall mean an inference where a central component of that inference is the fact that the inferred (purported) facts provide a putative explanation of the evidence or some part thereof. I shall treat 'Inference to the Best Explanation' (IBE) as a synonym for 'abductive inference', treating 'Inference to the Best Explanation' less as a description than as a name for a certain class of inferences that trade on the explanatory capacities of what is inferred. What exactly is involved in IBE is one of the issues to be discussed.

The application of reliabilism to IBE is thoroughly problematic. Typical accounts of IBE are what I shall call 'comparative'. They involve comparing putative explanations of some evidence with respect to their explanatory 'goodness'. Such accounts may permit or enjoin acceptance of that putative explanation which is comparatively better than all the others (hence inference to the 'best' explanation). They may possess other features, such as the requirement that the best be clearly better than its competitors and that it meet some minimum threshold of goodness. Nonetheless, the common feature is a comparison of competing possible explanations.

In order to give a reliabilist explanation of how comparative IBE can generate knowledge, two tasks must be carried out. The first is to explain what the goodness of an explanation is. The second task is to show that goodness correlates with truth. So, in fulfilling the first task, we might identify goodness with certain virtues of explanation such as simplicity, tendency to provide explanatory unification, or capacity to provide understanding (what Peter Lipton (1991: 61) calls 'loveliness')—we infer *the most virtuous of* the potential explanations. (A potential, or putative, explanation is, very roughly, something that would be the actual explanation if it were true.) Then, for such accounts to be plausible as explanations of inductive knowledge, the second task

[4] For a reliabilist account of Humean induction see Mellor (1987), who explicitly excludes Inference to the Best Explanation from his considerations.

requires showing that it is at least plausible that explanatory virtue is a reliable indicator of truth.

It is this second requirement that ought to be particularly worrying for the reliabilist. Many debates surrounding IBE question whether accounts of explanatory goodness make that goodness too subjective for it to be even possible for them to be correlated with truth. But even if we can show that simplicity, unification, loveliness, and the like are objective, that only shows that they might be correlated with truth, not that they are. For reliabilism to be a plausible account of knowledge via IBE, we should seek some evidence that there actually is such a correlation.

The problem is that such evidence is thin. Good explanations are frequently falsified and often replaced by less virtuous ones. The theory of relativity is less simple than the Newtonian mechanics it replaced, while many aspects of quantum theory are distinctly lacking in virtue and might even be regarded as explanatorily vicious (renormalization, non-locality, complementarity, and so on). The ancient theory of four elements was replaced by one with over one hundred elements. Even if the balance seemed to be restored by the discovery of the three subatomic components of atoms, it was put out of kilter by the subsequent discovery of a zoo of such particles. The Pessimistic Induction is overstated; nevertheless, it is true that good explanations are frequently falsified. They are often replaced by hypotheses that were earlier considered (or would have been considered) less virtuous ones. Thus although quantum theory is a better explanation of the current evidence than classical mechanics since the latter is falsified by current evidence, matters are reversed when we consider the old evidence, that available say in the middle to late nineteenth century. It is significant that explanatory goodness is frequently in due course overruled by the evidence.

Thus, in so far as explanatory goodness (e.g. simplicity, elegance) is independent of any specific set of evidence, we find that such goodness often decreases as theories change. Even if there is some degree of correlation between goodness and truth, that correlation is, I fear, too weak to reach a level of reliability required to generate knowledge. There are differences among reliabilists about what degree of reliability is required. Some urge that if the level of reliability is less than 100 per cent, then any such account of knowledge is liable to fall foul of Gettier-style cases. Clearly inference to the most virtuous explanation does not

achieve that level of reliability. It is also doubtful whether it meets even any lower threshold that would nonetheless be a plausible degree of reliability in a reliabilist account of knowledge. It is worth recalling that the comparison ought not be simply between pairs of competing hypotheses considered individually; if comparative IBE were to be reliable, the preferred hypothesis would also have to be more plausible than the disjunction of the remaining hypotheses.

4. DIRECT AND INDIRECT EVIDENCE

The following propositions:

(a) abductive inference is comparative IBE;
(b) abductive inference can be knowledge generating;
(c) the difference principle, (DIFF), is true;

are in tension. Although they are strictly consistent, they can be jointly true only in virtue of peculiar and unrepresentative inferences. As applied to the large majority of instances of IBE they cannot be jointly satisfied.

In a comparative IBE the various hypotheses under consideration are consistent with the evidence. Considered individually in relation to the evidence, each could be true. That is why the inference needs to take into consideration the relative goodness of each explanation (whether it be simplicity, explanatory power, etc.). Comparative IBE is ampliative—the evidence does not entail the conclusion. That a knowledge-generating inference is ampliative does not immediately entail that it doesn't satisfy (DIFF). Consider an inference made by S to a proposition such as 'S exists' or 'S has some evidence' (which we shall take to be true). (DIFF) asks us to consider S's evidence in a situation where the inferred proposition is false. Clearly S's evidence would be different (namely, none at all) in such situations. Hence any inference, ampliative or not, to such propositions will satisfy (DIFF). Hence the mere fact that comparative IBE is ampliative is insufficient to show that (a)–(c) are inconsistent. However, it is clear that the cases that allow ampliative inferences to satisfy (DIFF) are unusual and the conclusion propositions in question are not the sort that one would normally employ an IBE to ascertain. Consequently, the vast majority of actual IBEs, if they are comparative, and if they are knowledge-generating,

will not satisfy the difference principle. Correspondingly, and this is the conclusion that I will be focusing on later: if IBE is knowledge-generating, and if the difference principle is to be respected, then IBE cannot be comparative. Some other account of what is going on in (knowledge-generating) IBE must be found.

At first sight a strategy for combining comparative IBE with respect for a difference principle might be the following. The facts we normally think of as evidence for a hypothesis include the results of experiments and observations, previous theoretical conclusions, and so forth. Let us call this 'direct' evidence. In a case of comparative IBE, the direct evidence relevant to a set of hypotheses will be, principally, the facts explained by those hypotheses. It is tempting to think that direct evidence exhausts the relevant evidence, but on reflection it does not. There is also *indirect* evidence. For example, the simplicity or elegance of a hypothesis might be further evidence in its favour. In general, when employing IBE, facts such as the fact that one hypothesis is a better explanation of the (direct) evidence than its competitors can be known to the investigator and hence can be part of the investigator's total evidence.

Noting that indirect evidence exists is one way of repelling the claims of the underdetermination of theory by evidence. For if hypotheses differ, for example, in their simplicity, then there will be an evidential difference between those hypotheses. One might hope to apply this to the current problem of reconciling IBE and some difference principle as follows. If we add the indirect evidence to the direct evidence, then perhaps the total evidence the subject has for hypothesis h might be evidence that S could not have in any situation where h is false. This requires that the following should hold:

$$<\text{S has evidence } e> \text{ entails } <h>$$

where e is the total evidence, indirect evidence included. But even those who are alive to the importance of indirect evidence tend still to regard IBE as ampliative on this total evidence. It seems that one could always have the same total evidence and yet be mistaken in the conclusion that IBE presses upon us.

Of course the total evidence e might *determine* a single conclusion in the following sense. The total evidence, including the indirect evidence, can yield an unambiguous conclusion when IBE is applied to that evidence. It can be that only one conclusion is consistent with possession of the evidence and with the application of IBE. That means:

<S has evidence *e* & IBE is truth-preserving> entails <*h*>

can hold. Does this show how a difference principle may be respected? No it does not, unless the fact that IBE is truth-preserving is amongst S's evidence—difference principles concern themselves only with differences of evidence. If IBE is indeed truth-preserving it may be possible to know that it is. Yet we would not want such knowledge to be a condition of IBE's producing knowledge. For then we really would fall foul of the circularity charges that Hume urged.

5. THE CHALLENGE OF ABDUCTIVE INFERENCE

In §2 I presented a sceptical argument, proceeding from two premises (EVEQ) and (DIFF), and two strategies for resisting scepticism, corresponding to the denials of the two premises. The rejection of (DIFF) goes hand in hand with a reliabilist approach to knowledge. §3 argued that reliabilism is not a satisfactory way of explaining how IBE yields knowledge. Hence a resistance to scepticism as regards IBE ought to consider rejecting (EVEQ) and retaining (DIFF). In §2 I noted other reasons for wanting to retain (DIFF). However, §4 shows that if the difference principle is to be retained then at least those instances of IBE that are knowledge-yielding had better not be comparative. In the next section I shall present an account of knowledge-yielding IBE that is not comparative. Here I shall consider in general terms the strategy of rejecting (EVEQ) and retaining (DIFF), as applied to IBE.

Can we extend Williamson's anti-sceptical argument to abductive scepticism by rejecting (EVEQ)? There are obstacles to so doing. Consider an inference to the conclusion *h*. The most straightforward attempted application of the Williamson strategy would amount to claiming, ''The sceptic's premise (EVEQ) is false. It is false because in the good situation S will know *h* and so have different evidence (since E = K) from S* in the bad situation where S* does not know *h*.'' As a response to the sceptic this seems sadly inadequate. Scientific inferences of the sort we are interested in are those that are supposed to take us from a state of possessing evidence along with ignorance regarding some proposition to a state of knowledge concerning it. And so a response to the sceptic that compares the states of evidence *after* the inference seems to have missed the point. After all, if in the good situation the subject

does come to know h as a result of an inference, as the anti-sceptic maintains, h may perhaps then become evidence that can be used in favour of some other proposition. But it is no part of S's evidence for h itself. The abductive sceptic intends to compare S's evidence in the good situation with S's evidence in the bad situation *before* the inference. The claim is that it is possible for S's evidence to be the same in both. So what we need is a diachronic version of the sceptical argument, which will now proceed as follows.

A subject S with evidence e at t_0 reasons on the basis of e to a conclusion p at t_1 later than t_0. The good situation is one where e is true and p is true and the bad situation is one where e is true and p is false.

$(\text{EVEQ})_\text{d}$ S's evidence in the good situation at t_0 is the same as in the bad situation at t_0;

$(\text{DIFF})_\text{d}$ If S is to come know at t_1 that p, by inference from evidence possessed by S at t_0, then S's evidence must be different at t_0 from what it would have been in any situation where $\neg p$.

Now consider a case where S, who has evidence e, is considering rival and mutually inconsistent scientific hypotheses, p and q. Let it be that e entails neither p nor q, and furthermore, both p and q are consistent with S possessing evidence e. By $(\text{DIFF})_\text{d}$ S does not know that p at t_1.

Now the assumption of the diachronic version of the evidential equivalence claim, namely $(\text{EVEQ})_\text{d}$, is no longer question-begging. We apply (E = K) and then $(\text{EVEQ})_\text{d}$ tells us that S's *knowledge* in the good situation at t_0 is the same as in the bad situation at t_0. But that does not beg the question as to whether S comes to know p at the later time t_1 after the inference.

This would appear to cast doubt on the strategy for combating scepticism derived from Williamson.[5] Nonetheless, I do think that an

[5] For the response given above can be employed for any proposition that comes to be known as a result of an inference. And for most ampliative inferences it will appear that there can be a bad situation that makes $(\text{EVEQ})_\text{d}$ true. On the other hand, if the strategy is applied to propositions that are not inferred, but are believed directly (e.g. perceptual propositions), then it is not clear that the argument presented accurately characterizes the sceptic's position. For if the proposition is not inferred, an argument based on the nature and sufficiency of evidence seems an inappropriate way of spelling out the sceptical worry.

adaptation of Williamson's argument does play a valuable role, as I shall show later. Furthermore, the version of the sceptic's argument that it engenders, as we have just seen, prompts us to look for an account of IBE that permits the truth of the difference principle and the falsity of the evidential equality claim.

6. HOLMESIAN INFERENCE

We are apt to classify the inferences ascribed to Sherlock Holmes (such as identifying a criminal on the basis of the mud on a man's boot, the analysis of a cigar ash, and so on) as inductive. Yet Conan Doyle described Holmes's method as *deductive*. This appears to be a solecism.[6] On the other hand, Sir Arthur goes on to provide details of the method which allow for a reconciliation of this terminology. On several occasions Holmes tells Watson, ''Eliminate the impossible, and whatever remains, however improbable, must be the truth.''[7] That clearly is deductive. If Holmes starts by knowing that one of ten hypotheses is true and by dint of further evidence gathered in the course of his investigations comes to know of nine of them that each is false, then deduction tells him that the tenth must be true.

Holmes's method is deductive if and only two conditions are met:

(*a*) Holmes knows that one of the ten hypotheses is true;
(*b*) Holmes obtains evidence that is inconsistent with nine of the hypotheses.

As a procedure, what I shall call Holmesian inference has the following form. From initial evidence, e_i, Holmes gets to know that one of hypotheses $h_1, \ldots, h_n$ can be true. These hypotheses are explanatory hypotheses; they explain some subset e_s of the evidence. Holmes then collects additional evidence, e_a, such that e_a (given e_i) rules out $h_1, \ldots, h_{n-1}$. Hence Holmes may deduce that h_n is true. Holmesian inference requires three premises:

[6] This criticism is made e.g. by Lipton (2001).

[7] See, e.g., Conan Doyle 1953*b*: 94, 118; 1953*a*: 1089. Kitcher and Earman make favourable references to Holmes in support of eliminative induction (Kitcher 1993: 239; Earman 1992).

(i) the fact e_s has an explanation (*Determinism*);[8]
(ii) $h_1, \ldots, h_n$ are the only hypotheses that could explain e_s (*Selection*);
(iii) $h_1, \ldots, h_{n-1}$ have been falsified by the evidence (*Falsification*).

We need not see this solely in procedural terms. For we may say that at the end of the investigation Holmes's final evidence, $e_f = e_i + e_a$, entails the one hypothesis, h_n. In summary, Holmesian inference may be explained thus:

(HOLMES) S knows h by Holmesian inference from evidence e iff S deduces h from e, which includes the proposition s, where, for some $e_s \subset e$, s is the proposition that there is some explanation of e_s.

Holmesian inference is Inference to the Best Explanation. But it does not involve the selection of potential explanations according to their explanatory virtues. Knowledge by Holmesian inference is gained only when the evidence rules out all but one of the potential explanations—the best explanation is the *only* explanation of the evidence, or some part of the evidence, that is consistent with the evidence.

Now consider S who infers that p by Holmesian inference from what S knows. Since p is entailed by what S knows, it could not have been that in some other world S has the same evidence but p is false. Hence if there is a bad situation in which S infers p by Holmesian inference but p is false, S must have different evidence in that situation. Therefore (EVEQ) and $(\text{EVEQ})_d$ will not be true in this case and so this approach is consistent with Williamson's.

In what follows I shall argue that knowledge by Holmesian inference is possible. We may thereby have a view of abductive knowledge which coheres with the approach to scepticism that grants the sceptic a version of the difference principle.

The possibility of knowledge from Holmesian inference is of course controversial. It requires not only the possibility of the truth of each of the three premises *Determinism*, *Selection*, and *Falsification* but also possibility of (concurrent) knowledge of the three premises. I'll comment on each in turn. The premise *Falsification* ought to be the least

[8] I take the names of the first two assumptions, *Determinism*, and *Selection*, from von Wright 1951: 131.

controversial. On some occasions at least we are able to falsify hypotheses. If so we should sometimes be able to falsify all but one of the hypotheses in a finite set of mutually inconsistent hypotheses that includes the true hypothesis. Although the least controversial aspect of Holmesian inference, *Falsification* is not entirely uncontroversial. The (Duhem–)Quine thesis alleges that we always have the choice of avoiding a falsification by changing an auxiliary hypothesis. Knowledge from Holmesian inference requires that we know that we have falsified alternative hypotheses. Hence the Quine thesis presents a challenge to Holmesian inference if the thesis is taken to be the claim that we cannot *know* that any hypothesis is falsified by the evidence. Let us say that we do know that some observational proposition *o* and some relevant auxiliary hypothesis *a* are true, such that from *o* and *a* the falsity of the target hypothesis *h* is deducible.[9] Then the subject can know the falsity of *h*. Hence the denial that we can know that *h* has been falsified requires that we deny that the auxiliary hypothesis *a* is known (assuming knowledge of the observational proposition *o*). Consequently, the Quine thesis, if it is to be a challenge to Holmesian inference, must be regarded as stating that auxiliary hypothesis cannot be known to be true. Hence the thesis that we cannot know that a hypothesis has been falsified thus implies a general scepticism about the possibility of knowledge of (auxiliary) hypotheses. To the extent that the Quine thesis is used to question the possibility of knowledge from Holmesian inference, it begs the question by assuming a scepticism at least as strong as that which it seeks to establish.[10]

The premise *Determinism* should not be too controversial either. We need it because ruling out all but one of the potential explanations is not quite sufficient for abductive knowledge by Holmesian inference. The evidence may not require an explanation at all. So Holmesian inference

[9] Assuming E = K, we can say, as (HOLMES) requires, that the falsity of *h* is deducible from the subject's evidence. It is clearly little consolation to the supporter of the Quine thesis to argue that this is not falsification by evidence, by denying that E = K. *Falsification* by a known proposition is just as bad. This is why maintaining the Quine thesis requires denying *knowledge* of the auxiliary hypothesis. Parenthetically, I suggest that the Quine thesis is implicitly operating with a very restricted notion of evidence, for example, the phenomenal conception of evidence discussed in Williamson (2000*b*).

[10] Note that Susan Vineberg (1996) criticizes Kitcher's eliminativism (1993) on the ground that for him acceptable auxiliaries are determined by prior scientific practice. Here it is not merely prior practice that delivers the acceptable auxiliaries but rather the fact that the relevant auxiliaries are known.

requires that the subject know that there is *some* explanation or other. The subject must have some evidence that rules out the null hypothesis, that there is no explanation. In some cases the null hypothesis may indeed be true. We may take one lesson of quantum indeterminacy to be just that. But in many cases the existence of some explanation is not in doubt. The detective knows that the bullet didn't just materialize out of nothing in the brain of the victim and that the entry wound didn't just come from nowhere. Similarly, that there is some correct explanation or other for the extinction of the dinosaurs is not up for question even if the nature of that explanation is. The assumption of *Determinism* is not the false assumption that universal determinism holds, but the assumption that some part of the evidence in question has an explanation, which may be true for many cases. Van Fraassen (1980: 21) says that ''the realist will need his special extra premise that every universal regularity in nature needs an explanation, before the rule [of Inference to the Best Explanation] will make realists of us all.'' But such a strong, universal premise is not required. It is sufficient for Holmesian inference that we know of *some* facts of interest that they have an explanation.

The premise *Selection* of the Holmesian inference is the most controversial. Most philosophers of science are not willing to grant that the hypotheses that could explain some piece of evidence may be finite in number. They accept the thesis that theories are radically underdetermined by the evidence. I tackle this problem in §§ 8–11. Before that I shall compare Holmesian inference with another, related version of eliminative induction.

7. PAPINEAU ON NON-ENUMERATIVE INDUCTION

David Papineau (1993: §5.15) proposes a similar model of induction, based on Mill's methods.[11] I have characterized Holmesian inference as employing three kinds of premise: *Determinism* (the fact e_s has an explanation), *Selection* ($h_1, \ldots, h_n$ are the only hypotheses that could explain e_s), and *Falsification* ($h_1, \ldots, h_{n-1}$ have been falsified by the evidence). Papineau regards *Falsification* as the only premise in an argument that leads to the conclusion that some h_n is true. According

[11] Von Wright's account of eliminative induction (1951) also starts from a consideration of Mill.

to Papineau and von Wright *Determinism* and *Selection* are not premises of the subject's reasoning at all. This is why, on their view, inductive reasoning of this kind is ampliative and not deductive. Instead, according to Papineau, the subject is simply disposed to assert h_n once the subject knows $h_1, \ldots, h_{n-1}$ to be false. If *Determinism* and *Selection* are true (even if not known to be), then that disposition will be reliable. So overall, given knowledge of *Falsification*, the subject's belief in h_n will be reliably formed and hence will be knowledge.

Papineau's account is satisfactory only if we can regard the disposition in question as part of the process or rule whereby h_n was inferred *rather* than as masking undischarged premises of the form of (i) and (ii). This is a problem that requires a general answer. Consider a subject who reasons as follows: premises P, $P \rightarrow Q$, conclusion Q. If this subject knows P but does not know $P \rightarrow Q$ (even though $P \rightarrow Q$ is true), then we must deny knowledge to this subject of the conclusion Q. Now consider a second subject, who also does not know $P \rightarrow Q$ and who argues from the single premise P to the conclusion Q, being disposed that way. In the latter case, if Papineau's view is correct, the subject will get to know that Q. But is the second subject really entitled to the status of knowledge of Q that is denied to the first? Furthermore, the two cases are not so clearly distinct, since someone who believes that $P \rightarrow Q$ will be disposed to believe Q when they believe P (indeed on some accounts that disposition is partly constitutive of belief that $P \rightarrow Q$). So the case of believing $P \rightarrow Q$ will include the case of being disposed to infer Q from P. If the latter gives knowledge one might imagine that the former should also.

A further problem for Papineau concerns the nature of the disposition that has to exist in lieu of the premise (ii). We have many innate cognitive dispositions—many of these are typically fairly general in nature and can be explained by their evolutionary contribution to fitness. We acquire further cognitive dispositions through experience of the world. Typically these dispositions will be cognitive habits, acquired by repeated use or by repeated experience. In the sort of case we are considering neither of these apply. One of Papineau's examples is the identification of the human immuno-deficiency virus as the agent that causes AIDS. This he presents as being achieved by the elimination of other possible viruses as candidates for the agent. So the subject who thereby gets to know that HIV is the cause of AIDS must be disposed to infer that HIV causes AIDS when that subject knows that other viral

candidates have been eliminated. Such a disposition is clearly one that could not be innate. Could it be acquired by repeated explicit use or by experience of constant conjunction? Surely not—the disposition is far too specific for that. And in any case this is a disposition that must be able to exist before the subject makes the first inference of the relevant form. So where does the disposition come from? Presumably it comes from background beliefs the subject has concerning microbiology. It will be generated by beliefs such as the beliefs that only viral infections do not respond to antibiotics and that AIDS does not respond to antibiotics. It is difficult to see how a very specific disposition linking beliefs of a theoretical kind that is brought about by beliefs with theoretical content is itself very far short of being a belief, even if only a tacit belief. And if this is the case that belief or quasi-belief will look much closer to an undischarged premise than to a mere disposition, part of the process and not part of the content.

Why does Papineau want to avoid the suggestion that (i) and (ii) might be genuine premises? Presumably, I surmise, because he thinks that there are problems concerning knowledge of (i) and (ii). One ground for doubting knowledge of (ii) I address below—this is the concern that there might be too many competing potential explanations for it to be possible to know some premise that states that they are all the potential explanations there are. Papineau does not raise this problem. Presumably it would be difficult to have a disposition whose nature covered a vast range of hypotheses. Both Papineau and I require the range of hypotheses at stake to be manageable. Papineau's worry is instead a different one. Knowing that the cause of AIDS is one of viruses $v_1 \ldots v_n$ requires *knowing* that only viral diseases do not respond to antibiotics. Papineau does not deny that this could be known. But it will be known as the result of another eliminative inference of the same kind, one that asks which agents are responsible for antibiotic resistant infections. This seems to threaten some kind of regress, which is avoided by not requiring the subject to have (ii) as a premise. The philosopher can show that the disposition to assert h_n in response to knowing (iii) is a reliable one, in a naturalistic fashion, by citing the relevant facts concerning which agents are responsible for antibiotic resistant infections.

However, it is not clear that the regress in question is a vicious one. It seems that in many cases the relevant premises of the form of (ii) are known to the investigator, and in sophisticated cases, such as those just considered, they need to be. The regress is contained by noting

that for some Holmesian inferences the relevant premise (ii) need not be gained by an antecedent Holmesian inference. Later I shall consider some everyday cases where the limited range of hypotheses is knowable by simple common sense. The core of common sense may be innate knowledge or reliable intuition, which in this case constrains the range of possibilities we consider. It is clear that Holmesian inference must be supplemented by another source of knowledge of general propositions.[12]

8. THE UNDERDETERMINATION CHALLENGE TO KNOWLEDGE BY HOLMESIAN INFERENCE

Traditional approaches to inductive scepticism have ruled out anything like inductive knowledge by Holmesian inference. The common view is that inductive inference, including abductive inference, is ampliative. By definition, the conclusions of ampliative inferences are not entailed by the evidence from which they are inferred. Holmesian inference is not ampliative—the conclusion may be deduced from the three premises, *Determinism*, *Selection*, and *Falsification*. We have considered *Determinism* and *Falsification*. The claim that abductive reasoning is always ampliative typically rests on a thesis that rejects *Selection*, namely the thesis that hypotheses are underdetermined by the data. The precise nature of that thesis is itself debatable; the version I shall consider is:

(UD) There is always more than one explanatory hypothesis consistent with the evidence.

This thesis entails the claim that inductive inferences are ampliative. It is clear that knowledge by Holmesian inference is inconsistent with (UD). What reason is there to believe (UD)?

There are two considerations or kinds of consideration that are typically cited in favour of (UD). The first states that there must be such a *quantity* of distinct possible causal histories that however much evidence is gathered that rules out some of these, there will always remain more than one. The second consideration is that there be a *qualitative*

[12] One would like to show that such sources also satisfy the difference principle—or that it does not apply, e.g. by showing that the relevant knowledge is delivered by a quasi-perceptual faculty, as I suggest below.

difference between the evidence and the facts constituting the possible causal histories. If the number of possible disjoint causal histories were not too great, (UD) might still be true if the only possible evidence were of such a *kind* that it could not rule out any of these histories. I shall look at the qualitative thesis before returning to the quantitative thesis. In both cases I shall argue that the underdetermination thesis is false.[13]

9. THE QUALITATIVE THESIS OF UNDERDETERMINATION

Evidence will qualitatively underdetermine theory if evidence propositions are all of one kind and theoretical propositions are all of another kind, such that the latter are epistemically inaccessible from the former. For example, if our evidence propositions consisted solely of propositions from pure mathematics and the theoretical propositions in question concern organic chemistry, then one would not expect to be able to get knowledge of the latter by any form of inference from the former. Empiricism in one of its guises holds that our evidence propositions are always observational. A sceptical conclusion concerning the knowability of propositions concerning the unobserved may be drawn, employing the following argument:

(OBS) All evidence is observational;
(INF) From observational premises only observational conclusions may be rationally inferred;[14]

therefore

(SCEP) Only observational propositions can be known.

Many of the failings of empiricism have been adequately addressed elsewhere. Here I shall add to those arguments one that derives

[13] Other supporters of eliminative induction accept the underdetermination thesis but hold that we are able to pare down the infinite range of logically possible hypotheses to a manageably finite number. We have seen Papineau's appeal to a disposition that fulfils this function. For Kitcher (1993: 248) it is prior scientific practice that performs this function.

[14] To my mind this premise is itself highly questionable. It would be a little less questionable if we were to replace 'inferred' by 'deduced'. If we do so, then the sceptical conclusion follows only if we add a further premise to the effect that only deductive inferences can lead to knowledge (which would be acceptable to a supporter of (DIFF)). However, my strategy here is not to take issue with (INF) but with (OBS), for which reason I am happy for (INF) to be as strong as any sceptic could wish for.

from Williamson's strategy as outlined at the opening of this paper. We apply Williamson's equation of evidence with knowledge, (E = K), to the premise (OBS). This yields:

(OBS*) All knowledge is observational;

which is equivalent to the sceptical conclusion. Hence the very limitation of evidence to observational propositions is to assume what the sceptic is seeking to prove. The sceptical argument of qualitative underdetermination is question-begging.

This strategy generalizes to any argument of an analogous nature that appeals to an underdetermination of theory by evidence on the grounds of a difference in kind between evidence propositions and theory propositions. A qualitative underdetermination argument may have the following form:

(EV) All evidence is of kind *K*;
(INF) From premises of kind *K* only conclusions of kind *K* may be rationally inferred;
therefore
(SCEP) Propositions of a kind other than *K* cannot be known.

Applying (E = K) to (EV) gives (SCEP) immediately. Any such argument will be question-begging.

10. THE QUANTITATIVE THESIS OF UNDERDETERMINATION

This consideration in support of (UD) is that the range of possible explanations is too large for any possible gathering of evidence to pare that range down to one. While theory actually is often underdetermined thanks to insufficient evidence, it needs substantial argument to show that it always must be. While the qualitative consideration in favour of (UD) employed some principled (but flawed) arguments, this quantitative consideration is supported less by positive argument than by a sense, bred by familiarity with sceptical scenarios, that however much evidence one adduces in favour of an hypothesis, one could always imagine some competing hypothesis consistent with the same evidence.

A time-slice through the causal history of an explanandum constitutes in Hempel's terms a complete (rather than an elliptical) explanation.

Holmesian inquiry need not be expected to determine one complete explanation, but may hope to show that some possible fact of interest (a hitherto unknown, typically explanatory fact) is contained in every complete potential explanation that is consistent with the evidence. (UD) is true and rules this out only if, whatever one's evidence, for any fact there is always some complete potential explanation that does not include that fact. Let us suppose in accordance with the conclusions of the last section that there is no restriction on the kind of evidence available. And let us suppose, reasonably enough, that there is no finite upper limit on the quantity of evidence we may collect. (UD) then requires that for any possible fact of interest, F, there be an infinite number of distinct complete potential explanations of some explanandum not containing F, all of which are consistent with the evidence.

This is a strong requirement that we tend in everyday circumstances to think is false. A simple case is that where we know that some fact, F, exists but want to know whether it is part of the *causal* history of some explanandum, E. Mill's method of difference tells us to consider a parallel case, the foil, that has the same total history as E, except for the absence of F. If in the parallel case there is no parallel to E itself, we may deduce that either E has no explanation or that F is part of its causal history. Hence we may know the latter, if true, given that we know the premise *Determinism* (that E has some explanation). The foil, in effect, is a way of excluding all potential explanations that do not include F.

In other cases we may not know whether the possible fact in question occurred at all, and so Mill's method is not applicable. Even so, there are certainly occasions when we naturally think that one of only a finite number of potential explanations must be true. Detective stories of a kind less sophisticated than Conan Doyle's trade on this fact. Often they will involve a murder in an isolated country house or inaccessible island, where there are only so many guests, butlers, and detectives. The number of potential murderers is finite and even if we consider the possibility of more than one murderer, there is still only a finite number of mutually exclusive hypotheses concerning the identities of those responsible. Consider also the following more mundane example. I pour milk into a tall glass and leave it on the kitchen table. I leave the kitchen for a few moments and then hear a crash. I return to the kitchen to see a broken glass on the floor with milk spilt on the floor and table. The cat is standing on the table licking at the pool of milk. Let us now look for the explanation of the spilt milk. The obvious potential

explanation is that the cat knocked over the glass, which rolled off the table onto the floor. Holmesian inference says that we know that this is what happened only if there is no other explanation consistent with the evidence. Is there no other explanation here? Perhaps some other large object hit the glass and knocked it over—perhaps the dog or a cookery book falling from a shelf. But the dog is outside. So that hypothesis is falsified. All heavy objects like cookery books are in their place. Nothing like that is found on the table or floor. Furthermore the table is in the centre of the kitchen nowhere near the path of a falling object. The ceiling, by the way, is intact too. So that class of hypotheses is falsified as well. Perhaps something shook the table. Might it have been the cat? No, since the table is a heavy oak table, too heavy for the cat to move or judder. Perhaps an earthquake? But I know we don't have powerful earthquakes in South-West England, and even if we did, I would have felt one powerful enough to shake the table and knock the glass over, which I did not.

Just as in the typical country house murder, the simple example just given strongly suggests that what I know can rule out all explanations bar one. And for that reason, I know it was the cat that knocked the glass over. In contrast, consider for a moment an additional surmise, that the cat knocked the glass over as a result of trying to get the milk inside. It seems a pretty good explanation. There are other explanations, that the cat knocked the glass with its tail or just by sitting down on the table too close to the glass. These are not such good explanations. It may be that what I know makes such explanations unlikely, and may even justify my surmise. But, I suggest, if I know nothing that rules out these other explanations, then my surmise will not amount to knowledge.

11. UNDERDETERMINATION—SCEPTICAL SCENARIOS

An objector might try more abstruse explanations of the broken glass and spilt milk. Perhaps an evil demon is playing a trick on me; the cat is innocent and the demon pushed the glass over. Does anything I know rule that hypothesis out? Yes, what I know may very well rule out that hypothesis. Most obviously I might know that evil demons of that sort do not exist. Throughout science there is no evidence that such things exist. And, given the exhaustive nature of science, there would be such evidence if they did exist. Since there isn't, we know they don't. Perhaps

some might regard this as too optimistic; perhaps there is a demon who has decided to wait until precisely this moment to engage in interference with the world.

There may be enough evidence to rule out this demon hypothesis. Let us assume I know some basic physics—folk physics may be enough. If I know it, then, thanks to (E = K), it is part of my evidence. To knock a glass of milk over requires the transference of a certain minimum amount of energy to the glass. To generate that energy there must be a force acting over a certain distance. The dimensions of the kitchen and the fact that the doors and window are closed put an upper limit on the distance and so a lower limit on the force. But I know that there is nothing in the kitchen to generate a sufficient force. For instance, the energy could have been transferred by a largish object (the size of a cat) moving slowly. But no large object other than the cat was found in the kitchen. Perhaps the evil demon and its tools are invisible. But that too would require a violation of reasonably basic physical truths which I know. Similarly, momentum might have been transferred to the glass by a small object moving at greater speed (a squash ball, for instance, which has escaped my notice). But there is nothing to have accelerated the ball sufficiently. Such considerations are a convoluted way of illustrating the point that more bizarre explanations and sceptical alternatives just as much as the plausible potential explanations are ruled out by facts I know.

Furthermore, the existence of such a demon, even if not active, may be inconsistent with other knowledge that constitutes my evidence. On one view, for instance, knowledge requires reliability of nomic rather than statistical strength. The demon's existence therefore would render the mechanism of perceptual belief formation unreliable and so would rule out much perceptual knowledge. Since we are allowing perceptual knowledge, we may argue by modus tollens that such a demon does not exist. That is, if my evidence includes any perceptual evidence, then the demon explanation of the glass falling over is ruled out. Of course, the critic may suggest that the reliabilist's condition on knowledge is too strong, but if conditions on knowledge are weakened, then it may be that knowledge of the demon's non-existence may be obtained directly.

The sceptically minded critic might be tempted to point out that I might not know folk physics and other things I claimed to know in the last paragraph. Indeed those claims might all be false. Perhaps what we take to be the laws of physics are mere regularities that exist only

thanks to the will of the demon and which may be violated by the demon at will. Such an objection misses the point. Certainly, we might live in a world in which the things I have cited as evidence are not known and so are not evidence. In such a world I will not be able to get to know that the cat spilled the milk. But there are also worlds in which I do know these things, and so these things are evidence. The challenge we are currently considering is whether the sorts of thing we normally count as evidence, if they are evidence, could ever be enough to rule out all explanatory hypotheses but one. The claim I have argued for is the conditional one: *if* what we normally take to be evidence is evidence, *then* we can gain knowledge by Holmesian inference. A form of scepticism that argues that what we normally take to be evidence is not in fact known, does not undermine this conditional claim. A die-hard sceptic will regard the antecedent as excessively strong. But that reaction just adds to the plausibility of the conditional. Precisely because the antecedent is inconsistent with sceptical hypotheses, those hypotheses are ruled out by the assumption of the truth of the antecedent. Our 'normal' evidence not only rules out ordinary hypotheses; rather, the very possibility of normal evidence is incompatible with sceptical hypotheses also. Consequently those sceptical hypotheses do not support (UD) and do not undermine the possibility of Holmesian inference.

12. ATTENUATED VERSIONS OF HOLMESIAN INFERENCE

Even so, there may be some residual concern that without going as far as considering demon-laden sceptical scenarios there may nonetheless remain abstruse and unusual explanatory hypotheses that the investigator has not considered and which have not been ruled out by direct refutation. While I am not sure that this must always be the case, it is worth mentioning a possible response. This draws upon Peter Lipton's account of IBE (1991: 61). According to Lipton the investigation and ranking of hypotheses takes place only concerning live, plausible options. He thinks that there are indeed many other potential explanations out there, but these never get consciously considered. This does not matter; these are explanations that had they been considered would have been given a very low ranking. It might be that an intelligent and experienced investigator is reliably disposed to ignore only poor

potential explanations. Any potential explanation that would be a reasonably good explanation does get considered. To Lipton's picture we could add a reliabilist coda. It might be that IBE as traditionally considered, as a ranking of hypotheses according to virtue, is reliable at least as far as excluding bad explanations. I argued that we might have reason to think that traditional IBE is not sufficiently reliable to give us knowledge of its favoured hypotheses. But it may be reliable enough at giving us knowledge that very bad explanations are false. If that is the case, an investigator who ignored such explanations might not have his reliability impugned by that fact, so long as his ignoring them is reliably related to their being very bad explanations.

This attenuated version of Holmesian inference can account for knowledge by reliabilist criteria. But does it not give up on $(\text{DIFF})_{\text{d}}$ in the process? I am not sure that it does. As Lipton sees it, IBE is a two-step process. The first step in the process is the one of thinking up and selecting the plausible potential explanations and filtering out the implausible ones. The second step is that of ranking and selecting among them. (According to the Holmesian inference, the second step is that of eliminating all but one.) Assuming reliability in the first step, the investigator is in a position to know, before embarking on the second step, that the actual explanation lies among the potential explanations now under active conscious consideration; he knows that all the others are false. Now the first step need not be thought of as a process of *inference* at all. Ignoring the very bad potential explanations is a quasi-intuitive skill; it is the product of experience not of ratiocination (Papineau's remarks about dispositions to believe can apply here). Hence the knowledge that the actual explanation is among those under consideration can be seen as akin to acquiring new evidence by observation. No process of ampliative inference was used. $(\text{DIFF})_{\text{d}}$ is thereby respected.

For this to work, it had better be that the actual explanation is indeed not among the explanations not considered. That it very frequently will be amounts to the concern that Lipton calls 'underconsideration'. I believe that Lipton's answer to that problem is right (Lipton 1993). He argues that we could not have even a reliable ordering of considered hypotheses unless our background theories, used to assist in this ordering, were true or approximately true. If the background theories can be true that shows that in their cases we *did* consider the true hypothesis.

Hence, so long as we can reliably rank our hypotheses for goodness, underconsideration cannot be endemic. In my view IBE does not work simply by ranking hypothesis, instead it works by refuting them (all but one). But the argument is the same. Refutation will depend on auxiliary hypotheses. For refutation to be possible the auxiliary hypothesis must be true or approximately so. Hence if refutation of some hypotheses is possible, it cannot be that the true hypothesis is never among those considered.

Another attenuation one might make to accommodate abstruse and implausible hypotheses that are nonetheless consistent with the subject's evidence amounts to a weakening of the difference principle. We have been working with the following difference principle:

(DIFF) If S is to know p then S's evidence must be different from what it would have been in any situation where $\neg p$.

This regards as relevant situations that may be vastly unlike the actual one. However, a plausible line of epistemological thought suggests that it is a philosophical illusion that knowledge is sensitive to distant possibilities. If we think of knowledge in terms of (the denial) of luck or in terms of safety, we are not obliged to focus on any more than nearby possibilities. In which case we can employ the weaker difference principle:

($\text{DIFF}_{\text{weak}}$) If S is to know p then S's evidence must be different from what it would have been in any nearby situation where $\neg p$.

Such a conception of knowledge renders Williamson's anti-sceptical strategy redundant—but we've already seen reason to doubt its efficacy. More importantly, even this weaker difference principle gives us reason to prefer Holmesian inference to comparative IBE. The latter is insufficiently reliable for knowledge even when we restrict our attention to nearby worlds. That unreliability is revealed by, for example, the history of science rather than consideration of abstruse possibilities. Restricting our attention to nearby possibilities will allow us, in some cases at least, to ignore abstruse hypotheses and to hope for evidence that will falsify all remaining hypotheses but one. In the next section I will say more about the relationship between Holmesian inference and comparative IBE.

13. INFERENCE TO THE BEST EXPLANATION RECONSIDERED

The Holmesian picture of scientific inference allows for a greatly simplified understanding of Inference to the Best Explanation. Above I stated that there are two puzzles for accounts of Inference to the Best Explanation that regard such abductive knowledge as resulting from selecting an otherwise radically underdetermined theory on the basis of explanatory goodness. The puzzles were first to give an account of this goodness, and secondly to demonstrate a correlation between it and truth. We can now see that Inference to the Best Explanation considered as Holmesian inference may eliminate these difficulties. In the first place we may construe goodness simply as not being falsified by the evidence when other hypotheses are; the best explanation will be the only one that could be true. Secondly, Holmesian inference guarantees truth when arguing from known premises.

The description just given concerns the circumstance where all but one of the hypotheses actually entertained have been refuted. The Holmesian deduction permits us to infer the truth of the remaining hypothesis. When that is the case goodness entails truth. However, we may frequently want to make inferences where we have not yet refuted all hypotheses but one. Take a simple case where two hypotheses remain unrefuted, h_1 and h_2. Let it also be the case that there is some proposition p entailed by h_1 and denied by h_2. It might be that we are not in a position to ascertain the truth or falsity of p; nonetheless, independent knowledge tells us that p is highly unlikely. Thus we may not be in a position to know that h_1 is false, and h_2 true; but we can know that this has a high probability. Thus the structure of Holmesian inference allows room for probabilistic reasoning concerning hypotheses. Similarly, someone might have evidence which while it does not rule out a hypothesis, justifies the belief that it is ruled out. Correspondingly one might come to a justified belief, by Holmesian inference, that a hypothesis is true, even if that belief does not amount to knowledge.

The probabilistic use of Holmesian inference may be available when propositions such as p in the above are of a kind previously known to us. But where we are dealing with novel, unobservable, or previously unobserved circumstances, the background information required to give us a reasoned assessment of the likelihood of refuting facts may

be unavailable. As discussed, proponents of Inference to Best Explanation have appealed to considerations of explanatory goodness of a kind that do not entail truth in order to assess the chances of hypothesis being true. Let us call this sort of goodness 'virtue'. Virtues include features such as simplicity, unification, and Lipton's 'loveliness'. Our concern was that virtues correlate with truth too weakly to provide knowledge. It may be that in extreme cases, where one explanation is very much more virtuous than its competitors, we can know, by reliabilist criteria at least, that the loveliest explanation is true, in the absence of evidence refuting all competitors. In other cases we have at best only a justified epistemic preference weaker than knowledge. Recalling the distinction between direct evidence for a hypothesis, which is evidence, for example, that refutes a rival (or entails the hypothesis), and indirect evidence, which is evidence of explanatory virtue, then the relationship between Holmesian inference and inference to the most virtuous explanation (e.g. as understood by Lipton) may be characterized as follows. Clearly both sorts of evidence may be relevant. But direct evidence takes priority. Once refuted, a hypothesis is out of consideration, however explanatorily lovely. Indirect evidence is therefore relevant only amongst unrefuted hypotheses. Holmesian inference corresponds to the case where there is sufficient direct evidence that indirect evidence is not needed. These are the cases that yield knowledge of hypotheses. Inferences to the most virtuous explanation, such as Lipton's inference to the loveliest explanation, concern cases where there is insufficient direct evidence. These cases may yield a rational preference but typically not knowledge.

It is not to give in to scepticism to accept that many well-favoured hypotheses do not yet constitute knowledge. It is in the nature of scientific enquiry that theories concerning a subject are proposed well in advance of there being sufficient evidence to decide their truth. Early on it will be important to gather evidence that will assist in coming to such a decision. At the same time a fruitful and promising theory will become the basis of research into yet further hypotheses. Scientists, both as individuals and as a community, will need to decide whether their efforts should be put into confirming the basic theory or into research that takes that theory as a given. They will need to take a calculated gamble. The new research will be more exciting and provide greater opportunities for personal satisfaction and professional advancement. At the same time, there is the danger that they will be pursuing a

wild goose, should the basic theory turn out to be false. There is no reason to suppose that the point at which scientists decide to accept the gamble coincides with the point at which evidence is sufficient for the theory to be known, especially when we consider that the scientists will typically not know the point at which they come to know the theory. It seems a reasonable speculation that scientists will accept the gamble well *before* the theory becomes known. Even if all scientists in the field accept the gamble and give up the search for further confirming evidence, it is likely that the new research generates results that themselves are confirming of the theory, ruling out alternatives.

14. FALSIFICATIONISM

Holmesian inference may explain the surprising readiness of professional scientists to endorse Popperian falsificationism. That endorsement is surprising because, as so many have pointed out, falsificationism explicitly denies the possibility of inductive knowledge, at least as 'knowledge' is normally understood (as being factive, entailing truth), and implicitly entails an even greater degree of irrationalism about science than this. It cannot be that scientists are attracted by the whole package—it must be some label or slogan on the wrapping. There are two features of Popperianism that are attractive to scientists and which are shared by the model of Holmesian inference. The first is the simple idea that science proceeds by falsifying hypotheses. For scientists good and interesting evidence is not a pile of confirming instances. Rather it is evidence which might refute some live option. Only Popper's philosophy of science emphasized the falsification of theories. But Holmesian inference emphasizes falsification too, for it is only by the refutation of rival hypotheses that a given hypothesis can get to be known.

The second feature is that discussed in the preceding section. On the Holmesian model, scientific knowledge is available, but not too easily. Science may be in a position where our favoured theories are not known and so, in that sense, are tentative. It may be that Popper's scepticism struck a chord here too. Scientists are indeed reluctant to claim for their best theories the status of knowledge; instead they will say things like 'this is currently our best model'. For Popper himself tentativeness and

the lack of factive knowledge are features of *all* scientific belief. Which is surely absurd. There is no reason to think of the genetic theory of inheritance and the double-helix account of DNA as mere models. We shouldn't deny knowledge of the electro-magnetic nature of light or the atomic constitution of matter. But in fields still being explored, underdetermination may be the case and tentativeness will be the appropriate attitude to take. Furthermore, it may be added, even when evidence does rule out all but one hypothesis, scientists may not immediately be aware of the fact. That is, they may know a hypothesis is true, or be in a position to know it, while still being far from knowing that they know. Hence they may not be entitled to assert that they know (Williamson 1996).

15. CONCLUSION

The argument I have given rests in large part on accepting Williamson's equation of evidence with knowledge. Relative to some conceptions of evidence (e.g. that evidence is what one believes) this equation is restrictive. Clearly such a conception is inadequate, as Williamson shows. In any case, if it were true, there would be no reason to suppose that the underdetermination thesis is true. Trivially, a unique hypothesis can easily be determined by what I believe—so long as I am happy to believe enough. At the same time, Williamson's equation is reasonably generous. For example, it does not require certainty for evidence. Nor does it permit, as we have seen, a limitation of evidence to observational knowledge alone. It does permit inferred knowledge to be evidence.[15] I have thus assumed throughout that the propositions that we generally take ourselves to know can be regarded as among our evidence.

Armed with this reasonably rich stock of evidence we can tackle the alleged sceptical problem of inductive knowledge that argues that a difference condition on knowledge cannot be met thanks to radical underdetermination of theory by evidence. Denying the ubiquity of underdetermination of theory by evidence enables us to assert the

[15] Strictly, Williamson's arguments for E = K do not rule out a limitation of evidence to non-inferential knowledge. This lacuna is readily filled—see Bird (2004).

possibility of knowledge by Holmesian inference. While many favoured theories at the leading edge of science probably are currently underdetermined by data, that is a scientific not a philosophical problem. Some hypotheses are not underdetermined by evidence and so can be known by Holmesian inference while adhering to some version of the difference principle.[16]

REFERENCES

Bird, A. (2004) 'Is Evidence Non-Inferential?', *Philosophical Quarterly*, 54: 252–65.

Conan Doyle, A. (1953*a*) 'The Adventure of the Bruce-Partington Plans', *The Complete Sherlock Holmes*, ii (New York: Doubleday).

—— (1953*b*) 'The Sign of Four', *The Complete Sherlock Holmes*, i (New York: Doubleday).

Earman, J. (1992) *Bayes or Bust* (Cambridge, Mass.: MIT Press).

Kitcher, P. (1993) *The Advancement of Science* (New York: Oxford University Press).

Lipton, P. (1991) *Inference to the Best Explanation* (London: Routledge).

—— (1993) 'Is the Best Good Enough?', *Proceedings of the Aristotelian Society*, 93: 89–104; reprinted in D. Papineau (ed.), *Philosophy of Science* (Oxford: Oxford University Press, 1996).

—— (2001) 'Is Explanation a Guide to Inference? A Reply to Wesley Salmon', in G. Hon and S. Rakover (eds.), *Explanation: Theoretical Approaches and Applications* (Dordrecht: Kluwer), 93–120.

Mellor, D. H. (1987) 'The Warrant of Induction', in D. H. Mellor, *Matters of Metaphysics* (Cambridge: Cambridge University Press, 1991).

Papineau, D. (1993) *Philosophical Naturalism* (Oxford: Blackwell).

Reed, B. (2002) 'How to Think about Fallibilism', *Philosophical Studies*, 107: 143–57.

Van Fraassen, B. (1980) *The Scientific Image* (Oxford: Clarendon Press).

Vineberg, S. (1996) 'Eliminative Induction and Bayesian Confirmation Theory', *Canadian Journal of Philosophy*, 26: 257–66.

Von Wright, G. (1951) *A Treatise on Induction and Probability* (London: Routledge & Kegan Paul).

[16] I am grateful to Richard Fumerton and Timothy Williamson for helpful comments, as well as audiences at Dartmouth College, and the Universities of Nottingham, Cambridge, Reading, and Lund.

Williamson, T. (1996) 'Knowing and Asserting', *Philosophical Review*, 105: 489–523.
——(1997) 'Knowledge as Evidence', *Mind*, 106: 717–41.
——(2000*a*) *Knowledge and its Limits* (Oxford: Oxford University Press).
——(2000*b*) 'Scepticism and Evidence', *Philosophy and Phenomenological Research*, 60: 625.

This page intentionally left blank

2. The Fallacy of Epistemicism

James Cargile

1. A PROBLEM

There exists an infinite series, the natural numbers, such that any property which is possessed by 0 and by the successor of any number that has it, is possessed by all the numbers. This is equivalent to the Least Number Principle (LNP): for all m and n, $m < n$, if m has a property F and n does not, then there is an i, $m \leq i \leq n$, such that i is F and $i + 1$ is not—a ''breakpoint'' in the series for F. Some of us believe we can see that this is true by the light of pure reason. The ancient Sophists made light of this 'light'. It seems that LNP has instances that are absurd. 1 may name an obvious pile of sand. Removing one grain seems to make no difference to being a pile. There must of course be some difference if we are to have separate entries. But you cannot have a pile of sand if one grain removal (without also spreading out the rest or some such trick) results in not having such a pile. And yet n removals leave, not a pile, but only one grain. LNP seems to entail some removal must have made a difference after all, contrary to ''common sense''.

Defenders of pure reason could respond by challenging the assumption of the example. We might get to three grains on the flat with a fourth on the top in a pyramid such that close observers would hold that the four-grain pyramid was a pile, while the result of removing the top grain is not.[1] We would have located a breakpoint with no trouble. This is reason to modify the example. Poverty of examples clouds the view, but so do sketchy surveys. These considerations can be difficult to balance. Philosophers rely on imaginary examples, unlike the scientists with their labs. This may be disputed. We can always watch a pot of water come to a boil, etc. But even with an actual series, we do not

[1] This is pointed out by Hart (1992: 3).

proceed like scientists. This may well be disputed, but not by the scientists. Some philosophers regard the imagination as a threat to discipline. It is, but one we must confront. We cannot hide our heads in the sand.

Imaginary examples offer the compensation of an unlimited budget. We dig a hole 10 feet deep and 30 in diameter in level ground and fill it with sand grains identical in size and shape, and smooth this level with the adjacent ground. No one (who counts here) would say, ''That is a pile of sand''. We remove grains one at a time in such a way that no other grains are disturbed by the removal, making a perfectly exact copy of the (w)hole setup for each step. We arrive at a roughly conical figure, about 6 feet in diameter at the base and 4 feet in height with its top level with the surrounding ground, such that everyone would say, ''That is a pile of sand''. It seems this could be done in such a way that adjacent entries are not distinguishable by observation methods which should be adequate for judging pilehood.

Having reached a ''pile'' by *m* removals, we can go to a ''non-pile'' with *n* more removals, getting to a roughly flat surface in the hole at a depth of 4 feet. Every one of these entries is stored indoors. (A sand-pile out in the rain might freeze solid in winter and be pried up and rolled away. We might deny that what was rolling was a pile of sand, just as we might think that what the prisoner has in the corner is a pile of rags until we learn it is one escape rope.) We can also have these entries very readily available for observation.[2] The numbers will strain human powers of observation, but that is just a feature of sorites.

We have arranged a case in which it seems that 1 names something with the property of being a hole with one pile of sand in the middle and *n* names something without that property, with no relevant difference between adjacent steps. That would be a counterexample to LNP. Case 1 provides occasion for ''That is a pile'' and case *n* for ''That is not a pile'', but there is some trouble as to what is being indicated by ''that''. If it is the (w)hole array some might say case 1 is not a pile either. This can be made clear. The subjects in our series are numbers, but ''pile'' also requires some subject.

Some will say, ''Pure reason tells us that if 1 has the property of being a pile (case) and *n* does not, then there is a breakpoint *i*''. They will explain, ''That is an instantiation of LNP with the property of being a

[2] This setup should be easy to arrange in a building of only 200 square miles.

pile as the value of *F*''. This can distort the question at the outset. Consider, ''Pure reason tells us that if 1 has the property of being ardufafardable and *n* does not, then there is a breakpoint *i*''. It is highly probable that that is not, in my language community, an instantiation of LNP. It is a dummy instance with ''ardufafardable'' simulating a property-expressing term. This cannot be repaired with such a preface as ''If 'ardufafardable' expresses a property, then, . . .''. Some would recommend that we adjourn to a formal system where some semblance of LNP appears either as an axiom, or in first order number theory, an axiom scheme, and the idea of an instantiation can be made perfectly clear. There are no sorites problems there.

We will continue in the dangerous language, but warily. ''Pile'' is clearly a working term of our language in a way that ''ardufafardable'' is not. What that means for our problem remains to be seen. So, suppose that there were a breakpoint in our imaginary series. In that case, an angel could tell us the number and we could be presented with *i*, and $i + 1$ and asked to tell which one is the pile. Since we could not distinguish the two (without peeking at their numbers) this would be quite a mystery. We would of course be respectful of the angel and would bear in mind how tricky oracular riddles can be and how they can help reduce hubris. An extensive search to determine which is a pile might help our souls. But the solution might be to burst out laughing.

There could be indefinitely many other routes from pile to non-pile or the reverse. We might seem to see a huge pile of sand looming. Is this a pile we see before us? Lurching forward, we find it is nothing but air, a strange sand-colored cloud. It gets thicker so that we have to fight our way out. Eventually ''it'' settles into a routine big pile of sand, leaving us to enquire as to just when it qualified. Here it is not number of grains but density. This series is difficult to review, the entries lost with the passage of time and hard to number. We could start with a large pile of sod which is gradually transformed into a sod shanty. With this series, our judgment about an entry would and should be influenced by the results obtainable by families moving in and setting up house. That is too variable and takes too long to keep track of.

We tried to design our series so as to preclude ''conceptual tricks''. But if we actually encountered this series, no one of us could have built it. Suppose the angel tells us the break is at entry 37 and this seems impossible, obviously in pile territory in the series. Looking at three entries we are assured are 36, 37 and 38, we cannot detect any difference

and agree they are all piles. But when allowed to climb into the exhibits, we find that 38 is like sandstone, with even the surface grains difficult to dislodge. If adhesion between grains was steadily increasing, this could be a factor we neglected. Our piety might be strained by a temptation to think that the angel had cheated us. ''The only difference allowed is number of grains!'' And all consequences of that—adding a grain might increase the overall adhesion. Such tricks are common in popular ''brain-twisters''. By working in our imagination, we may fall into the illusion we can rule out tricks we fail to imagine.

Alternately, it might be scientists who come forward to tell us they can find the number with their pile detector. Starting at any point in the series, a man with the detector can always find the same breakpoint. Such success is trivially imaginable and it is imaginable that an explanation is offered which is about as comprehensible as many summaries in popular science. We would need to bear in mind that that is not imagining an explanation. If we did *imagine* one, we could claim it for philosophy.

In our sorites, imagination shows (I) 1 does not look like *n*, (II) every *i* looks like $i + 1$. Some will conclude (III) ''the relation 'x looks like y' is not transitive''. That would seem to prove by the definition of transitivity (T) that (IV) there are entries x, y, z such that x looks like y, and y like z, but x does not look like z. What if we cannot find such a triple? Another struggle between reason (I&II; therefore, III&IV) and imagination (I, II, $\sim$IV)? Our sorities threatens (T) just as it does LNP—not at all; except for those who assume that ''x looks like y'' expresses a relation, just as they presume that ''x is a pile'' expresses a property. But this can be hard to make clear. How can those forms have meaning but not express universals? For nominalists, who reject language-transcendent universals, that is very easy to answer. For Platonists it is harder.

2. LOCKE'S ANSWER

Locke offers an interesting response to this problem.

> I demand, what are the alterations may or may not be in a horse or lead, without making either of them to be of another species? In determining the species of things by our abstract ideas, this is easy to resolve. But if anyone will regulate himself herein by supposed real essences, he will, I suppose, be at a loss: and he

will never be able to know when anything precisely ceases to be of the species of a horse or lead. (Locke 1959: 24)

. . . where the nominal essence is kept to as the boundary of each species, and men extend the application of any general term no farther than to the particular things in which the complex idea it stands for is to be found, there they can be in no danger to mistake the bounds of each species, nor can be in doubt, on this account, whether any propositions be true or no. (Locke 1959: 253)

Translating freely, Locke is warning someone who says of case 1, ''That has the property of being a pile'', and of case *n*, ''That does not have the property'', that he is in for embarrassment in a run through our series. But if he said only, ''I feel like calling that case 'pile' '' or ''I don't feel like calling that case 'pile' '', then he is not vulnerable. LNP guarantees something the first claimant will not be able to find: a last entry with the property of being a pile.[3] What it guarantees for the second claimant is a cinch to find: a last case he calls ''pile''. If we ask the ''real essence'' man why he stopped at case *i*, calling it a pile while refusing to call $i + 1$ a pile, he could (perhaps) feel embarrassed if he implied a claim to recognitional ability. If we ask the ''nominal essence'' man, he can say, ''I don't know; I just don't feel like it''. He has only been telling what he feels like saying when he was asked, reporting the word (if any) that comes to mind.

Morris Lazerowitz (1992: 145) appeals to Locke in arguing that, ''The meaning of a word is not an essence, a common property''. He asks us to imagine a series of creatures, the first being a horse, ''changing by imperceptible gradation into a swan''. This could be set up like our sorites series, working with our trusty imaginations. He argues that,

If . . . there were a property . . . in virtue of having which the animal was a horse . . . it would be possible to know at exactly what point in the process of transformation the animal ceased to be a horse . . . The fact that there are no sharp lines of demarcation shows that there is no property common and unique to all things, actual or imaginable, to which the word 'horse' is applicable and the failure to have which makes the word inapplicable.

This misrepresents Locke. That claiming our usage is regulated by recognition of a real essence leaves us open to embarrassment about locating the breakpoint does not prove that no such point is marked by a real essence. It requires some imaginative power, not merely to

[3] This sentence features the sort of loose usage discussed in the last section.

construct the cases, but to find them embarrassing. A ''scientific'' advance might enable us to spot a breakpoint. (If we (adequately) imagine the advance, we count it for philosophy.) However, it requires no advance to know at exactly what point in the transformation the animal ceases to be called by ''horse''. Thus it seems that the view that there is such a property as being called by the term ''horse'' passes Lazerowitz's test.

Perhaps there can be sorites series such that, in case 1, someone definitely calls a thing by ''horse'' and, in case *n*, does not, with the breakpoint hard to find. There could be a series in which the word fades, or the speaker or writer goes from a human orator to a quacking duck. The possibility of such a series would show, according to Lazerowitz's rule, that there is no property of being called by a term. It is not a good rule, nor is it one Locke accepted.

If we ignore questions about intentions and focus simply on the production of the term in reference to a thing, it is very easy to tell if *A* has called *x* by a term. If someone is ordered at gunpoint to call his brother a ''fool'' he might be excused from the judgment due for calling a brother a fool. But he could say, ''I would rather die than call him that!'' Still, if he were hooked to a machine that electrically causes his body to produce the term in speech or writing while pointing to his brother, it seems incorrect to say he did the calling. We haven't proved that ''x called y, z'' names a relation. Nor has it been proved it does not. Either way, we can still use the pattern.

There has been puzzlement about ostension, beautifully reflected in Quine's (no doubt ironical) definition: ''The *ostended point* . . . is the point where the line of the pointing finger first meets an opaque surface'' (emphasis added). And there is always Cartesian scepticism. Someone saying (to himself), ''That is a pile'' may not, according to scepticism, be applying a predicate to anything. The phenomenalist reply elevates this risk to a new rule. The metaphysically proper predication is, ''This is a pile-encountering type of experience''. The (primary) subject will never be a mind independent thing.

On another view, the speaker may not be applying a predicate to any one thing, but only to ''a collection of atoms arranged pile-wise'' (which due to the alleged powers of ''plural quantification'' is not some one thing, the collection). These intrusions can remind us that there is as much philosophical conflict over subjects as predicates. ''The property of being a material object (of readily perceivable size)'' is called in question both by the phenomenalist and the atomist in a way that has large

implications for the relation of ''pile'' or similar terms to properties. They offer an alternative way of avoiding sorites, by getting rid of the troublesome subjects of predication. Locke preserves the subjects by weakening the predications. (These escapes are somewhat similar.)

3. LAZEROWITZ'S INTERPRETATION OF LOCKE

Let us suppose there are horses and we are observing one in the company of ML, who says, ''If . . . there were a property . . . in virtue of having which this animal was a horse . . . it would be possible to know something we cannot know''. We would rightly take this to entail that there is no property in virtue of having which this animal is a horse. This is an awkward saying. ''It is not a horse by virtue of being a horse'' is much easier to follow, as a conceptual truth about the use of ''by virtue of''.

However, the latter reading does not support the conclusion that ''there is no property common and unique to all things, actual or imaginable, to which the word 'horse' is applicable and the failure to have which makes the word inapplicable''. The conclusion could appeal to the fact that ''horse'' is applicable as a mass term to a certain narcotic. But sticking with the Budweiser Clydesdales, the fact that no one of them is a horse by virtue of being a horse does not entail that it is not by virtue of having the property of being a horse (or being a horse) that the word ''horse'' applies. ''Being rightly called by 'horse' by virtue of being a horse'', though it is a peculiar explanation we will need to criticize, does not have the obvious absurdity of ''being a horse by virtue of being a horse''.

This pushes us back to the awkward verdict. It is a trivial grammatical transformation to go from ''This is a horse'' to ''This has the property of being a horse''. Sydney Shoemaker (1997: 229) says, ''Philosophers sometimes use the term 'property' in such a way that for every predicate F true of a thing there is a property of the thing which is designated by the corresponding expression of the form 'being F' . . . It is natural, however, to feel that such properties are not 'real' or 'genuine' properties.'' This is an excellent point, though the formulation needs some qualification. All properties are genuine properties. Furthermore, we might agree that it was true to say ''Bill was sick yesterday'' and thus that the predicate ''was sick yesterday'' is true of Bill, but the related

property would not be ''being was sick yesterday''. The grammatical transformations involved in speaking of ''the property expressed'' by a given predicate are not easy to state in full generality. Let us simply speak of the ''grammatical switch'' from talk applying a predicate to talk of the instantiation of a property.

Representing Platonism as treating such grammatical transformations as straightforward logical inferences, forcing someone who believes in horses to accept properties, is a common device of its detractors. Platonists do not regard properties as trivial shadows of words. Still, it is not easy to make sense of ''This is a horse but there is no such thing as the property of being a horse''. Furthermore, the redundancy of property talk does not prove it never designates a property. Being redundant, unnecessary for the relevant purposes, superfluous, is not the same thing as not referring or expressing.

Philosophers who say the predicate of ''Doing X is good'' does not express a property, and that the form functions primarily to show the user favors doing X, will themselves favor some X. They will not say, without explaining, something such as ''Doing X is good but there is no such property as being good''. Their idea can be described in terms of ''taking a field linguist's point of view'' towards your own language use or that of others (warning that this is not based on knowing what a scientist actually called a ''field linguist'' does).

The ''field linguist'' observes creatures to explain their behavior. To call the behavior language production is a commitment. It is a further commitment to divide the sounds or marks produced into such as subjects, predicates, nouns, verbs, etc., and further still to describe the subject as applying a predicate to a specific thing or a kind *K*.[4] These descriptions may nonetheless be scientifically useful and at least something like them is needed to qualify as ''linguist''. By contrast, it is generally dubious to explain an agent A's applying a predicate term *t* to a thing *x* by appeal to the idea that *t* expresses a property *F* for A and *x* is an instance of *F*. The dubiety is brought out by sorites problems. For if A applies *t* to *x* on one occasion and not another and there is no relevant difference in *x* (as when *t* = ''haven't seen this before'') or in A (say if the gun has been removed) then *t* expressing *F* and *x* being an *F* is not a good explanation. Such irregularity in application is a primary feature

[4] Remember though, that this ''calling'' is minimalist. Coerced calling would presumably get a very different explanation from more ''spontaneous'' productions.

of sorites series. Thus, such series offer some reason not to explain application of a term by its expressing a property.

There remains the explanation by appeal to the idea that the term *t* expresses the property *F* for A and A believes *x* is *F*. The fact that someone is mistaken in thinking *x* is a horse does not prove that the term ''horse'' does not express for him the property of being a horse. Of course, if he calls everything ''horse'', we may suspect his term is best translated ''thing'', or if it's foxes, by ''fox''. But not every mistake is grounds for denying that his term expresses a property.

In reply, it could be said that the belief explanation, though a common explanation for someone's applying a term *t*, is really a mediocre one. A sorites is supposed to show A applying *t* to *x* one time and not doing so the next, with no *F*-relevant differences. A really good explanation for the applying would not settle for ''he believed it was *F*''. It would explain *why* he believed that, with an explanation not resting on a causal role for *F*. So why should *F* ever be needed in the best explanation? Suppose that there is a range of brain states of A, B_1 through B_n and a range of properties of *x* (and perhaps the surroundings) H_1 through H_m, such that there is a law *L* successfully describing which pairs B_i, H_j predict a calling by *t* of *x* by A. If *F* is just one of many Hs or even not among them, then the *L*-explanation makes the appeal to *F*-belief idle.

In so far as the field linguist explanation of A's applying *t* to *x* has no use for *t* expressing *F* for *x*, it has no use for the idea that *t* is ''true of'' *x*, where that means expressing a property of which *x* is an instance (a ''sense-reference'' theory). Pragmatic nominalism offers another meaning. Applying *t* may be more or less useful for the language community. Degree of usefulness might be called degree of truth, or a high degree of usefulness simply called truth. These are not competing theories of truth, but different meanings for ''true'', where meanings need not be properties, but various kinds of social forces. A native says, ''This is edible'', resulting in behavior in response. That doing what is called ''following the advice'' of one who applies ''edible'' will have good results can be marked by calling the predication ''true of'' its subject. There may also be a trivial grammatical expansion of ''That is edible'' to ''The proposition that that is edible has the property of being true''. The nominalist can call these sayings ''true'' (useful) or ''false'' without granting that any propositions or properties were expressed in the native talk.

To make sense of calling *x* a horse while denying there is any such property as being a horse, you could say ''That sure is a fine horse!'' and then ''But to get back to science, when I applied 'horse' just then, my doing so is not explained by use of a property description obtained from the term by common grammar. There is no such property (or 'property'). The vicissitudes of my application are best explained by the law *L*.'' You would speak of your ''applying 'horse''' rather than of your ''attributing the property of being a horse''. You may not know of any such law *L*. You could still believe there is one, on the grounds that sorites troubles show there must be such an explanation, and then on those grounds you could deliver the verdict from the field that your native use does not really express a property. That this is possible for the field linguist does not mean it is required. He may do fieldwork on mathematicians and credit them with expressing lots of properties. The perspective provides an option which is compatible with nominalism or Platonism. But it is concerned with causes of, rather than reasons for, application of predicates.

Locke does not need that option. He would allow the predicate user to claim a real essence for which unaided observation is unreliable. If you want to avoid possible embarrassment, then switch to a nominal essence. Locke might not have liked a Platonist harping on this, but both real and nominal essences are properties; he was not trying to make out predicating ''horse'' as not attributing any property.

4. INTERPRETING LOCKE'S PROPOSAL

Predicating ''pile'' of *x* does not entail calling *x* a pile, and conversely. We might solve a puzzle as to why A called a certain aspirin ''a pile'' (both in speech and writing) by learning that A belongs to a language community which uses the word ''pile'' the way we use ''pill'' (and where literacy and standards of spelling are high). (Some will say that then A's term ''pile'' is different from our term ''pile''; this is an unwarranted complication of the identification of terms.) Nonetheless, ''I feel like saying it's a pile'' would ordinarily be just as modest as ''I feel like saying 'pile'''. The ordinary speaker won't be fancy about quotation marks. The idea is that he need not, in either case, be claiming (or denying) that there is such a property as being a pile. He may be unable to note this himself, but the field linguist can treat either

one as a way of applying ''pile''. In so far as no property expressed by ''pile'' is involved in explaining the application, this is not an objectionable conflation.

By contrast, Locke's proposal is best seen as a recommendation for a contribution in Platonic dialogue, which is concerned with identifying properties, universals, and understanding them better in order to improve ourselves. Belief in the existence of such properties is Platonic realism, but excellent dialogue can take place with nominalists. They may do much better than the average non-philosopher, except that they will not be able to properly appreciate their merits.

Properties are universals, primarily, not in being common to many instances, but in being attributable or deniable of many. (You may correctly deny that x is an instance of the Washington Monument, for any x.[5] This only shows that being an instance of the Washington Monument is a property, not that the Washington Monument is. Being an instance of being an instance of x is being an instance of x, when x is a property. This does not hold for non-properties.) By this definition, a non-instantiable property, such as being a rational root of two, is a universal. This definition provides a basis for criticizing such notions as ''tropes'' or ''haecceities''. If I say of something ''This is this'' I am not saying something I can attribute at all, even falsely, to anything else.

A proposition is essentially connected to properties in being a thing which is *to the effect that* certain properties go together in certain ways. It is not a going together of properties, raising the problem of how they are held together, but is, when said or asserted, a saying to the effect that those properties go together in such-and-such a way. A simple proposition is essentially something which can be asserted or denied in dialogue (that is, in ideal philosophy), which requires that it be understandable to some good degree by a number of persons and thus public. Dialogue may be imagined in private thought, but if it is not self-deception it can be made public, not to every mob but to any dialectical gathering. Overemphasis on the private is why phenomenalists and conceptualists are not good at identifying universals.

There are also truth-functional compounds of simple propositions (though simple propositions are not anything like the mythical

[5] Including x = the Washington Monument, though someone inspired by Quine's device of identifying an individual with its unit set might propose something like that here.

''atomic propositions''). We can see this clearly and thus know that there are propositions we could not understand. Similar considerations apply to properties. It is essential to a (simple) property to be attributable or deniable in good dialogue, and to many, so that there is no such property as ''being this''. Not that properties do not exist independently of being publicly thought of. It is just that simple properties are essentially publicly thinkable. This means that an essential property of properties and propositions is essentially connected to the property of being possible. To the extent that our understanding of that latter property is clouded, so will our understanding of the nature of propositions and properties. Unfortunately, that is a considerable extent.

There is a significant difference between the two kinds of universals. One can hold, for example, that ''His negligence caused the accident'' expresses a proposition while denying that ''x caused y'' expresses a relation or that ''the relation of causing'' designates one. This leads to the analysis of the proposition to determine what properties are really attributed in asserting it. That a sentence expresses a proposition means that there are properties there but, freed of the simple grammatical transforms, finding the properties can be considerably more work.

Platonic realists are primarily interested in value properties and have sometimes resisted the idea of *low* properties, such as that of being a clot of mud or a ball of hair (if those are properties). But understanding *lowness* or indifference is also important in elevating. There is another type of realist, descended from Aristotle, the scientific realist, who seeks genuine properties to be expressed by descriptive, rather than evaluative terms. They have surrendered one of the principal tests we have for when a predicate expresses a property—that the predicate serve as a proper instantiation in the predicate place in (a statement of) a logical law. They deny that properties are closed under logical operations, deny, that is, that for all properties *F* and *G*, there is a property of being either *F* or *G*, a property of being both *F* and *G*, and of not being *F*. They, thus, leave logic to the nominalists. The compound predicates may be perfectly correct grammatical constructions, but according to the scientific realists these cannot all express properties. However, in dialogue, if you can attribute (or deny) being *F* and similarly being *G*, then you can attribute (or deny) any truth-functional compound you are able to grasp. Platonic realism differs from scientific realism in two essential respects: the emphasis on value properties, and the insistence on the universal applicability of classical logic to properties and propositions.

Scientific realists allow theoretical inferences about properties somewhat independently of our ability to reliably detect them. A scientific law may be established which licenses a counterfactual ''If we had done test *T* the verdict would have been property *F*''. (Some of these realists are unable to give examples because they are waiting for ultimate science. For our purposes ''having temperature *n* degrees'' should do.) For horse sorites, the solution, if any, might be something about internal genetic makeup.[6] A scientific realist might give up on ''horse'' and turn it over to the nominalists if the proper tests did not turn up. This makes for a conflict over philosophical method. For imaginary breakdowns are impossible to forestall, except with question-begging stipulations. Suppose that an immensely expensive scientific project finds what really determines horses. Something beyond phenotypes or genotypes; we might call it the ''longgreenotype'' (LGT). Confronted with a horse-to-swan sorites, the LGT test gives consistent answers along the most difficult border zones.

The imagination need not take this imaginary discovery lying down. Unafraid of there being any genuine horse-to-swan sorites, imagination can exploit the imaginary science to slip in some clinkers as breakpoints. We might just find that an entry all speakers agree is a routine horse (perhaps a recent Derby winner) comes out, on the LGT test, as a crocodile and a breakpoint in the series. People might get rebellious (''That's a croc, all right!'') but if the animal is not fertile with mares, but when crossed with a crocodile produces what appears to be a horsy-looking crocodile, they might quiet down and hope matters can be confined to a side-show. Empirical-theoretical ''criteria'' can conflict badly with ''conceptual criteria''. Both are vulnerable to surprises, but in different ways. Still, both kinds of realists can agree in resisting pragmatic popular usage as a basis for attributing the expressing of properties.

The basic question of Platonic dialectic is this: ''You say this is a case of *F* (justice, piety, love, beauty, a pile, etc.). Are you speaking of a property, or merely using a word you have been taught to produce on occasion, for reasons you may not understand?'' This might be put as, ''Why do you call this an *F*?'' But that can be badly misleading, because that is ordinarily used as a challenge or a request for a justification, so

[6] Here ''genetic'' is used in the way philosophers use ''atom'', and I have been using ''field linguist'' without appeal to genuine science.

that philosophers might be led to confuse, ''What are you saying in saying this is an *F*?'' with ''What is your justification for calling this an *F*?'' For most applications of terms, requests for justification are senseless, even when we step back and warn we are merely engaged in philosophy. Someone says, ''Watch out, don't step in that pile!'' ''Why do you call that a pile?'' may be a nonsensical response. You may succeed in explaining that you wanted to call a halt to the daily grind and do some philosophy. That will not save the request for justification. The philosophical truth may be that none was needed and that requesting justification was unjustified. The move to a philosophical perspective does not make the request for justification justified. Philosophical scepticism is fostered by ignoring this.

The question as to what property you were attributing may be equally out of place in everyday business, but it need not presuppose that there was any property being attributed (or denied). Of course the sceptic does not presuppose that there is justification either but he does assume that there being none is a defect. The verdict that no property is attributed is not any criticism of the performance. The field linguist, in refraining from describing a saying as expressing properties does not mean that as a criticism of such a saying. Perhaps, ''You have no justification for saying that'' can also be separated from any implication of criticism. The questions might be made analogous, one asking whether there was justification, one whether there was a property attributed. The answer is clearer with respect to justification, that is, that none was needed. The question whether a property was being attributed, and if so, what property, is quite cloudy. We have a perspective for making sense of saying such as that there are horses but no such property. That does not provide a basis for determining whether or when such verdicts are correct.

One symptom of this is the dispute between ''externalism'' and ''internalism'' about ''content''. There are various notions of ''content'' or ''meaning'' but one concerns how to credit ordinary speakers with expressing or thinking of properties. Some content theorists have peculiar reasons for overruling the grammatical transformations. A doppelganger from a ''Twin Earth'' phenomenally indistinguishable from our place, who by some miracle has recently arrived unknowingly, may say to us ''Thanks for the water'' and allegedly, unbeknownst to him and us, not mean what we do by ''water''. If we ask the poor newcomer what property he is attributing, ''the property of being

water'' won't do. (A doppelganger ''twinhorse'' would make a terrific sorites breakpoint.)

To determine the content of someone's speech or writing, why not ask him, as in philosophical dialogue? The debaters over ''content'' use the field linguist perspective. Here we may have a misunderstanding about ''meaning''. It has been argued as follows: ''The meaning of 'It is raining' does not change from day to day. The proposition expressed by various uses of 'It is raining' does change from day to day. Therefore, the meaning cannot be a proposition.'' The same argument can be run for properties. It involves an equivocation. Theorists of ''content'' (or ''meaning'') are welcome to the meanings that stay the same between nonequivalent assertions. They are a respectable topic for technical specialists. Propositions and properties are of interest to dialogue.

Of course, in many important language uses, the user is not reflecting on universals and his considered judgment as to the universal he is applying might be worthless for the task at hand. In lawsuits, what the court decides you said does not require your assent. What your guarantee of ''delivered promptly'' or ''made of aluminum'' or ''pure water'' requires of you will not be determined by your dialectical meaning (especially an initial offer). Dialogue can be a luxury which the needs at hand do not allow.

Again, if we ask Smith what he means in saying *L* ''is tangent to'' circle *C* at point *P*, he may respond that he means ''is a line in the same plane intersecting at just *P*''. That property will then be his initial dialectical meaning. When he is shown an extension of *L* intersecting a triangle *T* at a vertex *V* and then intersecting a spiral *S* at a point *P*′ such that *L* is the graph of the derivative of *S* at *P*′, he may become confused, surprised to find himself wanting to say *L* is tangent to *C* at *P* and *S* at *P*′ but not to *T* at *V*, which would be doubly inconsistent with his opening choice of meaning. This merely reinforces the point that the dialectical meaning is a poor candidate for explaining the agent's disposition to apply the predicate. This does not make dialectical meaning a bad candidate for the ''content'' of a judgment.

It does leave ''content'' vague, because a person's formulation of what he means by a term may change so much in the course of dialectical discussion and depend on the dialectical questions raised. You may decide that in calling this ''water'' you mean it has various observable characteristics. If you are asked whether you are assuming you are not a recently arrived doppelganger who has never been exposed to water

before, you could reply that you are assuming that, or that that is irrelevant; you are only concerned with the observable characteristics you have somehow associated with the term ''water'', perhaps even if you have just come into existence along with a feeling of having a past, etc. Your ''meaning'' can go a variety of ways, with the clarity as to which coming from what you are able to articulate to others listening with fairness and intelligence, and that evolves in the course of dialogue. Whether it was what you originally ''meant'' in some other sense is not the main concern. A content theorist could be after something else, but conflation of these projects is worth avoiding.

Locke saw that we all have introspectively detectable dispositions to apply words. In reply to, ''Why do you call that a chair?'', ''It's the word that comes to mind'' may be, in some cases, a fair reply to philosophers demanding more. Neurophysiologists may provide a lot on what causes the application of ''chair'', explaining *why* it comes to mind. With the application of, say, ''embezzler'' or ''blood type A'', the reply ''It's the word that comes to mind'' won't do. Normal people do not say ''embezzler'' simply on sight (unless observing a silent embezzling movie). Still, you could say, ''That person is such that my background mental condition makes me feel like calling him by the term 'embezzler' ''. Unlike the ''chair'' case this would rarely be an adequate reply. It is nonetheless a significant property available to be meant.

Locke does not present his account as an explanation of why the native applies the term, but as a recommendation for a dialectical offer of a sense. It would be odd to recommend to someone who shouts, ''My kingdom for a horse!'' that he ''keep to the nominal essence''. This makes sense as advice for an occasion when you are trying to determine what universal you may be justified in attributing.

What property *A* expressed with the term *T* may be determined differently in a dialectical discussion, a scientific explanation or a legal verdict. The result may not be well described as ''What *A* meant''. It may still be the relevant universal. Thus, this is not a general account of what is ''meant'' by an application of a general term, but rather, one way of replacing the unclear notion of a general account. ''Coming up with the property in dialogue'' will strike the unsympathetic as being too similar to just expressing the property, only worse, because the dialogue may be at a considerable remove from the actual utterance. The association of ''meanings'' with utterances can divide into a confusing variety of different games with language. This replacement gives up on

confidence about content where there is no associated dialogue to determine it unless there is a well established relevant law as to what would be arrived at if there were good dialectical discussion. A counterfactual in the absence of such laws would only cloud the question of content.

It is not that dialectic features some secret or mystical technique for determining whether a term expresses a property for a user. It is just that, having learnt by pure reason, from contemplation of such laws as LNP, some grasp of the property of being a property, we can appreciate the importance of grasping specific examples of properties. We can be guided by a sense of the special importance of value properties and take easier cases of properties as valuable in strengthening our understanding of the value properties. Mathematical properties are especially useful in this way since many of them are especially accessible to pure reason. But whether there is such a property as ''being a pile'' can also be a helpful question. Only property-expressing predicates can correctly instantiate principles such as LNP.

In dialogue we want to transcend the merely everyday pragmatic use of language but without alienation from it. When we say in practical life, ''This is love'', ''That was a noble deed'', ''I'm home!'', ''That is a pile'', etc., the practical functions served can be evaluated in a special way by determining whether there are properties that can or should be connected with key terms. In this undertaking the dialectician's only advantage is an idea of what he is seeking and an understanding of what is involved in being a genuine universal. There is no litmus test for when the grammatical transformations lead to a genuine description of a property. There is a kind of discussion where the problem is properly appreciated and it is recognized that success depends on communicability.

Scientific realists and Platonists care about properties in different ways. Both can agree with the triviality that it is good, or at least, accurate, to recognize the existence of every genuine property, that is, every one that exists. But there is one universal such that recognizing it is essential to Platonism: goodness. It is ironical to speak, for example, of ''mathematical Platonism'' as if the mere recognition of mathematical universals suffices. If there is no property of being good, Platonism is a wild goose chase, though that couldn't be bad, since badness goes with goodness. Even doubting there is such a property is problematic, since it exists necessarily. The nonexistence should be in some sense a

dialectical possibility, as when we discuss whether there is any such thing as justice. It is like the difficulties over ''epistemic possibility'', where we seem to seek a kind of possibility possessible by things that are impossible.

5. THE MODESTY OF LOCKE'S PROPOSAL

If Locke's account were translated into an account of applying truly, it would have it that a term *t* applies to *x* for A if *x* strikes A in such a way as to lead him to apply *t* to *x*. This is very different from suggesting that A claim to mean, by *t*, being a thing which strikes A so as to incline him to apply *t*. But with that account of the content, the sense-reference account will have the same verdict as to truth as that account of applying.

As a dialectical candidate, Locke's modest suggestion as to what you might claim to attribute with your use of a general term has a feature effectively exposed by Plato in the *Theatetus*. It leaves little room for error. If a rock falls, hitting you on the head and lands at your feet, and you point to it and sincerely say ''That's a pile'', meaning just that it struck you in such a way as to lead you to apply the term ''pile'', then you are right. You may say ''That shows what a wrong account of meaning this is!'' That is just a reminder that dialectical meaning is not the only ''meaning'' feature related to a statement. When it ''struck you as a pile'' you may have had no concern with what property, if any, you were attributing. When you reflectively consider this use in a dialogue you may be more interested in what you could or should have conveyed or even what you now see yourself as conveying. Good dialogue will require that you are no longer stunned.

That line as a suggestion for dialectical meaning is at least much better than as an account of applying. That encountering *x* inclined you to apply *t* would rarely suffice to make the application correct. But if the content of your application of the term is only that *x* inclined you to apply that term, then you were right. But is the claim then trifling? That is not a simple question. Locke is highly ambivalent about his ''real and nominal essences''. He sometimes recommends keeping to nominal essences as the right way to respond, for example, to problems about changelings, and various other cases. But he sometimes treats with contempt the course of keeping to the nominal essence and not

attempting to find a real essence, as in his comment on the question whether bats are birds, where he depicts the question in terms of real essences to be serious and worthwhile, and the nominal question to be verbal and trifling (Locke 1959: 107–8).[7]

Pragmatic nominalism offers extensive supplementation to the idea of nominal essence without forgoing the benefits for sorites problems. You may venture further that your application was the result of a stable disposition coordinated with other people so that you expect your applications,[8] even if not accepted by others, to qualify as within the community range (''sound rational''). Furthermore, there will never be anything but agreement and success with your application. To a Platonist, this is not enough. It is easy to imagine a story in which an innocent accountant dies and ''embezzler'' is a pragmatically successful predication, but false nonetheless. The truth transcends the language use. They may reply, ''If a rational observer had seen what the accountant did he would not have agreed with the charge''. There is the old problem about applying this to show the falsity of ''The incident had rational observers'', but anyway, counterfactuals about what would be said are not a good way to explain a meaning.

We may believe that whether the accountant embezzled cannot be a matter of how people use language, no matter how successful and long term the pattern of use may be. We think the nominalist is involved in a vicious regress. In reply, he can cite his freedom from sorites problems, easily noting that agreement in usage breaks down in sorites series and that this is merely a fact about humans and their language, not a paradox about language-independent properties, since there are no such things. We are apt to become deadlocked with the nominalist in a disagreement as to what it means to be language-independent.

The issue relates to whether ''Applying 'embezzler' ('pile', etc.) to x is a maximal pragmatic success, therefore x is an embezzler (a pile, etc.)'' is a valid argument. We feel that even the nominalist should admit that it is not. This would strongly encourage the impression that ''That is a pile'' must be attributing ''the property of being a pile''. For if it is a

[7] Locke has been misunderstood on this point, e.g. by Richard Boyd (1981: 70), who says, ''Locke holds that the question whether bats are birds is a purely verbal question.''

[8] Here we are talking about one application, so that there may be very little social impact to assess. We make more of it by talking of the impact of other similar uses. This will suggest talk of what impact other hypothetical uses would have if they occurred. This gets into the shade of counterfactuals, but if they are based on general laws that may be clear enough.

nontrivial point that ''That is a pile'' is not deducible from simpler property ascriptions, then it would seem that it must involve a stronger property ascription. This is not a correct inference. Again, it will seem that ''This is perfectly flat; therefore, this is not a pile'' is a genuinely valid argument. Since genuine validity is a relation between propositions, not sentences, it will be concluded that since ''there is a necessary condition for being a pile'' there must be the property for which it is a necessary condition. This is an illusion that warps much ''conceptual analysis'', leaving it adrift in the area between meaningless grammatical predicates such as ''ardufufardable'' and genuinely property expressing ones. Nominalists can do respectable ''conceptual analysis'' without accepting properties.

We still charge them with regress. Someone may say, ''Alpha is ardufufardable'' and explain that he means Alpha is such that applying ''ardufufardable'' to it is a pragmatic success. This seems likely to be false, but there is no telling what will catch on. Anyway, if it does not catch on, the claim will be false. For us Platonists, this would mean the application did not have the property of being a pragmatic success. For the nominalists, it would mean that applying ''pragmatic success'' to the initial applying was not a pragmatic success, or better, that applying ''not a pragmatic success'' was a pragmatic success. The nominalists will likely get the right verdicts, whatever we may think of their reasons. How could people agree in calling *x* ''ardufufardable'' and not agree in calling it ''called etc.''? But it's possible. People might be put off by the complexity and dismiss it as gibberish. Don't we now have the nominalist on the logical ropes? We may think we have scored a dialectical ace for our side, except that we are apt to hear, ''What 'logical ropes'?'' We need to discuss the nominalist view of logic. But first we will consider the nature of logical ropes.

6. THE RELEVANCE OF ''ALTERNATIVE LOGICS''

Sorites problems cloud LNP. They also involve such logical principles as the Law of Excluded Middle (LEM) but LNP is a preferable focus because it is math, and we do not have ''alternative math'' in the way we do have ''alternative logic''. Even the intuitionists, who reject the Law of Excluded Middle and classical analysis, accommodate mathematical induction. Their formulation has their characteristic quirks, but it entails

that it is absurd to deny that there is a breakpoint. This is enough to show that logical pressure against denying a breakpoint does not depend on LEM. Even their pressure on LEM is different from the usual borderline cases. Pi is not a borderline case of the property of having 13 consecutive 7s in its decimal expansion. Sorites slides and borderline cases raise similar, but distinguishable, problems. Borderline cases are problems for the LEM. There can be sorites with no borderline cases. We can focus on a borderline case. It can become famous and easily recognized. The characteristic trouble in sorites is inability to focus, blurring of cases.

The study of formal systems has produced many other ''alternatives to classical logic'' besides intuitionism. But they are merely alternatives to various formal systems that exhibit some syntactic resemblances to sentences that have been taken to express theses of classical logic. A formal system is definable and its syntax is part of the definition. A natural language isn't definable and has a ''syntax'' in a different sense. The study of formal systems has a tenuous connection to philosophy and the study of formal ''interpretations'' has an even more tenuous connection with philosophical interpretations. Such technical achievements as Post-*consistent*, formal systems of paraconsistent logic need not be associated with actual contradictions or with grammatically ''contradictory'' sentences of natural language.

Whether a string of marks qualifies as a sentence of English depends to some extent on the use English speakers make of it. The extent may be like that of a formal system, zero, when the string is sufficiently similar syntactically to established sentences. So there may be sentences that are grammatical English that are never used, just as with formal systems, where it is certain that the majority of wfs (well formed formulas) are not even physically possible to produce. But for some strings, the practice of speakers can add the string to English regardless of what syntactical rules have been in force so far. English may share with a formal system such a rule as that if P and Q are both grammatical sentences and can both be written out with ''and'' in the right way, then that conjunction is also grammatical. Or this might be true for at least some base sentences. This is a recursive rule. But the grammaticality or understandability of English sentences cannot be reduced to any static body of such rules.

This can be overridden by speaking of ''English as of time *t* among speakers *C*'' or the like, where it will then be a contingent fact that this is a certain system of syntax and perhaps ''semantics'' or ''pragmatics'', in an attempt to make the language analogous to a formal system. Such

''idealizations'' could well be useful for some purposes. It would be too harsh to say, '' 'Mathematical logic' has completely deformed the thinking of mathematicians and of philosophers, by setting up a superficial interpretation of the forms of our everyday language as an analysis of the structures of facts. Of course in this it has only continued to build on the Aristotelian logic.''[9] But this is worth quoting. In semantics for formal systems, there is no basis for worry as to what may be the interpretation of a wf. A fixed interpretation completes the syntax into a ''language''. It is easily granted that the same syntax may thus govern many distinct languages, but for a ''formal language'', the interpretation of the ''predicates'' is fixed. This is far from the case with natural language (that is, language), where we may well wonder what, if anything, some sentence means, a question which would be absurd with respect to an interpreted wf of Predicate Logic.

7. THE LOGICAL RESOURCES OF NOMINALISM

Pragmatic nominalism offers a different sort of ''alternative'' to classical logic. It offers an account of what it is for a sentence to be true or a predicate to be true of something which is not put systematically in terms of a recursive syntax. This is convenient, since natural languages do not have such syntax (even though they may include a lot of recursive rules), but would seem to be risky logically. ''He is and he isn't'' may be a pragmatic success. How does this square with the Law of Non-Contradiction?

For some of us, it is a truth of pure reason that an argument of the form ''If P then Q; therefore, if not-Q then not-P'' is valid and we call that the Law of Contraposition. But the ingenious linguist may cite such as ''If he doesn't live in Paris then he lives somewhere in France; therefore, if he doesn't live anywhere in France, then he lives in Paris'' as if it were a counterexample to this law. The premise may be a hit and the conclusion a bomb. We may reply that the premise is not really a conditional and the purported instantiation is merely a grammatical illusion. Here, as with ''He is and he isn't'' the nominalist has an easy answer.

[9] See Wittgenstein (1956: iv. 48).

First, he may ask how we know his example is not an instance of Contraposition or that ''He is and he isn't'' isn't really a contradiction. Do we have some effective syntactic test for being a conditional, as in a system of formal logic where Contraposition is an unproblematic axiom, theorem, or rule of inference? Or do we ask that we be consulted as to what is a conditional, so that we can use our magical ''dialectical'' method? That would seem high-handed. If we can effectively identify some system of expressions and state recursive rules, then we can have them as we choose. If the rules catch on with the general public, they may be ''true''. But we can be happy with them in our group whether or not that happens. If we like saying of some sorites series that 1 is *F* and *n* isn't and there is no breakpoint, then go ahead; we may get popular approval. If we can't stand that, we can always insist on formulations in first order arithmetic. Critics can't bother us there; they will just be disqualified as not playing our game. And everybody will agree with that verdict. It is very nice how much agreement there is about whether formal rules have been followed when we stop worrying about whether those rules have ''language-transcendent truth'' (and we are welcome to try to make that predicate popular).

From this perspective it is unlikely that the nominalist must worry about our charge of vicious regress. Suppose that *x* is *F* by virtue of (1) ''There is happy consensus in applying a predicate for *F* to *x*''. By what logic does it follow that the truth of (1) entails (2) ''There is happy consensus about (1)''? By classical logic applied to the pragmatist account of truth and the initial verdict for the case. But the infinite regress argument will need mathematical induction. The nominalist may not dismiss mathematical induction out of hand and yet choose not to apply it here.

This is outrageous, to some of us, and irrational, if we mean by ''irrational'' ''knowingly rejecting verdicts of classical logic''. But the pragmatic nominalist will not mean that. Of course that will not make him any less irrational, but when we make charges like that, we are likely to crave consensus ourselves and not be satisfied with mere language-transcendent truth. There is a more popular sense of ''rational'' which means, ''replying like a rational person''. Suppose Smith agrees that if Bill is a neat guy, then he should be admitted to the club and we subsequently get him to agree that Bill is a neat guy. We may feel that ''Smith is irrational'' if he refuses to agree that that is sufficient for admitting Bill. This is easier to justify by appeal to

consensus than to modus ponens. For we may hold that ''Bill is a neat guy'' does not express a proposition but merely expresses liking.

Smith may not agree that if he *likes* Bill, Bill should be admitted. He may insist that his liking could be in error and what is required is that Bill has the objective property of being a neat guy. ''But you admit that he has this objective property and that if he has it then he should be in; you are violating modus ponens!'' Smith replies ''How dare you claim I admit he has the objective property while holding that there is no such property to be admitting he has?''

Here we might have a better chance with appeal to a different standard of rationality. You can't reasonably make acceptances in the pattern: ''If P then Q, and P'' while not accepting Q (where ''accepting'' is interpreted in terms of the right kind of cooperative behavior). ''If blah-blah then whoop-dee-doo! And blah-blah; therefore, whoop-dee-doo!'' is not an instance of modus ponens, but the fact that people feel it has force is a consideration relevant to pragmatic rationality. Feeling pushed to its conclusion may even be a way of celebrating a certain ubiquity of logic.

Nominalists can look to such forces in persuasion while denying any force to charges of necessary falsity due to language-transcendent rules. They can ignore logical challenges at the boundaries of their picture and still play by logical rules in many exchanges and with great ingenuity. And they have a formidable offer for sorites problems, an offer roughly foreseen by Locke. If we work with even the objective property versions suggested by the nominalist account of application, we will have no sorites problems. Why not? Because the nominal properties are too weak to give mysterious breakpoints. That agreement about application of a predicate often breaks down is no mystery. Being unable to locate a breakpoint for ''agreement'' or ''pragmatic success'' is also not surprising. Breakpoints could be found with precise tests for agreement, as in public opinion polling. But it would be, in general, a waste of time.

8. AN OVERSIMPLE, ''REALIST'' RESPONSE: EPISTEMICISM

There is a position which has been called ''epistemicism'', such that most philosophers who think they understand it would take it to entail that, in a series such as the one imagined above, there has to be a first

non-pile, a breakpoint, though we may none of us *know* (this must be the source of the name) just which point this is. This has seemed incredible to many (and one of the specifications for the imaginary series was that this should seem incredible). Epistemicism represents the incredible result as a consequence of LNP.

Let us call a ''sorites series'' a series 1 to *n* such that most respectable observers agree that verbiage *V* is applicable to case 1 and is not applicable to case *n*. That is, these applications are pragmatic successes. The nominalist could easily agree that the applicability of *V* does not make clear what property, if any, is involved in the series, there being no such things as properties. LNP claims breakpoints for any number-property series. It is a trivial consequence of LNP, not deserving a separate name, that there is a breakpoint whether or not we can find it. The question is, in general, what property is involved in a sorites series? What merits the title ''epistemicism'' (as distinct from ''LNP'') is the view that, for every sorites series, there is the property describable by the grammatical transformations which meets the specifications of LNP, so that there is a breakpoint for that property. (The ''whether or not we can find it'' is still a trivial consequence.)

This is an abstract definition of ''epistemicism'' which forgoes scholarly citation of texts, following the traditional attitude of Platonism in this regard. One can safely assume that the term has other uses that are confusingly close to our topic. But this version is a respectable view, worth discussing. It arises from partially deferring to the nominalists. The big success of ''That is a pile'' for case 1 and ''That is not a pile'' for case *n* is taken as proving that these remarks are true. So far, this is just the pragmatist line. If we grant there is such a property as being a pragmatic success, then the word ''true'' could be requisitioned to mark it. But then it is concluded that, since they are true, they express true propositions and true predications, which must correctly attribute properties. The grammatical devices for identifying these ''properties'' then give the result that it is the property of ''being a pile'', said of 1 and denied of *n*. So, LNP applies and we have a breakpoint, discoverable or not.

Skepticism about ''epistemicism'' arises from its avoidance of nominalist strategies and presumption of speaking of properties that are not essentially connected to linguistic practice, for which the guarantee of a breakpoint seems preposterous. The weakness of epistemicism against nominalism should be of concern to Platonists, for it creates a pressure in favor of the nominalist rejection of properties or else a nominalist

version of them. The idea that properties are in general not independent of our language use is unwholesome. The scientific realists have an answer. They would leave the piles to the nominalists and look for genuine properties at the frontiers of science. But this is unduly pessimistic and cuts off the layman from reflections important to the soul, which were available even when science was in its infancy. Epistemicism is careless about the risk of accepting bogus universals, in such a way as to give Platonism about universals a bad name. But scientific realism is not the remedy.

When we are dealing with mere predicates for which a nominalistic treatment is appropriate, then it is fallacious to apply logical and mathematical rules as if we understand what properties we are talking about. This is a fallacy worth naming and bearing in mind. Jose Benardete once remarked that what is known in mathematics as mathematical induction is known outside mathematics as the Slippery Slope Fallacy. (Respectfully) on the contrary, the Slippery Slope Fallacy is the fallacy of thinking that, if you take the first step on the slope, you will inevitably slide to the bottom, therefore, you should not take that step.

For example, Nelson Goodman (1976: 186–7) argues that to be a performance of Beethoven's Fifth, there cannot be an error of even one note by any player in the whole orchestra. The argument is that we have no general rule for discriminating between one-note deviations (in terms of their relevance to qualifying as a performance), and if we follow a generic tolerance of one-note deviation, then enough applications of this rule would qualify a letter-perfect performance of ''Three Blind Mice'' (relative to some general accepted transcription of this folk tune) as being of the symphony. This style of argument could easily be presented as another sorites paradox. It differs from our present pile sorites in that it is being recommended that we refuse to count a one-note deviation on the grounds that we will then not have a principled basis for stopping the slide from Beethoven's Fifth to ''Three Blind Mice''. Unlike the ruling that removal of even one grain undoes a pile, this seems to be at least coherently enforceable by sufficiently careful scrutiny of performances, based on a plausible candidate for a paradigm case, not of a *good* performance, but of one scrupulously following the score.[10] Reviewers could say that the orchestra came

[10] Here we ignore the potential sorites for such as ''playing A = 440''. They would at least be even more difficult to construct than our ''pile'' version.

close to performing the symphony, etc. This is coherently enforceable but nonetheless a bad idea, deserving the title ''Slippery Slope Fallacy''. People actually argue against certain practices that they will inevitably lead to various horrors unless severely regulated, in a way guilty of this fallacy.

The fallacy lies in the presumption that we cannot pronounce on the applicability of ''The Fifth'' or ''Three Blind Mice'' unless we can formulate precise ''criteria''. This is obviously false. Slippery Slope is more formidable in legal cases where there can be pressure for precise criteria to reduce the amount of litigation. Anyway, ''note perfect'' is a precise criterion, albeit a ridiculous one. There is no corresponding perfect standard for a pile. A better title to introduce here is ''the Fallacy of Epistemicism''. That is the fallacy of instantiating a formal mathematical induction without ensuring that the predicate uniformly functions to express a property. One criterion for whether your topic of discussion is a real property and not a mere predicate, is that the predicate fits in the laws of logic and mathematics in the places calling for property expressors. Even property-expressing predicates do not do this by logical necessity. But many predicates in perfectly ''correct'' everyday uses do not make fit property expressors at all.

In law, millions may turn on whether a predicate applies, say whether this counts as a ''vessel'' under maritime law. To assume that this is a matter of whether it had ''the property of being a vessel'' as the objective ''fact of the matter'' invites confusion. The drafters of laws about vessels may have simply failed to think of that case, so it requires arbitration, not to be confused with arbitrariness, but not to be viewed as discovery about the presence or absence of an eternal form. Some will say, ''We all know *X* is a vessel! I say, *Y* has that very same property, the property of being a vessel as instantiated by *X*!'' To regard such a dubious maneuver as Platonism at work is benighted. Do you know *why* you called *X* a vessel? ''Because it had the property of being a vessel'' is no answer.

Properties have no borderline instances. This can be obscured by the practice of using ''*F*-ness'' and ''the property of being *F*'' interchangeably, for brevity (this practice has been followed here). Fatness, wisdom, justification, etc. appear to come in degrees in that two cases may both be fat or wise or justified, etc. while one is more so than the other. For the property (if any) of being fat, wise, justified, etc., this would be absurd. Properties do not come in degrees. We may so use ''*F*-ness

comes in degrees'' as to speak truly. It is only harmful if this is confused with speaking of the property (if any) of being *F*.

For predicates, there can be ''borderline cases''. It would be wonderful to define that phrase, but it can be useful without expressing a property. People may apply a term and agree in doing so without expressing a property. Natives may discover a Coke bottle and name it ''alpha'', thinking it unique. Discovering other coke bottles they may call them all ''alphas''. They discover a Pepsi bottle and are stumped. Should the law of excluded middle keep them arguing? Of course it will either be an accepted ''alpha'' or it won't (for determinate tribe and percentage of agreement and testing method). That's not merely words. But if money turns on whether the Pepsi bottle is an alpha, substitution of ''alpha'' for predicates in logical laws may do more harm than good.

9. AN ESSENTIAL PLATONIST RESPONSE

Since we grasp the idea of a property, if we can identify a *t* and there is the property of being a term and the relation of a term expressing a property, we can believe that *t* is a term and that *t* expresses some property and we can believe we name that property by the grammatical transforms from *t*. Sorites and borderline problems can provide good reasons to doubt that a certain *t* does express a property, but those reasons are defeasible. With too much doubt about our terms expressing properties we might give up and become pragmatic nominalists. Too little doubt is just as bad. Nominalist strategies are tremendously workable for many predicates, if we can avoid vicious regress. That can be achieved by granting that being called by *t* is a property a thing can have independently of agreement in applying ''called by etc.''

For example, Quine (1963: 10) says, ''One may admit that there are red houses, roses, and sunsets, but deny, except as a popular and misleading manner of speaking, that they have anything in common''. We can indeed, as long as we can recognize they do have in common the property of being called by the term ''red'' and that this is not plausibly made out as the property of being called ''called 'red' ''.

We Platonists can grant Quine's doubt about ''red'' without fear of regress. The nominal property is more secure, but still a property. We can have the benefits of nominalism without the regress. Furthermore, giving up on the property named by the simple grammar transforms is

not giving up on ''criteria'' for applying the term. They will just not be a real definition. Without the pretensions of analytical defining of properties, ''conceptual analysis'', construed with proper modesty, can be a valuable activity, not ''a mug's game''.

''Criteria'' can be properties, and not merely nominal properties. It would be a bad misunderstanding if the present rapprochement with nominalism were taken as a large restriction on properties. What philosophy offers more? How many properties have the scientific realists given us? We won't reject a single one of those, so you cannot seek a better deal there. We needn't worry about the nominalists. Who is more generous? The epistemicists? Yes, and the Meinongians (there is considerable overlap). For Platonists, knowing when we have failed to identify a universal is as important as knowing when we have succeeded. It is surely true that our application of terms to a thing is caused by complex combinations of many language-transcendent properties. They are universals because possible to attribute to many things truly or falsely. We are only cautious about claiming to have realized the possibility. This does not leave reality dependent on verbiage.

Scientific realists deny that ''grue'' expresses a property, on the grounds of its compound structure. The compound being logical, there is no problem about it. The only real problem is with the components, such as ''green''. There are indisputably distinct colors that are labeled ''shades of green''. The presumption that there is more to this labeling than a pattern of declining agreement with ''green'' should be regarded with suspicion. (In fact, most scientific realists do suspect it, so their doubts about ''grue'' are unnecessary defiance of logical closure.) Some philosophers say that the experience of greenness has a phenomenal quality and this phenomenal quality is what it is like to have that experience. But what it is like to experience one shade of green is different from what it is like to experience a distinct shade. Both cannot be what it is like to experience greenness. The idea of *the* phenomenal quality of greenness is an illusion and in so far as ''the property of greenness'' is a phenomenal quality so too is it an illusion. This is nothing against phenomenal qualities, but it is reason to doubt that they are what is attributed in predicating ''green'' (except that predicating ''green'' may be classed, trivially, as ''attributing a phenomenal quality'').

As the nominalists say, we can get along fine with the term ''green'' and its applications (or even refer to this complex as ''the concept''). The degree of agreement in usage is high enough to make the term quite

useful. And the applications have phenomenological associations in a much more regular way than ''grue''. Neither the nominalist data nor the conceptualist supplements provide grounds for identifying a non-linguistic property. A sorites series from a green page to a red one in a thousand steps, pairwise indistinguishable in color, only serves to show that people do not have a grasp of what exactly about a color leads them to call it ''green''. To say it is ''being green'' is vacuous.

The same considerations apply generally to the ''phenomenal qualities'', such as feeling cold. You know you feel cold at 6 a.m. and warm at noon. But different feelings at different times count as feeling cold. What you do not know about these feelings is what leads you to class them under a common term. You may call it ''feeling cold''. But the commonality thus created by this linguistic practice is best described nominalistically in a way which provides no basis for confidence that there is a mysterious breakpoint.

Then there is knowing, or believing. The proposition that 373 is prime is known by me, as is the proposition that I am now wearing socks. That means both those propositions have the property of being called ''known'' by me. It does not license the conclusion that they both have ''the property of being known by me''.

Such predicates as ''is believed by someone'' or ''known'' do not express properties and so, Kp and Bp, though representing recursive syntactic patterns, are not propositional functions. Even if $K^{100!}p$ corresponds to a grammatical sentence, it is nonsense. It could arbitrarily be given a sense. One way would be to adopt the ''KK thesis'' according to which KKp follows from Kp. Along with K$p \Rightarrow p$ this makes $K^{100!}p$ just amount to Kp.

Some will say that the KK thesis is false. On the contrary, it does not represent a definite proposition. It is not that p may have the property of being known while Kp lacks that property. Rather, there is no such property, that is, no such property captured by the general form. The detective knows that Smith did it. ''Yes, and he sure knows he knows!'' This can mean the detective is tiresomely smug about cracking the case, or something else. The disputes of philosophers over what it is that is known in knowing that Kp should be a lesson to us.[11] It has been said

[11] The disputes would include many admirably clever and profound lessons, and perhaps a number of interesting properties related to ''knowing''. But no property or relation has been settled upon.

that its being common knowledge that p entails an ''omega series'' of iterations, Kp, KKp, ... $K^{99999!}p$ ... (for Kp = everyone knows that p). This is a good indication that the meaning of the common locution ''common knowledge'' has been lost track of, and keeping track of its meaning is not a matter of connecting it with one common property.[12]

''If Kp expresses a proposition and $K^{100}p$ does not, then there is a breakpoint''. True, if there is such a property as expressing a proposition. There is such a property as being a proposition, just as there is such a property as being a property. But the alleged relation of expressing a proposition, between a proposition and some propositional vehicle, such as a sentence, is like the alleged relation of believing, or knowing, between an agent and a proposition. The words have meaning. An instance of ''He knows that he knows that p'' may express a proposition. It's just that another use with reference to the same person and the same proposition p and the same occasion may express a different proposition. There are such properties as being called a ''proposition expresser'' or ''cognitively meaningful sentence'' or ''statement''. That does not guarantee ''the property of expressing a proposition''.[13]

It is not that, for example, all there is to knowing is being called by ''knowing''. It is not true that all there is to being called by ''knowing'' is being called by ''knowing''. The calling may be justified by all sorts of considerations and be associated with all sorts of nonlinguistic features. That does not require ''the property (or relation) of knowing''. But, when we apply a term, we should be able to ask whether doing so is judicious and fair. That requires trust that there is such a thing as goodness which transcends the evaluative terminology. ''Good'' and ''stupid'' may sometimes move closer together, as Nietzsche claimed, but that is not good. Or there is such verbiage as ''morally good but not epistemically good''. ''Good'' can be equivocal, but the property of being good is one thing, the thing to which valuing must appeal, which it must seek if it is to be valuing well.

It can be good to replace the question whether x is an F with the question whether it is good for us to call it an F. This strategy fails for the question whether it is good to call x an F. Just as we cannot replace

[12] This is argued at (slightly) greater length in Cargile (1970: 151–5).

[13] I may be charged with sawing off the limb I am standing on and there are lots of such ''limbs'' in this paper. But the tree is still standing. This is no ''Hindu rope trick''. Being called ''proposition expressor'' is not nothing.

''*x* is called *F*'' with ''*x* is called 'called *F*' '', we cannot replace ''it is good (fitting, judicious, etc.) to call *x*, *F*'' with ''it is good to call it 'good to call it *F*' ''. We can recognize there is no universal behind our term ''pile'' and that ''truth'' of ''That is a pile'' is up to us without ending up holding that what is true is in general up to us. For we may want to be right and do the good and judicious thing in applying our terms and assigning consequences to their application. It is possible to take a field linguist's perspective on our use of value terms and this can even be a valuable exercise. But to become stuck in that perspective would be bad for us.

Philosophy has been deeply influenced by the project of understanding the use of evaluative terms scientifically by explaining the activity of valuing in naturalistic terms, a goal which can lead to needless difficulty in understanding the evaluative terms. To attempt to ''analyze'' a value term can be a pointless effort to delimit a range of possible uses of a word when in fact there are no clear limits. The value term will in some fortunate cases express a value property and if we can think about that property we need not worry about the fact that our term for it can be connected with other properties. This is particularly true in epistemology. One use for ''justified in doing *X*'' is just to credit you with having done a good thing in deciding to do *X*. To be justified in believing that *p* is to have done a good job of deciding whether *p* by the means available to you. A good use for ''knows that *p*'' is to attribute being right and having a good understanding of how you are justified in believing that *p*.[14]

Rather than saying ''*p* is known (by Jones, by everyone, etc.)'' we can say ''Jones has done a good job on the question whether *p*, got it right and has a good understanding of how he qualifies as having done a good job on that question.'' We could represent this as G*p* and enquire about the iteration GG*p*, etc. If you take G*p* as saying of the proposition *p*, that it has been the focus of what we may call, for brevity, ''a good epistemic performance by Jones'',[15] then we could attribute that property again to G*p* and be off to the indefinite. The absurdity of $G^{100!}p$ is sufficient to

[14] A good understanding of why you are justified'' should rule out Gettier examples without requiring, say, that to know someone has AIDS you must understand what causes AIDS.

[15] This use of ''epistemic'' for brevity is no concession to those who hold that epistemic goodness differs from, say, ''moral'' goodness in that the same thing may be ''morally good'' but ''epistemically bad'', etc.

prove this is a misrepresentation. ''Being known by Jones'' does not mark a property of a proposition. ''Being a good epistemic performance'' can (perhaps with further explanation) mark a property. But it is a property of a performance, not of a proposition. Jones's epistemic performance can be a good one, and this is expressed in a proposition to that effect. We can then assess the performance, by Jones or anyone else, of accepting this proposition. We may find that it, too, is good and thus get to GG*p*.

However, this is not being achieved automatically, cost-free, as it would be with a truth-functional property of propositions. It is dependent on a more complicated and demanding performance by Jones. It is easy to see why GGGG*p* would be unlikely and $G^{100!}p$ absurd. The latter requires a performance clearly beyond the range of human capacity. Iteration might be rescued here by appeal to the great preserver of bogus universals: the subjunctive conditional. Rather than waiting to evaluate Jones's performance, we could speak instead of how it would have rated had there been one. Starting from G*p*, we could ask ''If Jones had looked into the question whether he did a good job on the question whether *p*, would he have done a good job on that one? The subjunctive conditional answering in the affirmative would then serve as GG*p*: if Jones tackled the question whether G*p*, he would have done well. By this method we would seem to get indefinite iteration yielding propositions, true or false.

Nothing has been presented here to forestall such a maneuver, beyond appeal to the fact that indefinite iteration is obviously senseless so that anything that entails it has to be wrong. What would be needed is an argument against the existence of a relation of subjunctive conditionality independent of general laws and, in turn, against such notions as the ''possible world'' or ''maximal consistent set of propositions'' or ''maximal state of affairs'', which seem to offer a basis for such a relation. Platonism is strongly opposed to such Meinongian notions, but is often misrepresented as sympathetic to them. However, this cannot be argued here.

It can be important to trust that you have indeed encountered a property. The fact that such trust is absurd and fallacious about some series does not establish that it cannot be admirable about others. This is most important in the case of goodness. Our powers of recognizing this property are very limited. The border zones are extensive and varied and many seducers are waiting to take us there, to adjust our sense of

what is right to suit their purposes. Our application of evaluative terms in border zones is not best explained in terms of our recognizing the presence of a property expressed by the term. Some will take this as evidence that it is an illusion to think there is such a property as goodness. Perhaps this property need not be invoked in explaining our use of value terms. But it is crucial to evaluating such usage. It is not immodest to think that we can recognize this property in some very simple cases. And then we can opine that this property is present even in cases in which we admit we cannot be perfectly sure. We may be unsure when it is a great trouble to us not to know better. It is good to know what is good, but terrible not to know even that some things are good in spite of our not being able to reliably identify them. It may be hubris to choose this dialectical meaning in some cases. But we will still know what to ask for in our prayers.

We can know what we are inclined to say (as opposed to what we would say given only an abstract description) and we can wonder whether it is good to follow the inclination and be guided by consequences we attach to the application. We can give up claiming a property for the term as long as we can ask if the application is judicious. But then the property of goodness is there among the properties of language and usage. By grasping it we may be able to see when having those properties can be of value. This is just one way in which the form of the Good is the source of the intelligibility of being.[16]

REFERENCES

Boyd, Richard (1981) 'Materialism Without Reductionism: What Physicalism Does Not Entail', in Ned Block (ed.), *Readings in Philosophy of Psychology* (Cambridge), 67–107.

Cargile, James (1970) 'A Note on Iterated Knowings', *Analysis*, 30: 151–5.

Goodman Nelson (1976) *Languages of Art: An Approach to a Theory of Symbols*, 2nd edn. (Indianapolis).

Hart, W. D. (1992) 'Hat-Tricks and Heaps', *Philosophical Studies*, 33: 1–24.

Lazerowitz, Morris (1992) 'The Existence of Universals', in Andrew Schoedinger (ed.), *The Problem of Universals* (New Jersey), 135–5.

[16] Thanks are due to friends, students, teachers, and colleagues for helpful dialogue over many years. Hopefully, we are getting closer.

Locke, John (1959) *An Essay Concerning Human Understanding*, ed. Alexander C. Fraser (New York).
Shoemaker, Sydney (1997) 'Causality and Properties', in D. H. Mellor and Alex Oliver (eds.), *Properties* (Oxford), 228–54.
Quine, W. V. (1963) 'On What There Is', in *From a Logical Point of View* (New York).
Wittgenstein, Ludwig (1956) *Remarks on the Foundations of Mathematics* (Oxford).

This page intentionally left blank

3. Recent Debates about the A Priori

Hartry Field

1. BACKGROUND

At least from the time of the ancient Greeks, most philosophers have held that some of our knowledge is independent of experience, or ''a priori''. Indeed, a major tenet of the rationalist tradition in philosophy was that a great deal of our knowledge had this character: even Kant, a critic of some of the overblown claims of rationalism, thought that the structure of space could be known a priori, as could many of the fundamental principles of physics; and Hegel is reputed to have claimed to have deduced on a priori grounds that the number of planets is exactly five.

There was however a strong alternative tradition, empiricism, which was skeptical of our ability to know such things completely independent of experience. For the most part this tradition did not deny the existence of a priori knowledge altogether, since mathematics and logic and a few other things seemed knowable a priori; but it did try to drastically limit the scope of a priori knowledge, to what Hume called ''relations of ideas'' (as opposed to ''matters of fact'') and what came later to be called ''analytic'' (as opposed to ''synthetic'') truths. A priori knowledge of analytic truths was thought unpuzzling, because it seemed to admit a deflationary explanation: if mathematical claims just stated ''relations among our ideas'' rather than ''matters of fact'', our ability to know them independent of experience seemed unsurprising. So, up until the mid-twentieth century, a major tenet of the empiricism was that there can be no a priori knowledge of synthetic (non-analytic) truths.

But in a series of influential articles, W. V. Quine (1936, 1951*a*, 1951*b*) cast a skeptical eye on the manner in which the empiricists of his day were trying to explain a priori knowledge of logic and mathematics. His critiques led some (including Quine himself) to a full-blown empiricism in which there is no a priori knowledge at all, not even in logic and mathematics. Others (Bonjour 1998) were led in the

opposite direction, to a fuller-blooded rationalism: since logic and mathematics seem obviously a priori, and the empiricist attempts to explain this away seem dubious, we must conclude that the mind just does have the power to ascertain ''matters of fact'' independent of experience, perhaps by a faculty of rational intuition. Still others (Boghossian 2000; Peacocke 2000) have tried to base a priori knowledge on meaning in a subtler way than the empiricists did, so as to evade Quine's critique while avoiding the need for a full-blooded rationalism; and another strategy for accepting a priori knowledge while avoiding full-blooded rationalism will be mentioned below.

This is the cluster of issues to bear in mind in what follows.

2. A PRIORITY: WEAK AND STRONG, DOGMATIC AND UNDOGMATIC

We need to be a bit more precise about what 'a priori knowledge' means. Presumably someone knows that p only if p is true, she believes it, and she is *entitled* to believe it; the issue of a priority concerns the kind of entitlement that is in question. Roughly speaking, a priori entitlement is entitlement that is independent of experience.

But what is it for one's entitlement to be independent of experience? There are at least three issues here.

(i) Having the belief that p requires that we have the concepts involved in p. Experience is involved in the acquisition of concepts; if there is to be any realistic chance of there being a priori knowledge, experience involved merely in the acquisition of the concepts must ''not count''. Just how to allow it not to count is not entirely clear, especially given that learning concepts often involves the acquisition of information.

(ii) Proofs in logic and mathematics would seem to confer a priori entitlement if anything does. But long proofs need to be carefully checked, which normally involves looking at the written proof, and perhaps asking others to look for errors. Experience is clearly involved here, but this sort of experiential involvement must also ''not count''.

These first two issues will not be my concern here. They do point to the need of clarification, and it is certainly a respectable position that

adequate clarification is impossible to come by and that the concept of a priority is so hopelessly obscure that it ought to be simply abandoned. But let us make the working hypothesis that it is possible to clarify the concept in a way that doesn't either rule out a priority trivially or make it uninterestingly weak.

More important to the subsequent discussion is

(iii) 'Entitled independent of experience' is ambiguous. On a weak construal, to say that a person is entitled to believe that p, independent of experience, means only that she is entitled and *none of the experiences she's had* figure in making her so. On a strong construal, it also requires (roughly) that the fact that she *hasn't* had certain experiences plays no role in making her entitled: it requires that no experience she might come to have could defeat the entitlement.

The stronger notion is the more philosophically important: the philosophical interest of the claim that Euclidean geometry or classical logic is a priori would be much reduced if that claim were taken to be compatible with the claim that experience might undermine them. (When we come to consider the possibility of ''default entitlement'' in §§ 3 and 7, we'll see that the philosophical interest of weak a priority is indeed quite minimal.)

Why the parenthetical qualifier 'roughly' in stating the strong construal? Because of a problem analogous to that under (ii). Suppose we have carefully worked through a correct mathematical proof and thoroughly understand it. Still, as Kitcher 1983 (ch. 1) has observed, it is possible that we might acquire (misleading) evidence that we were suffering delusions every time we went through the proof, and/or that every respected mathematician regarded our ''proof'' as demonstrably fallacious. Few if any a priorists would deny that such experiences are possible and that they could undermine our entitlement, so we ought to rule them as ''not counting'' as regards strong a priority (or else stick to weak a priority). A rough stab at explaining why they shouldn't count—doubtless inadequate—is to put the empirical unrevisability requirement as follows: there is no possible empirical evidence against p which is ''direct'' as opposed to going via evidence of the reliability or unreliability of those who believe or disbelieve p. Whatever the merits of this, the problem of explaining why the experiences shouldn't count seems no worse than the analogous problems for weak a priority that arose under (i) and (ii).

Note that for p to be strongly a priori, it is not required that p be rationally unrevisable: if thought alone, unaided by evidence, could rationally undermine the belief in p, that has no bearing on the a priority of our entitlement to believe that p. (It needn't even undermine a priori *knowledge* that p, since the undermining thought could be rational but erroneous.) What is required for a priority is only that p be unrevisable *on empirical grounds*. But even this could mislead, for it is slightly ambiguous: it means, of course, that it is *possible* to rationally revise the belief on empirical grounds, but the question is whether we understand the possibility as ''genuine'' or ''merely epistemic''. The merely epistemic conception of possibility is the one we use when we say ''It is possible, though unlikely, that standard set theory might be inconsistent''. What we're saying is: this is something we can't totally rule out. By contrast, the ''genuine possibility'' that set theory is inconsistent would entail that set theory actually is inconsistent: for if set theory is consistent (i.e. it is not possible to derive an inconsistency in it), then it is necessarily consistent. Most of us believe set theory to be consistent, hence believe there to be no ''genuine possibility'' of its being inconsistent; but we are undogmatic in this belief, in that we believe that there are conceivable developments (e.g. the derivation of a contradiction within it) that would lead us to alter our opinions.

On the epistemic interpretation, strong a priority would involve the claim that there is no epistemic possibility whatever of revising our mathematics or logic on empirical grounds. Adhering to such an incredibly strong claim would seem like pure dogmatism: we may not be able to see how our mathematics or our logic could be revised on empirical grounds, but the track record of philosophers' pronouncements on epistemological matters is hardly good enough to inspire complete confidence that we might not be overlooking something. I take strong a priority in this dogmatic version to be indefensible: there's no way to *completely rule out* revising logic as the best way of accommodating, say, quantum mechanical evidence, even if we can't conceive the details of how such a revision would go.

But on the ''genuine possibility'' interpretation, no such dogmatism is required: the claim is merely that good standards of evidence wouldn't allow for empirical evidence against mathematics or logic, in any genuinely possible circumstances. Someone could believe this claim strongly, while retaining the fallibilist attitude that for this belief, as for all others, there is an epistemic possibility of someday rationally revising it.

3. PHYSICAL GEOMETRY

Why have philosophers often been skeptical of claims to a priority? In the case of claims about the number of planets, or even about the structure of physical space, it just seems obvious to many people that these aren't the sort of things that we could possibly be entitled to believe without evidence. But that is less an argument than an assertion of the doctrine to be argued.

Something closer to an argument can be given in the case of strong a priority (even of the undogmatic kind). Consider Euclidean geometry (viewed as a theory of physical space, which I believe is how it was almost universally viewed until at least the mid-nineteenth century). Despite the fact that none of its axioms is based on empirical evidence in any obvious way, still the system as a whole does have consequences that might be questioned on empirical grounds. An obvious example is that the surface area of a sphere is 4π times the square of the radius: it seems possible to imagine finding an object, verifying by repeated careful measurements that every point on its surface is indeed the same distance from a certain point (the center), and verifying by repeated careful measurements that the surface area was very different from 4π times the square of the radius (different by an amount far greater than could plausibly be attributed to systematic experimental error). Of course, we could explain this away by far-out hypotheses (systematic but undetectable expansion of our measuring rods, deceptive demons who made us misread the instruments, and so forth), but the possibility of saving a claim from empirical refutation by such far-out hypothesis is not normally thought to make that claim non-empirical, so why should it here?

Does this argument also go against the weak a priority of geometry? It would to someone who thought that you couldn't be entitled to believe a claim without having evidence for each of its empirical consequences, but that is generally implausible, and seems implausible even in the case at hand since most of us would think that Euclid was eminently entitled to his geometric beliefs without having made careful measurements of space.

Indeed, it isn't clear that we should doubt the weak a priority of geometry. It is not unreasonable to think that evolution might have endowed us with a tendency to believe Euclidean claims, barring evidence to the contrary, even in absence of arguments for them; and it

isn't clear why, if this is true, such beliefs shouldn't count as ones we're entitled to when there is no evidence against them. (And why they shouldn't count as known, if in addition they are true.) If so, this would seem to be a case of weak a priority. Perhaps claims to *weak* a priority shouldn't be regarded as such a big deal.

The suggestion here—which will play an important role in § 7—is that some of our beliefs count as entitled *by default*: we need no positive reason for them, experiential or otherwise, to count as entitled to believe them. If these beliefs are empirically undeterminable they won't be a priori in the interesting strong sense, but they are trivial cases of the weak a priori.

''But where does this default entitlement come from?'' It needn't ''come from'' anywhere: entitlement isn't a fluid whose creation needs explanation. Probably the best view is that we simply have an attitude of regarding some beliefs as entitled under some circumstances, others not; and we regard some of them as entitled in absence of evidence for or against, even though there might someday be evidence that disconfirms them. And to put it crudely, there are no ''facts about entitlement'', there is nothing beyond these attitudes; we can evaluate attitudes as good or bad, but such evaluation is not a ''factual'' enterprise.

4. LOGIC, MATHEMATICS, AND METHODOLOGY

Though physical geometry seems not to be a domain of strongly a priori knowledge, there are other candidates that fare better. Perhaps the best candidates are logic and pure mathematics. The reason for thinking of these as strongly a priori is evident: they don't seem to be based on empirical evidence, and it is hard to see how empirical evidence could undermine them. What possible empirical evidence could undermine the logical belief that *if snails exist then snails exist*, in the way that evidence of spatial measurements could and did undermine Euclidean geometry as a theory of physical space? I'll return to this in a moment.

Another case worth mentioning is empirical methodology itself: there are reasons for thinking that empirical methodology is strongly a priori, in the sense that its rules are rationally employable independent of evidence and can't be undermined by evidence. The impossibility of undermining evidence may be less evident in the case of empirical methodology than in the case of logic and mathematics. Presumably

our empirical methodology includes a bias for simplicity. We recognize that in so far as we can account for all past and present observations by our present body of theory T, we could account for it equally well by an alternative theory T* according to which T holds until 1 January 2004, after which Aristotelian physics, Lamarckian biology, etc., take over. Why do we rule out T*, and base our predictions instead on the approximate truth of T? We certainly have no evidence favoring T over T* (since they yield exactly the same probabilities for everything in the present and past), so presumably it's that T is a vastly simpler way of accommodating our evidence than is T*. But now it might seem that our methodology of choosing the simpler is empirically revisable (either by revising the principle ''choose the simpler'' or by revising the simplicity judgements that give this slogan its content). Suppose we had evidence that in each past year on New Year's Day, the laws of nature drastically changed; that would seem like good inductive evidence that they'd change on New Year's Day in 2004 too. Doesn't this show that our empirical methodology (our system of simplicity judgements and the methodological import we give them) is itself empirically revisable?

No, it doesn't show this at all. What it shows is only that we regard theories T** according to which the laws of nature change every year as more plausible than corresponding theories T^{***} according to which the laws change every year until 2004, but don't change then. It seems that we have two pre-existing biases: one for T over T*, which licenses a belief that the laws won't change in 2004 given evidence that they haven't changed in the past; the other for T** over T^{***}, which would license a belief that the laws will change in 2004 were we to be given evidence that they have changed each year in the past. So the fact that the laws of nature haven't changed drastically in the past is indeed inductive evidence that they won't change drastically in 2004; but this fact is based on a fixed bias (for T over T* and for T** over T***) which there is no obvious way to undermine by empirical evidence.

In the case of mathematics, it is hard to come up with even a prima-facie case for the evidence-based revision of an accepted theory (say, the theory of real numbers). We could, to be sure, imagine discovering that the structure of physical space was not accurately describable (even locally) as a product space of the real numbers; perhaps physical space is discrete, or countable, or whatever. But this would surely not be best thought of as showing that the theory of real numbers is wrong, but only that that theory is inapplicable to physical space.

The case of logic is more interesting, for there have been proposals to revise logic in order to solve certain anomalies in quantum mechanics, and the proposals have at least some prima-facie attractiveness even though no clear sketch has been given of just how the development of quantum mechanics in such a logic would go. As noted before, it seems dogmatic to insist in advance that there is no epistemic possibility that a case for such a revision could ever be made compelling; on the other hand, we certainly do not now understand even what it would be like to use such revised logics as our sole logic, let alone understand just how the case for switching from classical logic to the revised logic would be rationally argued, so these proposals do not yet constitute an argument that there is a genuine possibility of rationally revising logic on the basis of quantum mechanical considerations.

But suppose that such a proposal could be worked out in detail, and could lead to a rational revision in logic. Would that revision be on empirical grounds? It's hard to say: perhaps it would be a case showing logic to be revisable on the basis of quantum mechanical evidence, but perhaps it would be a case where quantum mechanical considerations pointed up the need for a conceptual revision that could have been made independent of evidence. (Consider someone who is led to a logic that allows for negative existence claims involving names by the empirical discovery that there is no Santa Claus.) It seems idle to speculate whether, were it possible to work out the details of the case, the revision would be empirical or conceptual: that's rather like the question of what features an inconsistency in set theory is likely to have should one be discovered. In each case there's no way to answer the question, absent a clearer idea of what the alleged possibility might be like.

If it does make sense to suppose that logic might be rationally revised on empirical grounds, that might give reason to think that mathematics could too: after all, proof in mathematics goes via logic! To my mind this is the only serious possibility for revising mathematics empirically. But even here, it is not obvious that we would have a case for an empirical revision of mathematics, for not all revisions of logic would be relevant to mathematics. Consider proposals to revise logic on non-empirical grounds, for example, the proposal to abandon the law of excluded middle (*B or not B*) as a general principle so as to deal with the Liar paradox. Such proposals allow keeping all instances of excluded middle that don't involve 'true' (or other predicates that give rise to analogous pathologies), and in particular excluded middle can be

assumed for all mathematical sentences (though it may be demoted to the status of a *non-logical* axiom schema). No revision of mathematics need ensue. The same point would seem to arise for a revision of logic on empirical grounds, if that is possible: if experience tells us that the distributive law doesn't apply generally, still it may (not as a matter of logic but for other reasons) apply to many special objects (e.g. those that can't undergo quantum superpositions), and mathematical objects seem like very good candidates for being among those to which the distributive law would still apply.

5. THE BENACERRAF PROBLEM FOR MATHEMATICS

Even in cases, like mathematics, where strong a priori knowledge (of an undogmatic sort) seems highly plausible, there are puzzles about it. The most famous one was articulated by Benacerraf (1973). (He raised it not as a problem specifically about *a priori* knowledge of mathematics, but about *any* sort of knowledge of mathematics; but those who take knowledge of mathematics to be empirical, e.g. Hart (1996), often claim that by doing so they have a way around the argument.)

I will not consider Benacerraf's own formulation—it relies on a causal theory of knowledge that simply seems inapplicable to a priori knowledge—but rather, will try to capture its general spirit. The key point, I think, is that our belief in a theory should be undermined if the theory requires that it would be a huge coincidence if what we believed about its subject matter were correct. But mathematical theories, taken at face value, postulate mathematical objects that are mind-independent and bear no causal or spatiotemporal relations to us, or any other kinds of relations to us that would explain why our beliefs about them tend to be correct; it seems hard to give any account of our beliefs about these mathematical objects that doesn't make the correctness of the beliefs a huge coincidence.

Of course, no one would propose that we reject mathematics on the basis of such arguments; Benacerraf's point was simply to raise a puzzle about why not. There are various answers to this that seem satisfactory. Some of these (e.g. Field 1989; Yablo 2000) involve fictionalism about mathematics: on these it is simply not the function of mathematical theories to be true, so the puzzle just doesn't arise. (So we have no knowledge at all of mathematics, a priori or otherwise.) Others

(Balaguer 1995; Putnam 1980; perhaps Carnap 1950) solve the problem by articulating views on which though mathematical objects are mind-independent, any view we had had of them would have been correct. (In Balaguer's case that's because the mathematical universe is so plenitudinous that, whatever view we had had of it, there is some part of the mathematical universe of which it would have been true; and we are talking about whichever part makes our theory true.[1]) Unlike fictionalist views, these views allow for a priori knowledge in mathematics, and unlike more standard Platonist views, they seem to give an intelligible explanation of it.

Those who argue that Benacerraf's problem doesn't arise for the empiricist seem in considerably worse shape: although they say that empirical evidence bears on mathematical claims, they have not offered (and could not easily offer) even a clear sketch of how the experiences that allegedly might overturn our mathematics are reliable symptoms of the facts about mathematical objects. The problem isn't the indirectness of the evidence, or the fact that its being evidence depends on theoretical assumptions: evidence for black holes shares these characteristics, but raises no Benacerraf problem because there's a straightforward causal story that explains the correlation between the facts about black holes and the evidence for them. In the mathematical case such a story is lacking, which seems embarrassing to an empiricist view.

6. A BENACERRAF-LIKE PROBLEM FOR LOGIC?

Many philosophers think that to whatever extent there is a Benacerraf problem for mathematics, there is also one for logic: the fact that mathematics deals with special objects and logic doesn't is, in their view, an irrelevant difference. At first blush this seems reasonable: the worry would seem to be that there is no obvious explanation of how our logical beliefs can depend on the logical facts, and this should engender skepticism that they do depend on the logical facts. It would seem that only a huge coincidence could have made our logical beliefs accurately reflect the logical facts.

[1] In Putnam's case, it is because there is no constraint on the extension of our mathematical predicates other than that it be such as to make our mathematical beliefs true; so that they are bound to be true as long as there are infinitely many mathematical objects. Carnap's view is open to more than one interpretation.

This isn't really an optimal formulation of the problem about logic. After all, logic seems primarily concerned not with ''logical beliefs'' but with *inferences*.[2] Inferences connect claims (not primarily about logic) to other claims (not primarily about logic); they involve *conditional commitments*, which are distinct from beliefs. So ''logical beliefs'' don't enter the picture in any very direct way.[3] It would be better, then, to put the Benacerraf problem in terms of the lack of an explanation of how our logical *inferences* depend on the logical facts. And here we should presumably take the logical facts to involve meta-properties of the inference: for example, the fact that the inference is (necessarily) truth-preserving. To say that the inference from *A* to *A or B* is (necessarily) truth-preserving just means that (necessarily) if *A* is true then so is *A or B*; on a minimal notion of truth, that's just equivalent to the claim that (necessarily) *if A then either A or B*.

So the way to put a Benacerraf problem for logic is something like the following:

(i) it seems in principle impossible to explain such things as how our acceptance of the inference from *A* to *A or B* depends on the logical fact that necessarily *if A then either A or B*;[4]

(ii) without such an explanation, to believe in a correlation between our accepting the inferences we do and the logical facts requires belief in a massive coincidence;

(iii) the need to believe in such a massive coincidence undermines the belief in the correlation, which in turn should undermine our acceptance of the inference.

[2] Harman (1973) questions this, on the basis of the fact that when an argument leads us from antecedently believed premises to an antecedently disbelieved conclusion we may reject a premise rather than accept the conclusion. The view of inference in the next sentence is designed to accommodate his point.

[3] They may enter indirectly, in one of two ways. First, in classical logic and most of the popular alternatives to it, some claims are assertible without premises: e.g. in classical logic any claim of form *A or not A*, and in most logics any claim of form *If A, then either A or B*. If we employ a logic of this sort, these will be ''logical beliefs''. Second, we can take ''logical beliefs'' to mean meta-claims about the inferences involved: e.g. the claim that the inference is valid, or necessarily truth-preserving, or necessarily preserving of some other semantic status. But in either case, the logical beliefs seem to have a status secondary to the inferential behavior.

[4] There are logics in which one can accept the inference from $X_1, \ldots, X_n$ to *Y* without accepting the claim that *if* X_1 *and* ... *and* X_n *then Y*. In those logics, one does not accept the claim that the inferences in the logic are truth-preserving on a minimal notion of truth, and so if a Benacerraf problem can be raised at all it must be raised in a different way.

At first blush this may seem as compelling as the corresponding problem about mathematical objects is (on the naive Platonist picture for which the Benacerraf problem is genuinely a problem).

At second blush, the logical case seems very different from the mathematical case. For in the logical case, isn't it clear that evolution provides the answer? Isn't it clear that the correct logical beliefs are selected for (i.e. creatures whose logical beliefs didn't reflect the logical facts would die out)? In the mathematical case, on the other hand, it is hard to see how such selection could work: given that mathematical objects have no causal, spatio-temporal, etc. relations to us, what mechanisms could select for correctness of beliefs in that case?

At third blush, though, the evolutionary explanation is not obviously satisfactory in the logical case either. For in the mathematical case, it isn't in principle problematic to see how *a particular* mathematical theory *T* might have been selected for: perhaps belief in *T* leads to a subtle odor which our predators found repugnant. What is problematic is to figure out the connection between what is selected for and the actual mathematical facts. Doesn't this affect the situation for logic too? We could easily tell some sort of story (at least as plausible as the one about repugnant odors!) on which there were selection pressures for the acceptance of classical logic. But what we need is a story on which there is a selection pressure for acceptance of *the correct logic, whichever one that happens to be*. And it isn't so obvious that we can do that, so the Benacerraf problem for logic seems to remain.

At fourth blush (Field 1998), one might question the distinction between

(i) selection pressure for acceptance of a given logic, which is in fact correct,

and

(ii) selection pressure for acceptance of the correct logic, whichever one that happens to be.

In the mathematical case, such a distinction seems quite clear: we can see that in the odor story the mathematical facts themselves played no role in our survival (it isn't as if they had a relevant role in producing the odor), so in this case there is no doubt that the selection pressure was for the acceptance of a particular theory rather than for whichever one is true. But part of what makes this clear is that we can assume, with at least some degree of clarity, a world without mathematical objects, or a

world in which the particular theory *T* of them that we happen to believe in isn't true; and with ordinary logic we can then argue that the belief in *T* would still produce odors, so that the theory selected for would be a false one. But how are we to argue what would be selected for in a world with an alternative logic? We would apparently need to conduct the argument in the alternative logic in question, and we have so little idea how to do this that the counterfactual begins to look nonsensical. This casts serious doubt on the intelligibility of the distinction between (i) and (ii).

If that and similar distinctions really are unintelligible, that may itself provide an answer to the Benacerraf problem for logic, though not an evolutionary one (Field 1998). The Benacerraf problem in mathematics or logic seems to arise from the thought that we would have had exactly the same mathematical or logical beliefs, even if the mathematical or logical facts were different; because of this, it can only be a coincidence if our mathematical or logical beliefs are right, and this undermines those beliefs. In the mathematical case there is a reasonably clear content (at least prima facie) to the thought that we would have had exactly the same mathematical beliefs even if the mathematical facts were different; that's what gives the Benacerraf problem its initial bite in the mathematical case. But in the logical case, we have no idea how to determine what we would have believed had the logical facts been different: reasoning about what our beliefs would be in alternative circumstances requires logic, and if we contemplate a radically altered logic we have no idea how to conduct the reasoning. This seems to undermine the intelligibility of the counterfactual (about what we would have believed given different logical facts); in which case we have undermined, not just the evolutionary solution to the Benacerraf problem for logic, but the problem itself.

7. JUSTIFICATION, DISAGREEMENT, AND MEANING

Many of our beliefs and inferential rules in mathematics, logic, and methodology can be argued for from more basic beliefs and rules, without any circularity. But this is not so for the most basic beliefs and rules: we must be, in a sense, entitled to them by default. At the end of § 3 it was suggested that we don't have to regard our being default-entitled to them as a mysterious metaphysical phenomenon: it's

basically just that we *regard it as* legitimate to have these beliefs and employ these rules, even in the absence of argument for them, and that we have no other commitments that entail that we should not so regard them. (Of course, there are things we can say about *why* we regard it as legitimate to have these beliefs and employ these rules, and why anyone who didn't would be worse off; but the things we can say would be disputed by anyone who didn't have those beliefs and employ those rules, so the justification is circular. The circularity is broken by our attitudes—by what we *regard as* legitimate. See Field 2000 for more details.)

Many philosophers think more needs to be said to explain default-entitlement: they think that the only way we can be entitled to anything is for there to be some ''source'' for the entitlement, and since basic features of our logic and empirical methodology and perhaps mathematics can't have their source in a non-circular argument for them, they must have some other kind of source. One possibility (Boghossian 2000; Peacocke 2000) is that the meanings of our concepts serve as the desired source of entitlement.

At least in the case of logic and of empirical methodology, a prima-facie reason for thinking a source of entitlement needed is the possibility of alternative views that are in genuine conflict. The possibility of genuine conflict is clear in the case of empirical methodology: our broadly inductive methodology conflicts with counterinductive methodologies, and with skeptical methodologies that don't license the belief in anything not yet observed, and with innumerable methodologies that while broadly inductive also differ in the extent of the conclusions licensed about certain matters. There seems to be an issue as to whose empirical methodology is *more reasonable*. We presumably think ours the more reasonable, but they think the same of theirs; if ours *really is* more reasonable, doesn't there need to be a source of this reasonableness? Doesn't there need to be some kind of non-question-begging justification, even if in a sense of justification in which justifications needn't be arguments?

In the case of mathematics there may be no such genuine conflict between alternative theories (at least when the alternative theories are not based on different logical views): it's natural to think that different mathematical theories, if both consistent, are simply about different subjects. (That's why the pluralist views of Balaguer and Putnam, cited earlier, are as plausible as they are.) Because of this, the need for

justification (other than justification of the consistency of the theory) doesn't seem as pressing in the mathematical case. Or maybe, instead of lessening the need for justification, it means that the justification for consistent mathematical theories comes relatively cheap: by the purely logical knowledge that the theory is consistent. One way to develop this idea is to say that the axioms implicitly define the mathematical terms, and that consistent implicit definition in mathematics guarantees truth, so that only justification of the consistency of the theory is required.

But logic seems more like inductive methodology than like mathematics in this regard. In the first place, an implicit definition approach seems to face a serious limitation in the case of logic: it is only *consistent* implicit definition that could with any plausibility be held to guarantee truth, so we need an antecedent notion of consistency not generated by implicit definition; and what justifies a belief about consistency? (Admittedly, the notion of consistency required here may be one on which proponents of different logics may agree, so if this were the only point to be made it might seem that the implicit definition strategy could at least serve as a justification of the parts of logic about which controversies are likely.)

A more fundamental point is that those who advocate the use of alternative logics (and advocate them as more than just algebras for dealing with special subjects, but as systems for general reasoning) seem to be in genuine disagreement with us. There seems to be an issue as to which view is right (or at the very least, as to which is better); one that can't be removed by simply saying ''They're using their concepts, we're using ours''.

There are cases where this isn't so: for instance, someone might agree with us that there's no way for an argument from *A and not A* to *B* not to be truth-preserving, but call the argument invalid nonetheless *simply because it fails to respect some relevance condition that she imposes on consequence*. Here the disagreement seems to be a purely verbal one about the meaning of 'consequence'. But I take such cases not to be the interesting ones. What would be interesting is if someone rejected the rule because she thought it wasn't truth-preserving: that's the view of ''dialetheists'' (Priest 1998), who think that some claims of form *A and not A* are true. Dialetheists do seem to be in *genuine* (not merely verbal) disagreement with advocates of classical logic. So do ''fuzzy logicians'', who refuse to accept ''Either Harry is bald or he isn't'' in cases where Harry seems a borderline case. (''Fuzzy logic'' can be regarded as a

weakening of classical logic; it yields full classical logic when the law of excluded middle, *B or not B*, is added.[5])

Admittedly, locating what it is that proponents of different logics disagree about is tricky. For instance, the advocate of classical logic and the ''fuzzy logician'' both agree that full classical logic, including excluded middle, is valid with respect to the standard two-valued semantics; they also agree that fuzzy logic is valid with respect to the Lukasiewicz continuum-valued semantics and that classical logic isn't. Where then do they disagree? Do they disagree as to whether instances of excluded middle are true? Not really: on a minimal notion of truth, a fuzzy logician won't deny the truth of any instance of excluded middle, he'll just refrain from asserting some. Do they disagree as to whether instances of excluded middle are *necessarily* true? If that's all we can say, it's hard to see why the distinction isn't verbal: the fuzzy logician might just be employing a more restrictive notion of necessity. I think in the end the only way to make sense of the distinction is in terms of the laws they take to govern rational belief: for example, the fuzzy logician is willing to tolerate having a low degree of belief in instances of excluded middle, the classical logician isn't.

But again: if we are entitled to take a stand one way or the other on this (either following the fuzzy logician in tolerating a low degree of belief in excluded middle, or following the classical logician in not tolerating it and using the assumption of excluded middle in reasoning), mustn't there be a source for this entitlement? But what can it be?

It is sometimes claimed that meaning provides such entitlement. There are two conceptions of meaning one might invoke: truth-theoretic and inferential role. An advocate of the law of excluded middle might ''justify'' this using a truth-theoretic conception of meaning as follows:

> If *B* is true then *B or not B* is certainly true. And if *B* is not true then *not B* is true, so again *B or not B* is true. So either way, *B or not B* is true.

But this is grossly circular: the 'So either way' disguises a use of excluded middle at the meta-level, that is, it assumes that *B* is either

[5] Of course, if we take ''logic'' to include attributions of logical truth and their denials, then fuzzy logic is no weakening of classical logic: it conflicts with classical logic in claiming that instances of excluded middle are not logical truths.

true or not true. The fact is that the advocate of classical logic and of fuzzy logic can agree on the same compositional rules about truth; whether these laws make all instances of excluded middle come out true depend on whether one assumes excluded middle. A similar point holds for inferential rules, like modus ponens or the inference from *A and not A* to *B*: the truth rules guarantee that the inference is truth-preserving *if you assume a logic that employs the rule*, but don't guarantee this otherwise. (Note that the circularity for inferential rules seems no less noxious than that for beliefs, contrary to some proponents of ''rule-circular justifications''.)

The way that an inferential semantics would provide a justification or source of entitlement is different: here the claim would be (i) that the acceptance of certain logical beliefs or inferences is central to the meanings of the connectives, and (ii) that this somehow guarantees the legitimacy of those beliefs or inferences.

If this is to help with our example, the law of excluded middle must be one of those that are central to the meanings of 'or' and 'not'. Moreover, for (i) to support (ii) it must be interpreted to require that any alteration of the beliefs and inferences that are central to the meaning of the connective engenders a change in the meaning of the connective. On this interpretation, (i) is somewhat questionable: certainly a classical logician would have no better translation of a fuzzy logician's 'not' or 'or' into his own idiolect than the homophonic translations 'not' and 'or'. At any rate, if this is a change of meaning, it is not what Putnam (1969) called a ''mere change of meaning'', a mere relabeling: rather, the fuzzy logician would have to be seen as regarding the use of connectives with the classical meanings as *illegitimate*, and substituting new connectives that her opponent takes to be illegitimate in their place.

In any case, the key issue is (ii): why should the fact, if it is one, that certain beliefs or inferences are integral to the meaning of a concept show that those principles are correct? Why should the fact, if it is one, that abandoning those beliefs or inferences would require a change of meaning show that we shouldn't abandon those beliefs or inferences? Maybe the meaning we've attached to these terms is a bad one that is irremediably bound up with error, and truth can only be achieved by abandoning those meanings in favor of different ones (that resemble them in key respects but avoid the irremediable error).

There is reason to think that this must be a possibility: in earlier days if not now, the principles of the naive theory of truth were probably

central to the meaning of the term 'true' and the principles of classical logic central to the meaning of the connectives; but we know now that we can't consistently maintain *both* naive truth theory and classical logic, so at least some of the meanings we attached to our terms *must* have been bound up with error.

I doubt, then, that the appeal to the meaning of logical terms really serves the justificatory purpose to which some have tried to put it. It's worth remarking that any argument for thinking that we need a source of entitlement for our basic logical principles would seem to be a special case of a more general argument that we need a source of entitlement for all of our basic methodological principles, for instance our inductive rules. There, too, alternative rules are possible: not just counterinductive rules, but alternative rules with a broadly inductive character yet significantly different in details. (They might, for instance, allow for more rapid inductions to the next instance, and slower inductions to generalizations.) Here too it seems impossible to straightforwardly argue for one inductive rule over the other; and here the idea that one rule can be validated over another by being integral to the meaning of some of our concepts seems even less promising.

What, then, is involved in justifying a logic, or an inductive policy? To repeat two points:

(A) Our entitlement to use a logic or inductive policy can't depend on our having an argument for it; we are entitled ''by default''.

(B) Nor need our entitlement depend on there being some kind of justification other than argument (''source of entitlement''). The entitlement doesn't ''flow out of'' anything; in saying that we're default-entitled to our logic and methodology, I'm merely expressing an attitude of approval toward the use of the logic or methodology even by those who have no arguments on their behalf.

But to add a new point:

(C) There is still room for justification: questions of justification can arise when considerations are advanced *against* our logic or methodology. For instance, it is certainly possible to argue (whether persuasively or not, I won't here consider) that classical logic runs into trouble in dealing with certain domains, such as vagueness or the semantic paradoxes; and defending classical logic against such arguments is one form of justification. Positing

> our default entitlement to, say, the rules of classical logic by no means makes classical logic sacrosanct, it merely allows classical reasoning to be legitimate until a world view sufficient for reasonable debates about the principles of classical reasoning has been built.

There are puzzles about how debates about logic and methodology are ultimately to be conducted, puzzles that are beyond the scope of the present article. I think that the debates involve quite holistic considerations: the consequences of changing logical opinions, for example, about excluded middle, can be far-reaching, and we need to look at quite diverse consequences of the change and decide whether the benefits of the change to our overall world-view would outweigh the costs. As Quine (1951*a*) pointed out in response to Carnap, the fact that such debates are pragmatic does not preclude them from being *factual*: all high-level factual debates are pragmatic in this sense. Indeed, in the case of logic it is hard to see how such debates could be regarded as anything but factual. But the fact that debates about logic and methodology are holistic and pragmatic does not show that such debates are in any way *empirical*; and as argued in § 4, it is very hard to imagine how empirical evidence could be deemed relevant to such debates.[6]

REFERENCES AND FURTHER READING

Balaguer, M. (1995) 'A Platonist Epistemology', *Synthese*, 103: 303–25.

Benacerraf, P. (1973) 'Mathematical Truth', in Benacerraf and Putnam (1983: 403–20).

——and H. Putnam (1983) *Philosophy of Mathematics: Selected Readings*, 2nd edn. (Cambridge: Cambridge University Press).

Boghossian, P. (2000) 'Knowledge of Logic', in Boghossian and Peacocke (2000: 229–54).

——and C. Peacocke (2000) *New Essays on the A Priori* (Oxford: Oxford University Press).

BonJour, L. (1998) *In Defense of Pure Reason* (Cambridge: Cambridge University Press).

Carnap, R. (1950) 'Empiricism, Semantics and Ontology', in Benacerraf and Putnam (1983: 241–57).

[6] I thank Paul Boghossian, Paul Horwich, Chris Peacocke, and Stephen Schiffer for useful discussions.

Field, H. (1989) *Realism, Mathematics and Modality* (Oxford: Blackwell).
—— (1998) 'Epistemological Nonfactualism and the A Prioricity of Logic', *Philosophical Studies*, 92: 1–24.
—— (2000) 'A Priority as an Evaluative Notion', in Boghossian and Peacocke (2000: 117–49).
Harman, G. (1973) *Thought* (Princeton: Princeton University Press).
Hart, W. D. (1996) 'Introduction' to Hart (ed.), *The Philosophy of Mathematics* (Oxford: Oxford University Press).
Kitcher, P. (1983) *The Nature of Mathematical Knowledge* (Oxford: Oxford University Press).
Peacocke, C. (2000) 'Explaining the A Priori: The Program of Moderate Rationalism', in Boghossian and Peacocke (2000: 255–85).
Priest, G. (1998) 'What is So Bad about Contradictions?', *Journal of Philosophy*, 95: 410–26.
Putnam, H. (1969) 'Is Logic Empirical?', in R. Cohen and M. Wartofsky (eds.), *Boston Studies in the Philosophy of Science*, 5: 199–215.
—— (1980) 'Models and Reality', in Benacerraf and Putnam (1983: 421–44).
Quine, W. V. (1936) 'Truth by Convention', in Benacerraf and Putnam (1983: 329–54).
—— (1951*a*) 'Carnap on Logical Truth', in Benacerraf and Putnam (1983: 355–76).
—— (1951*b*) 'Two Dogmas of Empiricism', *Philosophical Review*, 60: 20–43.
Yablo, S. (2000) 'Apriority and Existence', in Boghossian and Peacocke (2000: 197–228).

4. Our Knowledge of Mathematical Objects

Kit Fine

I have recently been attempting to provide a new approach to the philosophy of mathematics, which I call 'proceduralism' or 'procedural postulationism'.[1] It shares with traditional forms of postulationism, advocated by Hilbert (1930) and Poincaré (1952), the belief that the existence of mathematical objects and the truth of mathematical propositions are to be seen as the product of postulation. But it takes a very different view of what postulation is. For it takes the postulates from which mathematics is derived to be imperatival, rather than indicative, in form; what are postulated are not propositions true in a given mathematical domain, but procedures for the construction of that domain.

This difference over the status of the posulates has enormous repercussions for the development and significance of such a view. The philosophy of mathematics is faced with certain fundamental problems. How are we capable of acquiring an understanding of mathematical terms? How do we secure reference to mathematical objects? What is the nature of these objects? Do they exist independently of us or are they somehow the products of our minds? What accounts for the possibility of applying mathematics to the real world? And how are we able of acquire knowledge of mathematical truths? The procedural form of postulationism, in contrast to the propositional form, is capable of providing plausible answers to each of these questions. By going procedural, we convert a view that is beset with pitfalls to one that is worthy of serious consideration.

In what follows I shall focus on the last question concerning our knowledge of mathematics (although this will inevitably involve the other questions). I do this not because this question is the most

[1] First broached in Fine (2002: 36, 56, 100).

interesting or even because it provides the most convincing illustration of the value of our approach, but because it helps to bring out what is most distinctive—and also most problematic—about the approach. If one can go along with what it recommends in this particular case, then one is well on the way to accepting the view in its entirety.

As with the 'big three' traditional approaches to the philosophy of mathematics—logicism, formalism, and intuitionism—the present approach rests upon a certain technical program within the foundation of mathematics. It attempts to derive the whole of mathematics—or a significant part thereof—within the limitations imposed by its underlying philosophy. Since the viability of the underlying philosophical view largely depends upon the possibility of carrying out such a program, it will be helpful to give a sketch—if only in the barest form—of what the program is and of how it is to be executed. I hope elsewhere to provide a much more extensive development of the view in both its philosophical and technical aspects.

1. THE LANGUAGE AND LOGIC OF POSTULATION

Under standard forms of postulationism, what is postulated is the truth of a proposition. Thus it is something that might be expressed by means of an indicative sentence, such as 'every number has a successor'. A mathematical theory is then given through a suitable set of indicative sentences or 'axioms'. Under our approach, by contrast, what is postulated, or prescribed, is a procedure for the construction of the domain. These procedures are more appropriately signified, not by indicative sentences, but by imperatives or 'rules'; and a mathematical theory is to be given by a suitable set of rules for the construction of its domain.[2]

We might compare the rules, as so conceived, to computer programs. Just as a computer program prescribes a set of instructions that govern the state of a machine, so a postulational rule, for us, will prescribe a set of instructions that govern the composition of the mathematical domain; and just as the instructions specified by a computer program will tell us how to go from one state of a machine to another, so the instructions specified by a rule will tell us how to go from one 'state'

[2] It has been pointed out to me that 'postulation' is not altogether an appropriate term for what I have in mind. But I know of no better term.

or composition of the mathematical domain to another (one that, in fact, is always an expansion of the initial state). Indeed, so arresting is this analogy that it will be helpful to pretend that we have a genie at our disposal who automatically attempts to execute any procedure that we might lay down. The rules are then the means by which we tell the genie what to do.

If we are to make the above idea precise then we need to specify a 'programming language' within which the instructions to the genie might be stated. This language, at least in its most basic incarnation, is very simply described. The programs or rules are of two kinds, *simple* and *complex*: simple rules are not built up from other rules; complex rules are. There is only one form of simple rule:

(i) *Introduction* !x.C(x);

and it may be read:

introduce an object x conforming to the condition C(x).

In response to the prescription of such a rule, the genie will introduce an object into the domain that conforms to the condition C(x) if there is not already such an object in the domain and otherwise he will do nothing. For example, in response to the prescription:

!x.$\forall$y(x > y),

he will introduce an object that stands in the >-relation to the pre-existing objects in the domain (unless such an object already exists).

There are four kinds of complex rule:

(ii) *Composition*: Where β and γ are rules, then so is $\beta;\gamma$. We may read $\beta;\gamma$ as: do β and then do γ; and $\beta;\gamma$ is to be executed by first executing β and then executing γ.

(iii) *Conditionality*: Where β is a rule and A an indicative sentence, then $A \rightarrow \beta$ is also a rule. We may read $A \rightarrow \beta$ as: if A then do β. How $A \rightarrow \beta$ is executed depends upon whether or not A is true: if A is true, $A \rightarrow \beta$ is executed by executing β; if A is false, then $A \rightarrow \beta$ is executed by doing nothing.

(iv) *Universality*: Where $\beta(x)$ is a rule, then so is $\forall x\beta(x)$. We may read $\forall x\beta(x)$ as: do $\beta(x)$ for each x; and $\forall x\beta(x)$ is executed by simultaneously executing $\beta(x_1)$, $\beta(x_2)$, $\beta(x_3)$, ..., for each value x_1, x_2, x_3 , ...of x (within the given

domain). Similarly for the universal rule $\forall F\beta(F)$, where F is a second-order variable ranging over any plurality of objects from the given domain.

(v) *Iteration*: Where β is a rule, then so is β^*. We may read β^* as: iterate β; and β^* is executed by executing β, then executing β again, and so on for any finite number of times.

All of the postulational rules from our simple language may be obtained by starting with the simple rules and then applying the various clauses stated above for the formation of complex rules. Each simple rule prescribes a procedure for introducing at most one new object into the domain, suitably related to itself and to pre-existing objects. A complex rule prescribes multiple applications of these simple rules, performed—either successively or simultaneously—to yield more and more complex extensions of the given domain. Thus the only simple procedure or 'act' that the genie ever performs is to introduce a single new object into the domain; everything else that he does is a vast iteration, in sequential or simultaneous fashion, of these simple acts of introduction.

Let us see how our simple postulational language might be used to prescribe procedures for the construction of various familiar mathematical domains. We consider two examples: arithmetic[3] and a version of set theory (to be exact: cumulative type theory).

Arithmetic

Read Nx as 'x is a number' and Syx as 'y is the successor of x'. Rules for arithmetical domain are then given by:

ZERO: $!x.Nx$
SUCCESSOR: $\forall x(Nx \rightarrow !y.(Ny \,\&\, Syx))$
NUMBER: ZERO; SUCCESSOR*.

ZERO says: introduce an object x that is a number. In application to a domain that does not contain a number, it therefore introduces a new object that is a number. We may take this new object to be 0. (Its uniqueness will be guaranteed under our theory by the fact that it is

[3] This is arithmetic à la Dedekind. It is also possible to provide a postulational treatment of arithmetic à la Frege. I make no stand here as to which approach corresponds more closely to our intuitive conception of number.

the first number to be introduced into the domain through procedural postulation. But the general question of the identity of such objects is not one that we shall pursue.) SUCCESSOR says: for each object x in the domain that is a number, introduce a number y that is the successor of x (unless such an object already exists). NUMBER (which we might pronounce 'Let there be numbers!') says: first perform ZERO, i.e. introduce 0, and then keep on introducing the successor of numbers that do not already have a successor.

It should be clear that in response to NUMBER, as applied to a domain that contains no numbers, the genie should introduce an ω-progression of numbers 0, 1, 2, . . . , with each but the first standing in the successor-relation to its immediate predecessor. However, in order to establish this result 'formally', we need to make two general assumptions about how the genie complies with a rule. In the first place, he follows a policy of conservativity; in extending a given domain, he never makes any internal change to the domain itself. Thus, once he has introduced the number zero, this policy will rule out his complying with SUCCESSOR by letting zero be its own successor. Second, he follows a policy of economy; he never does any more than is necessary to comply with a given rule. Thus, this policy will rule out his complying with ZERO by introducing two or more objects that are numbers. With these assumptions in force, the rule NUMBER will then have the intended effect.[4]

Set Theory

We now read Sx as 'x is a set' and xεy as 'x belongs to y'. There are two rules:

POWER: $\forall F!y.(Sy \;\&\; \forall x(x\varepsilon y \equiv Fx))$; and
SET: POWER*.

POWER says: for any plurality F of pre-existing objects, introduce an object that is a set and has exactly the objects in the plurality as members. SET (also pronounced 'Let there be sets!') says: keep on performing POWER, i.e. adding sets corresponding to any given plurality of objects.

[4] Strictly speaking, we should provide a formal semantics for our postulational language and then establish, under this semantics, that NUMBER will result in the intended domain. But I shall slide over such technical details in the exposition that follows.

In application to any given domain that contains no sets, POWER will then introduce all sets (of finite rank) that may be constructed from the objects of the domain.

I would argue that we can obtain all intuitively given mathematical domains in a similar way. These include the cumulative hierarchy of ZF, the various extensions of the number system to integers, rationals, reals, and complex numbers, and Euclidean geometry of any given dimension. For some of the set-theoretic cases, we need to make use of a stronger form of iteration, one in which the iteration of a postulate β^* can proceed into the transfinite. But, with this difference aside, the basic forms of postulation can remain the same. This is clearly a very bold claim; and its defense lies beyond the scope of this paper.

For epistemological purposes, we not only need a characterization of the mathematical domain in terms of a postulational rule, we also need to show that the characteristic axioms for the domain can be derived from that rule. We therefore need to develop a logic within which such a derivation can be carried out. This cannot be a logic of a standard sort, since it derives propositions from procedures rather than propositions from propositions. The characteristic form of inference of such a logic might be represented as follows:

$$\frac{A_1, A_2, \ldots, A_n}{B}\alpha \text{ (from } A_1, A_2, \ldots, A_n, \text{ given } \alpha, \text{ we may infer B)},$$

where $A_1, A_2, \ldots, A_n$ are indicative sentences (expressing propositions) and α is a rule (prescribing a procedure). Such an inference is then valid if the execution of α converts a domain in which $A_1, A_2, \ldots, A_n$ are true to one in which B is true.

Although I shall not go into details, it is possible to develop a logic of postulation along these lines.[5] Roughly speaking, each form of postulational rule will be associated with rules of inference telling us what the effect of complying with the rule will be. Consider the compositional rule $\beta;\gamma$, for example. Then, given that the inferences:

$$\frac{A}{B}\beta \qquad \frac{B}{C}\gamma$$

[5] There are similarities, which I shall not explore, between postulational logic and various forms of dynamic programming logic that have been developed in computer science (see Harel *et al.* (2000) for a survey).

are valid, we should expect the inference:

$$\frac{A}{C}\beta;\gamma$$

to be valid. If β makes B true, given an initial domain in which A, and γ makes C true, given an initial domain in which B, then β;γ should make C true, given an initial domain in which A. So, for example, if ZERO makes true that there is a number (∃xNx), given an initial domain in which there are no numbers (∼ ∃xNx), and SUCCESSOR makes true that there is a successor of a number ($\exists x\exists y(Nx \,\&\, Syx)$), given an initial domain in which there is a number, then ZERO; SUCCESSOR should make true that there is a successor of a number given an initial domain in which there are no numbers.

Using a postulational logic of this sort, it is then possible to show that the standard axioms for a given domain can indeed be derived from the postulational rules for that domain. From NUMBER above, for example, we can derive the standard (second-order) axioms for the theory of number and from SET we can derive a standard set of axioms for set theory. And similarly, so I would argue, for the other standard axiomatic theories of mathematics. Thus the postulational rules are not only sufficient to characterize the intuitively given domains of mathematics, but also sufficient to derive the standard axioms for those domains.

We obtain in this way a kind of axiom-free foundation for mathematics. The various axioms for the different branches of mathematics are derived, not from more basic axioms of the same sort, but from postulational rules. The axioms, which describe the composition of a given mathematical domain, give way to the stipulation of procedures for the construction of that domain. We therefore obtain a form of logicism, though with a procedural twist. The axioms of mathematics are derived from definitions and logic, as in the standard version of logicism, but under a very different conception of definition and of logic, since the definitions take the form of postulational rules and the logic provides the basis for reasoning with those rules. Moreover, in contrast to the logic required by the standard forms of logicism, our logic is ontologically neutral. We do not assume that there are any objects and, indeed, the whole of mathematics can be derived under the assumption that there are no objects; whatever objects we need can be generated from the prescription of suitable procedures.

2. THE PROBLEM OF CONSISTENCY

After this very brief sketch of the underlying technical program, we return to the original question. How are we capable of achieving mathematical knowledge and how, in particular, are we capable of acquiring knowledge of mathematical objects? No current epistemology of mathematics is altogether satisfactory; and so we would do well to consider to what extent the current postulational approach is able to shed any additional light on this question.

The prospects might look encouraging. Consider the case of numbers. We may lay down the postulational rule NUMBER above. From this, by means of the logic of postulation, we may then derive the standard axioms of arithmetic. We thereby obtain the axioms of arithmetic without any apparent epistemic cost. They are derived on the basis of logic, that is itself without existential import, and sheer stipulation.

Unfortunately, things are not so simple. There are two main problems —one concerning the input side to the purported derivation and the other the output side. The first of the problems is that we are not free to prescribe anything we like—at least, if we think of this as entitling us to assert what would thereby be rendered true under the prescription. Suppose I introduce an object x that is both a number and not a number ($!x.(Nx \,\&\, \sim Nx)$). It would then follow, within the logic of postulation, that something was both a number and not a number! Another difficulty arises, not from performing a given step prescribed by a procedure, but from completing *all* of the steps. Suppose, for example, that I lay down the *indefinite* iteration of POWER, i.e. POWER* under the strong, transfinite, reading of*. Then completion of the procedure prescribed by POWER* would require a domain in which unrestricted comprehension holds ($\forall F \exists y \forall x(x \in y \equiv Fx)$). And so, by the reasoning of Russell's paradox, we would again be saddled with a contradiction.

It is clear that a necessary condition for us to be entitled to prescribe a given postulational rule is that we should show it to be consistent in its consequences. But how are we to do this?

The problem is already familiar from the traditional forms of postulationism; and *there* it appears to have no satisfactory solution. Consider the case of arithmetic, for example. We wish to show that some standard set of axioms for arithmetic—say those of Dedekind—are consistent.

But we cannot appeal to the existence of a model for the axioms, since that is a mathematical fact which requires a mathematical proof of a sort whose justification is already in question. For the same reason, we cannot appeal to the consistency of the axioms within some formal system, since this requires reference to formulas and formal proofs, which is just as bad as reference to numbers. Nor can we appeal to the truth of the axioms under their standard interpretation, since this is what we hoped to establish. It seems that the best we can do is appeal to the fact that we have so far failed to derive a contradiction from the axioms. The evidence, in other words, is inductive.

But such inductive evidence hardly does justice to the degree of confidence that we feel entitled to place in the consistency of arithmetic and various other mathematical theories. For it provides the consistency of those theories with no better credibility than that of a well-confirmed mathematical conjecture. We have inductive evidence in favor of the consistency of Quine's New Foundations. So why do we feel entitled to place greater confidence in the consistency of arithmetic? Indeed, we seem to be entitled to a confidence in the consistency of the axioms of arithmetic simply on the basis of an 'intuition' that they are true and prior to any consideration of the inductive evidence. This is completely inexplicable under the inductivist view.[6]

One might think that the procedural postulationist is no better placed to solve this problem than his rivals. But before jumping to this conclusion, we should consider more carefully how, for him, the question of consistency is to be construed. For the prescription of a postulational rule will be consistent in its consequences if the procedure that it prescribes can indeed be executed. Thus for the procedural postulationist the question of consistency is, in effect, the question of executability. We might, if you like, talk of the rule α itself being consistent, though we should recognize that this is a consistency in what can be *done* rather than in what can be *true*.

Given a postulational rule α and an indicative statement A, let us use the indexed modal claim $\Diamond_\alpha A$ to indicate that it is possible to execute the procedure prescribed by α in such a way that A is then true. And similarly, let us use the indexed modal claim $\Box_\alpha A$ to indicate that it is necessary, however the procedure prescribed by α is executed, that A is

[6] There are other objections to the view that I have not considered. See Field (1989: ch. 4) for an espousal of inductivism within the context of nominalism.

then true. We may also use the unindexed modal claims $\diamond A$ and $\Box A$ to indicate that A is true under some possible (respectively, every possible) execution of an admissible procedure. Let $\top$ be any logical theorem (such as $\forall x(x = x)$). The postulational consistency of α might then be formalized as:

$$\diamond_\alpha \top$$

since this indicates that it is possible to execute the procedure prescribed by α in such a way that $\top$ is then true, which is simply to say that it is possible to execute the procedure prescribed by α. We might call $\diamond_\alpha$, $\Box_\alpha$, $\diamond$, and $\Box$ the *postulational* modalities, since they relate to what can be true under the possible execution of a procedure.

Now what is remarkable is that, once consistency claims are formulated in this way, it is possible to provide convincing demonstrations of their truth that are purely modal in character and that make no appeal either to models or proofs or to any other kind of abstract object. Consider the postulate NUMBER, for example. This is of the form ZERO; SUCCESSOR*, where ZERO is the postulate !x.Nx and SUCCESSOR the postulate $\forall x(Nx \rightarrow !y.(Ny \,\&\, Syx))$. Say that a postulational rule is *strongly* consistent, or *conservative*, if it is necessarily consistent ($\Box \diamond_\alpha \top$), that is, consistent no matter what the domain. Then it should be clear that the simple rule ZERO is conservative and that the simple rule !y.(Ny & Syx) is conservative whatever the object x. For these rules introduce a single new object into the domain that is evidently related in a consistent manner to the pre-existing objects (or else they do nothing). It should also be clear that each of the operations for forming complex rules will preserve conservativity. For example, if β and γ are conservative then so is $\beta; \gamma$, for β will be executable on any given domain and whatever domain it thereby induces will be one upon which γ is executable. But it then follows that NUMBER is consistent, as is any other rule that is formed from conservative simple rules by means of the operations for forming complex rules.[7]

The contrast with the standard postulational approach is striking. There is nothing in the axiomatic characterization of a basic mathematical domain that enables us to determine its consistency and, in

[7] Consistency can also be demonstrated for the higher reaches of set theory but since $*$, in the strong sense, will no longer preserve conservativity, we must make special assumptions concerning the executability of transfinite procedures.

particular, the consistency of a conjunction of axioms cannot be inferred from the consistency of its separate conjuncts. But once the present postulational approach is adopted, the consistency (and, indeed, the conservativity) of a rule can be read off 'compositionally' from its very formulation.

Moreover, the method of proof can be extended to show that the standard propositional axioms for a theory are also consistent. Suppose that α is a postulational rule and that A is a corresponding axiom for the resulting domain (when α is NUMBER, for example, A might be taken to be the conjunction of axioms for second-order arithmetic). We may demonstrate the consistency of α, i.e. $\Diamond_\alpha \top$, as above. Within the logic of postulation, we can then derive A from α, that is, we can show:

$$\frac{}{A}\,\alpha$$

(using no assumptions concerning the initial domain). This translates into a proof of $\Box_\alpha A$. From $\Box_\alpha A$ and $\Diamond_\alpha \top$, we can derive $\Diamond_\alpha A$ by ordinary modal reasoning; and from this follows the consistency of A, i.e. $\Diamond A$. (We might think of the axioms A as constituting a 'specification' for a program α. A proof of the above sort then constitutes a 'verification' that the program α does indeed meet the specification.)

I also believe, though this is not something I shall pursue, that the rules for the construction of a mathematical domain may be taken to represent our intuitive grasp of that domain and that a demonstration of the above sort may then be seen to represent the role that intuition can play in vindicating the consistency of the axioms for that domain. Thus we are able, on our approach, to account for the special kind of confidence that intuition is able to provide in the consistency of a mathematical theory.

3. THE PROBLEM OF EXISTENCE

Our aim is to establish the conjunction A of the axioms $A_1, A_2, \ldots, A_n$ of some mathematical theory in the following schematic way:

$$\frac{\Diamond_\alpha \top}{A}\,\alpha$$

We first demonstrate the consistency of an appropriate postulational rule α; this entitles us to prescribe α; and, on the basis of the prescription of α, we may then establish the conjoined axioms A. We have seen from the previous section how we might demonstrate the consistency of the rule α. The question remains as to how we might then justify the inference to A.

In general, the 'axioms' A that may be derived under the prescription of α will enable us to make existential claims that we were not able to make prior to its prescription. They may imply, for example, that there is an infinitude of numbers even though we previously had no warrant for asserting the existence of numbers. This makes the status of the inference highly problematic. For all that appears to license the prescription of the rule is the demonstration of its consistency. We know that the procedure specified by the rule *can* be executed. But what then justifies us in proceeding as if it *had* been executed? It is as if we were to infer from the possibility of pulling a rabbit out of the hat that the rabbit was already there!

In order to answer this question, we must delve more deeply into the nature of procedural postulation. It is important, in the first place, to appreciate that the inference in question is not simply from the consistency of $\alpha(\diamond_\alpha \top)$ to the truth of A, an inference which we might represent as:

$$\frac{\diamond_\alpha \top}{A}$$

For our inference works through the intermediary of α. What the consistency of α licenses is not A but the prescription of α. The prescription of α then effects a change in the interpretation of the domain of discourse; and it is this change in the interpretation of the domain that then justifies us in inferring A.

In order to understand how the inference might be justified, we must therefore understand how the prescription of α might be capable of effecting a change in the interpretation of the domain. Now one way in which it might do this is familiar and unproblematic. One domain may be understood as the restriction of another. If a given domain is understood as ranging over all professors, for example, we may restrict it to the subdomain of male professors. Now a procedural postulation is meant to effect an expansion in the domain of discourse and so one way of understanding how it might do this is to suppose that it relaxes a

restriction on the domain that is already in force. It could do this either by lifting the current restriction—going from *male professor*, for example, to *professor*—or by loosening the current restriction—going from *male professor* to *male or female professor*—or perhaps in some other way.

If this is how procedural postulation is meant to work, then I see no way in which it might plausibly be taken to provide us with the kind of justification for existential claims that we are after. For there is nothing in the nature of relaxing a restriction that might warrant us in supposing that there are objects not subject to the restriction. In making the transition from *male professor* to *professor*, for example, we are not entitled to assume that there are any female professors; and, likewise, in allowing a domain to include numbers, say, or sets, we are not entitled to assume that there actually are any numbers or sets.

However, I believe that there may be a radically different way of understanding how postulation might be capable of effecting an expansion in the domain of discourse. Let us call a domain of discourse *unrestricted* if it is not to be understood, either explicitly or implicitly, as the restriction of some other domain. The quantifiers in an unrestricted domain of discourse will be understood as ranging over everything that there is, since their ranging over anything less would be tantamount to a restriction on the domain. Now it seems to me that there are methods of domain-expansion that have application to unrestricted domains; and if this is so, then they cannot be understood on the previous model as a form of de-restriction.

Perhaps the most convincing way of demonstrating the possibility of expansion in the case of an unrestricted domain is to show that there are methods of expansion that work *whatever* the domain might be. For there do appear to be such methods. Consider the following rule, for example:

$$\text{RUSSELL:} \quad !y.\ \forall x(x \in y \equiv\sim x \in x),$$

for introducing the 'set' whose members are exactly those pre-existing objects that are not members of themselves. It seems evident that, whatever the domain of discourse, we may legitimately suppose there to be such an object; and yet we cannot, on pain of contradiction, suppose that it already is in the domain.[8]

[8] This issue and its connection with Russell's paradox is further discussed in my paper 'Relatively Unrestricted Quantification', to appear in Agustin Rayo and Gabriel Uzquiano (eds.), *Unrestricted Quantification: New Essays*, to be published by Oxford University Press.

Let us be a little more exact. Suppose that our initial understanding of the quantifier is given by $\forall$ and $\exists$ and that our new understanding of the quantifier is given by $\forall^+$ and $\exists^+$. It then seems reasonable to suppose that, given our initial understanding of the quantifier $\forall$, we can so understand the quantifier $\exists^+$ that $\exists^+y\forall x(x \in y \equiv\sim x \in x)$ is true; and it also seems reasonable to suppose that this new understanding of the quantifier can be secured through something like the prescription of RUSSELL. But then, by the reasoning of Russell's paradox, it follows that $\exists^+y\forall x(x \neq y)$; and so, even though the initial quantifier may have been unrestricted, an expansion in its range will still have been achieved.

My opponent may object that the existence of the new interpretation of the quantifier, as given by $\forall^+$ and $\exists^+$, shows that the initial interpretation, as given by $\forall$ and $\exists$, must already have been restricted. Let D be a predicate that picks out the objects from the initial domain of discourse. Then $\forall$ and $\exists$, it will be claimed, must be understood as the restriction of $\forall^+$ and $\exists^+$ to D.

But this objection confuses two different senses of 'restriction'. One quantifier is a restriction of another in the *extensional* sense if every object in its range is in the range of the other, while one quantifier is a restriction of another in the *intensional* sense if it is to be understood *as* a restriction of the other. Thus an unrestricted quantifier in this latter sense is one that is not to be *understood* as the restriction of some other quantifier. With this distinction at hand, we see that the objection merely shows that the initial quantifiers $\forall$ and $\exists$ are restrictions in the extensional sense of the new quantifiers $\forall^+$ and $\exists^+$. But our concern was with how the quantifiers are to be understood; and here the order of explanation goes in the opposite direction—the new quantifiers $\forall^+$ and $\exists^+$ are to be understood as expansions of the initial quantifiers $\forall$ and $\exists$. Thus even though there are restrictions of quantifiers $\forall^+$ and $\exists^+$ with the same range of values as the initial quantifiers, these restricted quantifiers do not provide us with the required understanding of the initial quantifiers. Rather, these restricted quantifiers themselves must ultimately be understood in terms of the initially unrestricted quantifiers, $\forall$ and $\exists$.

If this is right, then there are essentially two different ways in which one domain of discourse may understand in terms of another. One is *classical* or *restrictive*; a wider domain of discourse is presupposed and, in so far as an act of postulation succeeds in expanding a given domain, it will do so by relaxing some restriction that is already in force. The new

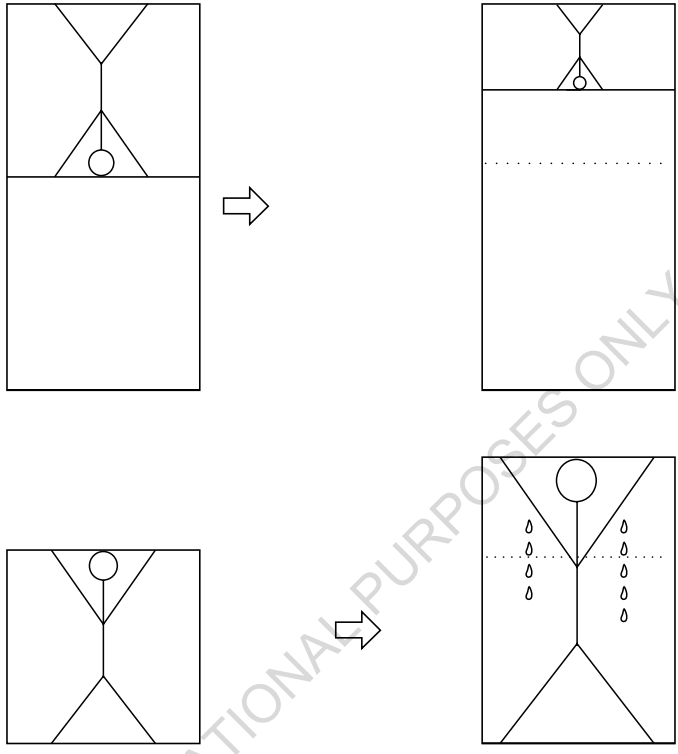

Figure 4.1. Expanding domains of discourse

domain is understood from the outside, as it were. The other method is *creative* or *expansive*; no wider domain is presupposed and postulation works by effecting a genuine expansion in the given domain. The new domain is understood from the inside, as it were. We might picture the difference as in Figure 4.1. In the former case, the given domain is seen as residing within a larger domain and its boundaries are then pushed outwards within that larger domain while, in the latter case, the given domain is already taken to constitute everything that there is and its boundaries are pushed outwards without presupposing that there is already a domain within which the expansion takes place (as a concession to my opponent, I have depicted the latter as involving some extra effort).

We may now return to the question with which this section began: what justifies us in drawing existential conclusions from a postulational

rule once the rule has been shown to be consistent? We may have shown the rule NUMBER to be consistent, for example. What reason do we then have for supposing that we can legitimately prescribe the rule and thereby establish the existence of 'numbers'? Now it seems to me that if the prescription of NUMBER can be used to establish the existence of numbers then it must be because it is, like RUSSELL, a creative act of postulation, one which genuinely serves to expand the given domain. Our question may therefore be raised in the following form: given the general possibility of creative postulation, as typified by a rule like RUSSELL, say, then what are the conditions under which it is legitimate? Is consistency of the rule to be laid down enough? Or is something else required?

In the case of RUSSELL it is hard to see what besides the consistency might be used to legitimate the inference to there being a Russell-set $(\exists^{+}y\forall x(x \in y \equiv\sim x \in x))$. It is legitimate to expand the domain with the Russell-set simply because there is no inconsistency in supposing that there is such a set. But my opponent is unlikely to think that this is enough. Surely, he will argue, we want not just that the object or objects *might* exist but that they *do* exist.

However, it seems to me that, in the context of creative postulation, this objection is misplaced and depends upon confusing creative postulation with other, more orthodox, forms of definition. For let us ask what this requirement of existence is meant to be. It cannot be that there exists a Russell-set in the *given* sense of what exists $(\exists y\forall x(x \in y \equiv\sim x \in x))$, since a negative answer has no bearing upon the legitimacy of the postulation. Indeed, the whole point of the postulation was to go from a sense of 'exists' in which there did not exist such a set to one in which there did. Nor can the requirement be that there exist a Russell-set in the *new* sense of what exists $(\exists^{+}y\forall x(x \in y \equiv\sim x \in x))$ since this sense is not yet available to us. Of course, my opponent wants to ask, in an absolutely unrestricted sense of 'exists', whether there exists such a set. But the given quantifier '$\exists x$' is already unrestricted and so there is no other, less restricted, form of quantification to which he can appeal. Thus it seems that any formulation of the existential requirement is either irrelevant to the question of legitimacy or ineffable.

It might be thought that the existential requirement is something that should be imposed *after* the rule has been prescribed, but not before. The thought is that something as weak as consistency might indeed legitimate the prescription of a rule but that the prescription then

results in a domain in which it is an open question whether or not objects of the required sort exist. Thus, on this view, the prescription secures an intermediate interpretation of the quantifier, one that makes room for the objects of the required sort but without guaranteeing that they exist.

The problem with this suggestion is to understand what this intermediate interpretation of the quantifier might be. How could RUSSELL, for example, lead to an interpretation of the quantifier in which it is a genuinely open question whether or not the Russell-set over the given domain exists? One obvious way in which the question of existence may be open is when the domain is given by some condition. The objects of the domain are all those that satisfy the condition; and so for an object of a given sort to exist is for there to be an object of that sort that satisfies the condition. Suppose now that the given domain is given by the condition $D(x)$. Then can we not take the new domain to be given by the disjunctive condition $D^+(x) = (D(x) \vee \forall y(y \in x \equiv y \notin y))$? And is it not then an open question whether there is an object satisfying the second disjunct $\forall y(y \in x \equiv y \notin y)$?

However, one should bear in mind that the free variables in the condition $D^+(x)$ are as equally subject to interpretation as its bound variables; and, given that the bound variables already range over everything that there is, there is nothing better that the free variable can do. They cannot reach out, as it were, to objects that are somehow not susceptible to quantification. But this means that the disjunctive condition $D^+(x)$ is not genuinely weaker than the original condition $D(x)$ and that it is an illusion to think that the intermediate interpretation of the quantifier can somehow be understood by relaxing the requirement on membership in the domain.

How then is the new interpretation of the domain to be understood? There is, it seems to me, no alternative but to understand it *as* an extension of the old domain. In the case of RUSSELL, for example, we must understand the new domain to be one in which there is an object to which all non-self-membered sets of the old domain stand in the relationship of membership. And, in general, we must understand the new domain in terms of how its objects relate to one another and to the objects in the old domain. It is as if we were to draw a diagram of how the new objects relate to the old objects; and the new interpretation of the domain is simply to be understood as one in which that diagram is realized.

There is therefore no room for an intermediate interpretation of the quantifier; and we see that no sensible requirement of existence can be imposed either as a condition on postulation itself or as a condition on what can be inferred from a postulational rule, once it has been prescribed.

4. THE SCOPE OF POSTULATION

I wish, in conclusion, to consider two other epistemological objections to our method of postulation. One is to its generality—that, once the method is allowed, it cannot be properly contained. The other is to its viability—that the method is either incoherent or without significant application. Our answers to both objections will enable us better to understand the nature of postulation and its place within the realm of rational enquiry.

The first objection goes as follows. The postulational rule

UNIVERSAL-SET: $!y.\forall x(x \in y)$

is consistent. So, according to our method of postulation, it may legitimately be prescribed; and, from the prescription of the rule, the existence of a universal set over the given domain may then be inferred. But is not

UNIVERSAL-LOVER: $!y.\forall x(\text{Person}(x) \supset y \text{ loves } x)$

also consistent? So what is to stop us from prescribing it and thereby establishing the existence of someone who loves everyone? Or to take a more familiar example:

GOD: $!y.\text{Divine}(x)$

is presumably consistent under a suitable understanding of 'divine'. So what is to stop us from prescribing it and thereby using something like the ontological argument to establish the existence of God?

Clearly, UNIVERSAL-LOVER and GOD should not be countenanced as postulational rules or, if they are, then they should be declared to be inconsistent on the grounds that there is no genuine *postulational* possibility of someone's loving everyone or of something's being divine. But on what basis do we distinguish between the cases, like RUSSELL and UNIVERSAL-SET, which we want to admit, and cases, like UNIVERSAL-LOVER and GOD, which we want to dismiss?

The difference appears to lie in the predicates. It is legitimate to postulate by means of predicates such as '$\in$' or '$<$' or 'successor', but not by means of predicates such as 'person' or 'loves' or 'divine'. But what then is the relevant difference in the predicates?

I should like to suggest that it lies in how the predicates are to be understood. Predicates of the first kind are in a certain sense formal; they are *simply* to be understood in terms of how postulation with respect to them is to be constrained. Thus our understanding of S (set) and $\in$ (membership) is entirely given by the fact that they conform to the following constraints:

Extensionality	$\forall x \forall y[Sx \,\&\, Sy \,\&\, \forall z(z \in x \equiv z \in y) \supset x = y]$;
Sethood	$\forall y[\exists x(x \in y) \supset Sy]$;
Set-Rigidity	$\forall y \forall F[Sy \,\&\, \forall x(x \in y \supset Fx) \supset \Box \forall x(x \in y \supset Fx)]$.

Thus according to the first constraint, no two sets are to be postulated to have the same members; according to the second, anything that is postulated to have members is to be postulated to be a set; and according to the third, no *members* of pre-existing sets are to be postulated (this explains the sense in which sets are formed from their members but not members from their sets). Similarly, our understanding of N (number) and S (successor) is entirely given by their conformity to the following three constraints:

Uniqueness	$\forall x \forall y \forall z(Syx \,\&\, Szx \supset y = z)$;
Numberhood	$\forall x \forall y(Sxy \supset Nx \,\&\, Ny)$;
Successor-Rigidity	$\forall x(Nx \supset (\sim \exists y Syx \supset \Box \sim \exists y Syx) \,\&\, \forall y(Syx \supset \Box \forall w(wx \supset w = y)))$.

According to the third of these, we cannot postulate *predecessors*, just as we cannot postulate *members*.

Say that a predicate is *postulational* if its meaning is entirely given by a set of postulational constraints. Our view is that we are entitled to postulate by means of postulational predicates, as long as we stay within the constraints by which they are defined. However, we are not entitled to postulate by means of any other predicates, since there is then an independent question, not to be settled by postulation alone, of what their application should be. Thus it is only when the predicates have no content beyond their role in postulation that they may legitimately be used as vehicles of postulation.

It should be noted that, on the present view, the usual existential assertions of mathematics will not be analytic; there is nothing in our understanding of the terms that they involve which will guarantee that they are true. It will not be analytic, for example, that there are numbers. For all that our understanding of number-predicates will guarantee is the correctness of the postulational constraints by which it is governed. But this is entirely without existential import. Of course, once we have demonstrated the consistency of an appropriate postulational rule, such as ZERO, we are then entitled to lay down the rule and thereby infer the existence of a number. But there is nothing in our understanding of the number-predicates themselves which either entitles us to assert the consistency of the rule or which obliges us to lay it down.

The second objection to our approach is more radical. Our whole defense of the method of postulation has been conditional in form: if one accepts creative postulation as legitimate, then one should accept consistency as a basis upon which it may proceed. But it might be denied that creative postulation is legitimate—either on the grounds that it is incoherent or on the grounds that, even though coherent, it has no significant application.

This objection strikes me as being essentially sceptical in spirit; for it is at odds with the commonly accepted epistemic facts. We *do* postulate. And here I do not merely have in mind the somewhat controversial case of sets. The actual practice of mathematics, before it was sanitized by logicians, contained numerous examples of postulation. The complex number i, for example, was postulated as a number for which $i^2 = -1$; and $+\infty$ was postulated as a number greater than all reals. The philosopher who rejects postulation must reject standard (or, at least, what *was* once standard) postulational practice.

I have no answer to scepticism—in this case or in any other. All I can say is that the sceptic has an unduly narrow conception of how we might come to know what we do. Each kind of object has its own way of being known. It is a peculiarity of perceptible objects that we may get to know of them through perception; it is a peculiarity of the theoretical entities of science that their existence is to be justified by way of inference to the best explanation; and it is a peculiarity of mathematical and other abstract objects that their existence is to be justified by way of postulation. In recent times, many philosophers have been attracted to an 'assimilationist' model of mathematical knowledge; they have supposed

that we know of mathematical objects in something like the way we know of other objects—either directly through some form of perception or apprehension or indirectly through inference to the best explanation. If the present approach has any value, it lies in its making clear the *distinctive* way in which we may acquire our knowledge of mathematical objects, one that is not reducible to other, more familiar methods and is in keeping with the peculiarly a priori character of mathematical thought.

REFERENCES

Field, H. (1989) *Realism, Mathematics and Modality* (New York: Blackwell).
Fine, K. (2002) *Limits of Abstraction* (Oxford: Clarendon Press).
Harel, D, D. Kozen, and J. Tiuryn (2000) *Dynamic Logic* (Cambridge, Mass.: MIT Press).
Hilbert, D. (1930) *Grundlagen der Geometrie*, 7th edn. (Leipzig: Open Court Press).
Poincarě, H. (1952) *Science and Method* (New York: Dover).

I should like to thank the members of seminars at UCLA, Harvard and NYU for very helpful comments. I should also like to thank Paul Boghossian, Ruth Chang, Bob Hale, Tony Martin, Derek Parfit and Alan Weir.

This page intentionally left blank

5. Sleeping Beauty Reconsidered: Conditioning and Reflection in Asynchronous Systems

Joseph Halpern

1. INTRODUCTION

The standard approach to updating beliefs in the probability literature is by conditioning. But it turns out that conditioning is somewhat problematic if agents have *imperfect recall*. In the economics community this issue was brought to the fore by the work of Piccione and Rubinstein (1997), to which was dedicated a special issue of the journal *Games and Economic Behavior*. There has also been a recent surge of interest in the topic in the philosophy community, inspired by a re-examination by Elga (2000) of one of the problems considered by Piccione and Rubinstein, the so-called *Sleeping Beauty problem*.[1] (Some recent work on the problem includes Arntzenius 2003; Dorr 2002; Lewis 2001; Monton 2002.)

I take the Sleeping Beauty problem as my point of departure in this paper too. I argue that the problems in updating arise not just with imperfect recall, but also in *asynchronous* systems, where agents do not know exactly what time it is, or do not share a global clock. Since both human and computer agents are resource-bounded and forgetful, imperfect recall is the norm, rather than an unusual special case. Moreover, there are many applications where it is unreasonable to assume the existence of a global clock. Thus, it is important to understand how to do updating in the presence of asynchrony and imperfect recall.

Work supported in part by NSF under grant CTC-0208535, by ONR under grants N00014-00-1-03-41 and N00014-01-10-511, by the DoD Multidisciplinary University Research Initiative (MURI) program administered by the ONR under grant N00014-01-1-0795, and by AFOSR under grant F49620-02-1-0101. A preliminary version of this paper appears in *Principles of Knowledge Representation and Reasoning: Proceedings of the Ninth International Conference*.

[1] So named by Robert Stalnaker.

The Sleeping Beauty problem is described by Elga as follows:

> Some researchers are going to put you to sleep. During the two days that your sleep will last, they will briefly wake you up either once or twice, depending on the toss of a fair coin (heads: once; tails: twice). After each waking, they will put you back to sleep with a drug that makes you forget that waking. When you are first awakened, to what degree ought you believe that the outcome of the coin toss is heads?

Elga argues that there are two plausible answers. The first is that it is 1/2. After all, it was 1/2 before you were put to sleep and you knew all along that you would be woken up (so you gain no useful information just by being woken up). Thus, it should still be 1/2 when you are actually woken up. The second is that it is 1/3. Clearly if this experiment is carried out repeatedly, then in the long run, at roughly one-third of the times that you are woken up, you are in a trial in which the coin lands heads.

Elga goes on to give another argument for 1/3, which he argues is in fact the correct answer. Suppose you are put to sleep on Sunday, so that you are first woken on Monday and then possibly again on Tuesday if the coin lands tails. Thus, when you are woken up, there are three events that you consider possible:

- e_1: it is now Monday and the coin landed heads;
- e_2: it is now Monday and the coin landed tails;
- e_3: it is now Tuesday and the coin landed tails.

Elga's argument has two steps:

1. If, after waking up, you learn that it is Monday, you should consider e_1 and e_2 equally likely. Since, conditional on learning that it is Monday, you consider e_1 and e_2 equally likely, you should consider them equally likely unconditionally.
2. Conditional on the coin landing tails, it also seems reasonable that e_2 and e_3 should be equally likely; after all, you have no reason to think Monday is any more or less likely than Tuesday if the coin landed tails. Thus, unconditionally, e_2 and e_3 should be equally likely.

From these two steps, it follows that e_1, e_2, and e_3 are equally likely. The only way that this could happen is for them all to have probability 1/3. So heads should have probability 1/3.

Suppose that the story is changed so that (1) heads has probability .99 and tails has probability .01, (2) you are woken up once if the coin lands

heads, and (3) you are woken up 9900 times if the coin lands tails. In this case, Elga's argument would say that the probability of tails is .99. Thus, although you know you will be woken up whether the coin lands heads or tails, and you are initially almost certain that the coin will land heads, when you are woken up (according to Elga's analysis) you are almost certain that the coin landed tails!

How reasonable is this argument? The second step involves an implicit appeal to the Principle of Indifference. But note that once e_1 and e_2 are taken to be equally likely, the only way to get the probability of heads to be 1/2 is to give e_3 probability 0, which seems quite unreasonable. Thus, an appeal to the Principle of Indifference is not critical here to argue that 1/2 is not the appropriate answer.

What about the first step? If your probability is represented by Pr then, by Bayes' Rule,

$$\Pr(heads \mid Monday) = \frac{\Pr(Monday \mid heads)\ \Pr(heads)}{\Pr(Monday \mid heads)\ \Pr(heads)\ + \Pr(Monday \mid tails)\ \Pr(tails)}.$$

Clearly $\Pr(Monday \mid heads) = 1$. By the Principle of Indifference, $\Pr(Monday \mid tails) = 1/2$. If we take $\Pr(heads) = \Pr(tails) = 1/2$, then we get $\Pr(heads \mid Monday) = 2/3$. Intuitively, it being Monday provides stronger evidence for heads than for tails, since $\Pr(Monday \mid heads)$ is larger than $\Pr(Monday \mid tails)$. Of course, this argument already assumes that $\Pr(heads) = 1/2$, so we can't use it to argue that $\Pr(heads) = 1/2$. The point here is simply that it is not blatantly obvious that Pr ($heads \mid Monday$) should be taken to be 1/2.[2]

To analyze these arguments, I use a formal model for reasoning about knowledge and probability that Mark Tuttle and I developed (Halpern and Tuttle 1993—HT from now on), which in turn is based on the ''multiagent systems'' framework for reasoning about knowledge in computing systems, introduced in Halpern and Fagin (1989); see Fagin *et al*. (1995) for motivation and discussion. Using this model, I argue that Elga's argument is not as compelling as it may seem, although not for the reasons discussed above. The problem turns out to depend on the difference between the probability of heads conditional on it being Monday vs. the probability of heads conditional on *learning* that it is Monday. The analysis also reveals that, despite the focus of the

[2] Thanks to Alan Hájek for making this point.

economics community on imperfect recall, the real problem is not imperfect recall, but asynchrony: the fact that Sleeping Beauty does not know exactly what time it is.

I then consider other arguments and desiderata traditionally used to justify probabilistic conditioning, such as frequency arguments, betting arguments, van Fraassen's (1984) Reflection Principle, and Savage's (1954) Sure-Thing Principle. I show that our intuitions for these arguments are intimately bound up with assumptions such as synchrony and perfect recall.

The rest of this paper is organized as follows. In the next section I review the basic multiagent systems framework. In §3, I describe the HT approach to adding probability to the framework when the system is synchronous. HT generalized their approach to the asynchronous case; their generalization supports the ''evidential argument'' above, giving the answer $\frac{1}{2}$ in the Sleeping Beauty problem. I also consider a second generalization, which gives the answer $\frac{1}{3}$ in the Sleeping Beauty problem (although not exactly by Elga's reasoning). In § 4, I consider other arguments and desiderata. I conclude in § 5.

2. THE FRAMEWORK

2.1. The Basic Multiagent Systems Framework

In this section, we briefly review the multiagent systems framework; see Fagin *et al.* (1995) for more details.

A *multiagent system* consists of n agents interacting over time. At each point in time, each agent is in some *local state*. Intuitively, an agent's local state encapsulates all the information to which the agent has access. For example, in a poker game, a player's state might consist of the cards he currently holds, the bets made by the other players, any other cards he has seen, and any information he may have about the strategies of the other players (e.g. Bob may know that Alice likes to bluff, while Charlie tends to bet conservatively). In the Sleeping Beauty problem, we can assume that the agent has local states corresponding to ''just woken up'', ''just before the experiment'', and ''just after the experiment''.

Besides the agents, it is also conceptually useful to have an ''environment'' (or ''nature'') whose state can be thought of as encoding everything relevant to the description of the system that may not be

included in the agents' local states. For example, in the Sleeping Beauty problem, the environment state can encode the actual day of the week and the outcome of the coin toss. In many ways, the environment can be viewed as just another agent. In fact, in the case of the Sleeping Beauty problem, the environment can be viewed as the local state of the experimenter.

We can view the whole system as being in some *global state*, a tuple consisting of the local state of each agent and the state of the environment. Thus, a global state has the form $(s_e, s_1, \ldots, s_n)$, where s_e is the state of the environment and s_i is agent i's state, for $i = 1, \ldots, n$.

A global state describes the system at a given point in time. But a system is not a static entity. It is constantly changing over time. A *run* captures the dynamic aspects of a system. Intuitively, a run is a complete description of one possible way in which the system's state can evolve over time. Formally, a run is a function from time to global states. For definiteness, I take time to range over the natural numbers. Thus, $r(0)$ describes the initial global state of the system in a possible execution, $r(1)$ describes the next global state, and so on. A pair (r, m) consisting of a run r and time m is called a *point*. If $r(m) = (s_e, s_1, \ldots, s_n)$, then define $r_e(m) = s_e$ and $r_i(m) = s_i$, $i = 1, \ldots, n$; thus, $r_i(m)$ is agent i's local state at the point (r, m) and $r_e(m)$ is the environment's state at (r, m). I write $(r, m) \sim_i (r', m')$ if agent i has the same local state at both (r, m) and (r', m'), that is, if $r_i(m) = r'_i(m')$. Let $\mathcal{K}_i(r, m) = \{(r', m') : (r, m) \sim_i (r', m')\}$. Intuitively, $\mathcal{K}_i(r, m)$ is the set of points that i considers possible at (r, m); these are the states that i cannot distinguish on the basis of i's information at (r, m). Sets of the form $\mathcal{K}_i(r, m)$ are sometimes called *information sets*.

In general, there are many possible executions of a system: there could be a number of possible initial states and many things that could happen from each initial state. For example, in a draw poker game, the initial global states could describe the possible deals of the hand by having player i's local state describe the cards held by player i. For each fixed deal of the cards, there may still be many possible betting sequences, and thus many runs. Formally, a *system* is a nonempty set of runs. Intuitively, these runs describe all the possible sequences of events that could occur in the system. Thus, I am essentially identifying a system with its possible behaviors.

There are a number of ways of modeling the Sleeping Beauty problem as a system. Perhaps simplest is to consider it as a single-agent problem,

since the experimenter plays no real role. (Note that it is important to have the environment though.) Assume for now that the system modeling the Sleeping Beauty problem consists of two runs, the first corresponding to the coin landing heads, and the second corresponding to the coin landing tails. (As we shall see, while restricting to two runs seems reasonable, it may not capture all aspects of the problem.) There are still some choices to be made with regard to modeling the global states. Here is one way. At time 0, a coin is tossed; the environment state encodes the outcome. At time 1, the agent is asleep (and thus is in a ''sleeping'' state). At time 2, the agent is woken up. If the coin lands tails, then at time 3, the agent is back asleep, and at time 4, is woken up again. Note that I have assumed here that time in both of these runs ranges from 0 to 5. Nothing would change if I allowed runs to have infinite length or a different (but sufficiently long) finite length.

Alternatively, we might decide that it is not important to model the time that the agent is sleeping; all that matters is the point just before the agent is put to sleep and the points where the agent is awake. Assume that Sleeping Beauty is in state b before the experiment starts, in state a after the experiment is over, and in state w when woken up. This leads to a model with two runs r_1 and r_2, where the first three global states in r_1 are (H, b), (H, w), and (H, a), and the first four global states in r_2 are (T, b), (T, w), (T, w), (T, a). Let $\mathcal{R}_1$ be the system consisting of the runs r_1 and r_2. This system is shown in Figure 5.1 (where only the first three global states in each run are shown). The

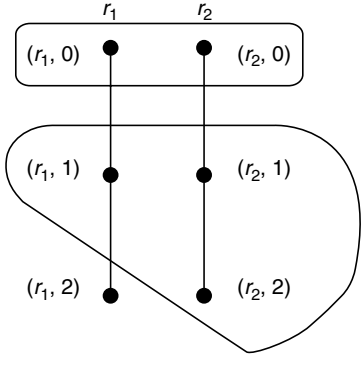

Figure 5.1. The Sleeping Beauty problem, captured using $\mathcal{R}_1$

three points where the agent's local state is w, namely, $(r_1, 1)$, $(r_2, 1)$, and $(r_2, 2)$, form what is traditionally called in game theory an *information set*. These are the three points that the agent considers possible when she is woken up. For definiteness, I use $\mathcal{R}_1$ in much of my analysis of Sleeping Beauty.

Notice that $\mathcal{R}_1$ is also compatible with a somewhat different story. Suppose that the agent is not aware of time passing. At time 0 the coin is tossed, and the agent knows this. If the coin lands heads, only one round passes before the agent is told that the experiment is over; if the coin lands tails, she is told after two rounds. Since the agent is not aware of time passing, her local state is the same at the points $(r_1, 2)$, $(r_2, 1)$, and $(r_2, 2)$. The same analysis should apply to the question of what the probability of heads is at the information set. The key point is that here the agent does not forget; she is simply unaware of the time passing.

Various other models are possible:

- We could assume (as Elga does at one point) that the coin toss happens only after the agent is woken up the first time. Very little would change, except that the environment state would be $\emptyset$ (or some other way of denoting that the coin hasn't been tossed) in the first two global states of both runs. Call the two resulting runs r'_1 and r'_2.
- All this assumes that the agent knows when the coin is going to be tossed. If the agent doesn't know this, then we can consider the system consisting of the four runs r_1, r'_1, r_2, r'_2.
- Suppose that we now want to allow for the possibility that, upon wakening, the agent learns that it is Monday (as in Elga's argument). To do this, the system must include runs where the agent actually learns that it is Monday. Now two runs no longer suffice. For example, we can consider the system $\mathcal{R}_2 = (r_1, r_2, r^*_1, r^*_2)$, where r^*_i is the same as r_i except that on Monday, the agent's local state encodes that it is Monday. Thus, the sequence of global states in r^*_1 is (H, b), (H, M), (H, a), and the sequence in r^*_2 is (T, b), (T, M), (T, w). $\mathcal{R}_2$ is described in Figure 5.2. Note that on Tuesday in r^*_2, the agent forgets whether she was woken up on Monday. She is in the same local state on Tuesday in r^*_2 as she is on both Monday and Tuesday in r_2.

Yet other representations of the Sleeping Beauty problem are also possible. The point that I want to emphasize here is that the framework

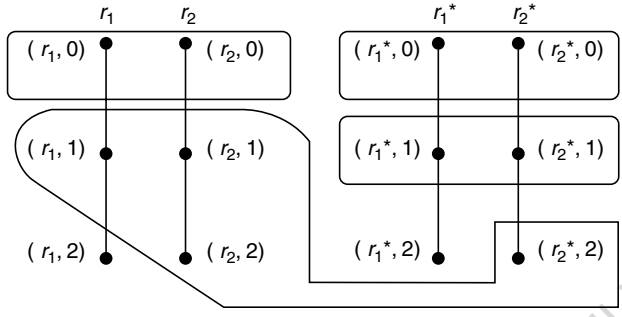

Figure 5.2. An alternative representation of the Sleeping Beauty problem, using $\mathcal{R}_2$

has the resources to capture important distinctions about when the coin is tossed and what agents know.

2.2. *Synchrony and Perfect Recall*

One advantage of the multiagent systems framework is that it can be used to easily model a number of important assumptions. I focus on two of them here: *synchrony*, the assumption that agents know the time, and *perfect recall*, the assumption that agents do not forget (Fagin *et al.* 1995; Halpern and Vardi 1989).

Formally, a system $\mathcal{R}$ is *synchronous for agent i* if for all points (r, m) and (r', m') in $\mathcal{R}$, if $(r, m) \sim_i (r', m')$, then $m = m'$. Thus, if $\mathcal{R}$ is synchronous for agent i, then at time m, agent i knows that it is time m, because it is time m at all the points he considers possible. $\mathcal{R}$ is *synchronous* if it is synchronous for all agents. Note that the systems that model the Sleeping Beauty problem are not synchronous. When Sleeping Beauty is woken up on Monday, she does not know what day it is.

Consider the following example of a synchronous system, taken from (Halpern 2003):

Example 2.1: Suppose that Alice tosses two coins and sees how the coins land. Bob learns how the first coin landed after the second coin is tossed,

but does not learn the outcome of the second coin toss. How should this be represented as a multiagent system? The first step is to decide what the local states look like. There is no ''right'' way of modeling the local states. What I am about to describe is one reasonable way of doing it, but clearly there are others.

The environment state will be used to model what actually happens. At time 0, it is $\langle\,\rangle$, the empty sequence, indicating that nothing has yet happened. At time 1, it is either $\langle H\rangle$ or $\langle T\rangle$, depending on the outcome of the first coin toss. At time 2, it is either $\langle H, H\rangle$, $\langle H, T\rangle$, $\langle T, H\rangle$, or $\langle T, T\rangle$, depending on the outcome of both coin tosses. Note that the environment state is characterized by the values of two random variables, describing the outcome of each coin toss. Since Alice knows the outcome of the coin tosses, I take Alice's local state to be the same as the environment state at all times.

What about Bob's local state? After the first coin is tossed, Bob still knows nothing; he learns the outcome of the first coin toss after the second coin is tossed. The first thought might then be to take his local states to have the form $\langle\,\rangle$ at time 0 and time 1 (since he does not know the outcome of the first coin toss at time 1) and either $\langle H\rangle$ or $\langle T\rangle$ at time 2. This choice would not make the system synchronous, since Bob would not be able to distinguish time 0 from time 1. If Bob is aware of the passage of time, then at time 1, Bob's state must somehow encode the fact that the time is 1. I do this by taking Bob's state at time 1 to be $\langle tick\rangle$, to denote that one time tick has passed. (Other ways of encoding the time are, of course, also possible.) Note that the time is already implicitly encoded in Alice's state: the time is 1 if and only if her state is either $\langle H\rangle$ or $\langle T\rangle$.

Under this representation of global states, there are seven possible global states:

- $(\langle\,\rangle, \langle\,\rangle, \langle\,\rangle)$, the initial state,
- two time-1 states of the form $(\langle X_1\rangle, \langle X_1\rangle, \langle tick\rangle)$, for $X_1 \in \{H, T\}$,
- four time-2 states of the form $(\langle X_1, X_2\rangle, \langle X_1, X_2\rangle, \langle tick, X_1\rangle)$, for $X_1, X_2 \in \{H, T\}$.

In this simple case, the environment state determines the global state (and is identical to Alice's state), but this is not always so.

The system describing this situation has four runs, $r^1, \ldots, r^4$, one for each of the time-2 global states. The runs are perhaps best thought of as being the branches of the computation tree described in Figure 5.3. ■

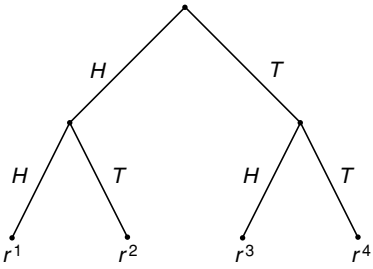

Figure 5.3. Tossing two coins

Modeling perfect recall in the systems framework requires a little care. In this framework, an agent's knowledge is determined by his local state. Intuitively, an agent has perfect recall if his local state is always ''growing'', by adding the new information he acquires over time. This is essentially how the local states were modeled in Example 2.1. In general, local states are not required to grow in this sense, quite intentionally. It is quite possible that information encoded in $r_i(m)$—i's local state at time m in run r—no longer appears in $r_i(m+1)$. Intuitively, this means that agent i has lost or ''forgotten'' this information. There are often scenarios of interest where it is important to model the fact that certain information is discarded. In practice, for example, an agent may simply not have enough memory capacity to remember everything he has learnt. Nevertheless, although perfect recall is a strong assumption, there are many instances where it is natural to model agents as if they do not forget.

Intuitively, an agent with perfect recall should be able to reconstruct his complete local history from his current local state. To capture this intuition, let *agent i's local-state sequence at the point* (r, m) be the sequence of local states that she has gone through in run r up to time m, without consecutive repetitions. Thus, if from time 0 through time 4 in run r agent i has gone through the sequence $\langle s_i, s_i, s_i', s_i, s_i\rangle$ of local states, where $s_i \neq s_i'$, then her local-state sequence at $(r, 4)$ is $\langle s_i, s_i', s_i\rangle$. Agent i's local-state sequence at a point (r, m) essentially describes what has happened in the run up to time m, from i's point of view. Omitting consecutive repetitions is intended to capture situations where the agent has perfect recall but is not aware of time passing, so she cannot distinguish a run where she stays in a given state s for three rounds from one where she stays in s for only one round.

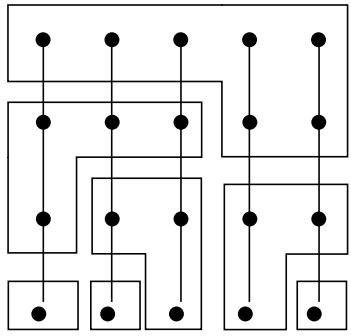

Figure 5.4. An asynchronous system where agent i has perfect recall

An agent has perfect recall if her current local state encodes her whole local-state sequence. More formally, *agent i has perfect recall in system $\mathcal{R}$* if, at all points (r, m) and (r', m') in $\mathcal{R}$, if $(r, m) \sim_i (r', m')$, then agent i has the same local-state sequence at both (r, m) and (r', m'). Thus, agent i has perfect recall if she ''remembers'' her local-state sequence at all times.[3] In a system with perfect recall, $r_i(m)$ encodes i's local-state sequence in that, at all points where i's local state is $r_i(m)$, she has the same local-state sequence. A system where agent i has perfect recall is shown in Figure 5.4.

The combination of synchrony and perfect recall leads to particularly pleasant properties. It is easy to see that if $\mathcal{R}$ is a synchronous system with perfect recall and $(r, m+1) \sim_i (r', m+1)$, then $(r, m) \sim_i (r', m)$. That is, if agent i considers run r' possible at the point $(r, m+1)$, then i must also consider run r' possible at the point (r, m). (Proof: since the system is synchronous and i has perfect recall, i's local state must be different at each point in r. For if i's local state were the same at two points (r, k) and (r, k') for $k \neq k'$, then agent i would not know that it was time k at the point (r, k). Thus, at the points $(r, m+1)$, i's local-state sequence must have length $m+1$. Since $(r, m+1) \sim_i (r', m+1)$, i has the same local-state sequence at $(r, m+1)$ and $(r', m+1)$. Thus,

[3] This definition of perfect recall is not quite the same as that used in the game theory literature, where agents must explicitly recall the actions taken (see Halpern (1997) for a discussion of the issues), but the difference between the two notions is not relevant here. In particular, according to both definitions, the agent has perfect recall in the ''game'' described by Figure 5.1.

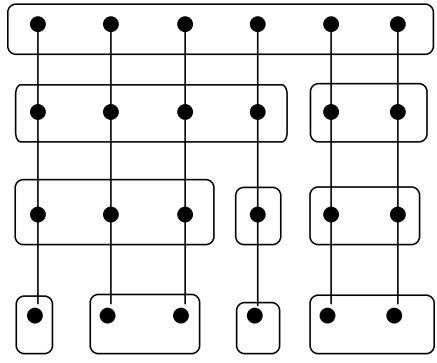

Figure 5.5. A synchronous system with perfect recall

i must also have the same local-state sequence at the points (r, m) and (r', m), since i's local-state sequence at these points is just the prefix of i's local-state sequence at $(r, m+1)$ of length m. It is then immediate that $(r, m) \sim_i (r', m)$.) Thus, in a synchronous system with perfect recall, agent i's information-set refines over time, as shown in Figure 5.5.[4]

Note that whether the agent has perfect recall in the Sleeping Beauty problem depends in part on how we model the problem. In the system $\mathcal{R}_1$ she does; in $\mathcal{R}_2$ she does not. For example, at the point $(r_2^*, 2)$ in $\mathcal{R}_2$, where her local state is (T, w), she has forgotten that she was woken up at time 1 (because she cannot distinguish $(r_2, 2)$ from $(r_2^*, 2)$). (It may seem strange that the agent has perfect recall in $\mathcal{R}_1$, but that is because in $\mathcal{R}_1$, the time that the agent is asleep is not actually modeled. It happens ''between the points''. If we explicitly include local states where the agent is asleep, then the agent would not have perfect recall in the resulting model. The second interpretation of $\mathcal{R}_1$, where the agent is unaware of time passing, is perhaps more compatible with perfect recall. I use $\mathcal{R}_1$ here so as to stress that perfect recall is not really the issue in the Sleeping Beauty problem; it is the asynchrony.)

[4] In the language of probabilists, in synchronous systems with perfect recall, information sets form a *filtration* (Billingsley 1986, Section 35). The importance of assuming that the information-sets form a filtration in the context of the Sleeping Beauty problem is emphasized by Schervish *et al.* (2004). However, my analysis in the asynchronous case applies despite the fact that the information-sets do not form a filtration.

3. ADDING PROBABILITY

To add probability to the framework, I start by assuming a probability on the set of runs in a system. Intuitively, this should be thought of as the agents' common probability. It is not necessary to assume that the agents all have the same probability on runs; different agents may use different probability measures. Moreover, it is not necessary to assume that the probability is placed on the whole set of runs. There are many cases where it is convenient to partition the set of runs and put a separate probability measure on each cell in the partition (see Halpern (2003) for a discussion of these issues). However, to analyze the Sleeping Beauty problem, it suffices to have a single probability on the runs. A *probabilistic system* is a pair $(\mathcal{R}, \Pr)$, where $\mathcal{R}$ is a system (a set of runs) and Pr is a probability on $\mathcal{R}$. (For simplicity, I assume that $\mathcal{R}$ is finite and that all subsets of $\mathcal{R}$ are measurable.) In the case of the Sleeping Beauty problem, the probability on $\mathcal{R}_1$ is immediate from the description of the problem: each of r_1 and r_2 should get probability 1/2. To determine a probability on the runs of $\mathcal{R}_2$, we need to decide how likely it is that the agent will discover that it is actually Monday. Suppose that probability is α. In that case, r_1 and r_2 both get probability $(1 - \alpha)/2$, while r_1^* and r_2^* both get probability $\alpha/2$.

Unfortunately, the probability on runs is not enough for the agent to answer questions like ''What is the probability that heads was tossed?'' if she is asked this question at the point $(r_1, 1)$ when she is woken up in $\mathcal{R}_1$, for example. At this point she considers three points possible: $(r_1, 1)$, $(r_2, 1)$, and $(r_2, 2)$, the three points where she is woken up. She needs to put a probability on this space of three points to answer the question. Obviously, the probability on the points should be related to the probability on runs. But how? That is the topic of this section.

As the preceding discussion should make clear, points can be viewed as possible worlds. In HT, a modal logics of knowledge and probability is considered where truth is defined relative to points in a system. Points are somewhat analogous to what Lewis (1979) calls *centered* possible worlds, since they are equipped with a time (although they are not equipped with a designated individual). Runs can then be viewed as uncentered possible worlds. Lewis (1979) argued that credence should be placed not on possible worlds, but on centered possible worlds. The key issue here is that, in many applications, it is more natural to start

with a probability on uncentered worlds; the question is how to define a probability on centered worlds.[5]

3.1. The Synchronous Case

Tuttle and I suggested a relatively straightforward way of going from a probability on runs to a probability on points in synchronous systems. For all times m, the probability Pr on $\mathcal{R}$, the set of runs, can be used to put a probability $\Pr^m$ on the points in $\mathcal{R}^m = \{(r, m): r \in \mathcal{R}\}$: simply take $\Pr^m(r, m) = \Pr(r)$. Thus, the probability of the point (r, m) is just the probability of the run r. Clearly, $\Pr^m$ is a well-defined probability on the set of time-m points. Since $\mathcal{R}$ is synchronous, at the point (r, m), agent i considers possible only time-m points. That is, all the points in $\mathcal{K}_i(r, m) = \{(r', m') : (r, m) \sim_i (r', m')\}$ are actually time-m points. Since, at the point (r, m), the agent considers possible only the points in $\mathcal{K}_i(r, m)$, it seems reasonable to take the agent's probability at the point (r, m) to the result of conditioning $\Pr^m$ on $\mathcal{K}_i(r, m)$, provided that $\Pr^m(\mathcal{K}_i(r, m)) > 0$, which, for simplicity, I assume here. Taking $\Pr_{(r,m,i)}$ to denote agent i's probability at the point (r, m), this suggests that $\Pr_{(r,m,i)}(r', m) = \Pr^m((r', m)|\mathcal{K}_i(r, m))$.

To see how this works, consider the system of Example 2.1. Suppose that the first coin has bias 2/3, the second coin is fair, and the coin tosses are independent, as shown in Figure 5.6. Note that, in Figure 5.6, the edges coming out of each node are labeled with a probability, which is intuitively the probability of taking that transition. Of course, the probabilities labeling the edges coming out of any fixed node must sum to 1, since some transition must be taken. For example, the edges coming out of the root have probability 2/3 and 1/3. Since the transitions in this case (i.e. the coin tosses) are assumed to be independent, it is easy to compute the probability of each run. For example, the probability of run r^1 is $2/3 \times 1/2 = 1/3$; this represents the probability of getting two heads.

[5] As a cultural matter, in the computer science literature, defining truth/credence relative to centered worlds is the norm. Computer scientists are, for example, interested in temporal logic for reasoning about what happens while a program is running (Manna and Pnueli 1992). Making time part of the world is necessary for this reasoning. Interestingly, economists, like philosophers, have tended to focus on uncentered worlds. I have argued elsewhere (Halpern 1997) that centered worlds (represented as points) are necessary to capture some important temporal considerations in the analysis of games.

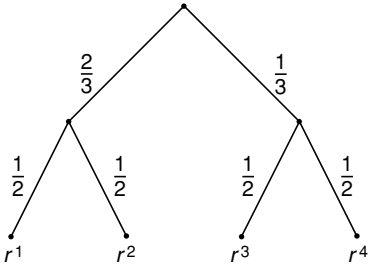

Figure 5.6. Tossing two coins, with probabilities

3.2. The General Case

The question now is how the agents should ascribe probabilities in arbitrary (not necessarily synchronous) systems, such as that of the Sleeping Beauty problem. The approach suggested above does not immediately extend to the asynchronous case. In the asynchronous case, the points in $\mathcal{K}_i(r, m)$ are not in general all time-m points, so it does not make sense to condition $\Pr^m$ on $\mathcal{K}_i(r, m)$. (Of course, it would be possible to condition on the time-m points in $\mathcal{K}_i(r, m)$, but it is easy to give examples showing that doing this gives rather nonintuitive results.)

I discuss two reasonable candidates for ascribing probability in the asynchronous case here, which are generalizations of the two approaches that Elga considers. I first consider these approaches in the context of the Sleeping Beauty problem, and then give the general formalization.

Consider the system described in Figure 5.1, but now suppose that the probability of r_1 is β and the probability of r_2 is $1 - \beta$. (In the original Sleeping Beauty problem, $\beta = 1/2$.) It seems reasonable that at the points $(r_1, 0)$ and $(r_2, 0)$, the agent ascribes probability β to $(r_1, 0)$ and $1 - \beta$ to $(r_2, 0)$, using the HT approach for the synchronous case. What about at each of the points $(r_1, 1)$, $(r_2, 1)$, and $(r_2, 2)$? One approach (which I henceforth call the *HT approach*, since it was advocated in HT), is to say that the probability β of run r_1 is projected to the point $(r_1, 1)$, while the probability $1 - \beta$ of r_2 is projected to $(r_2, 1)$ and $(r_2, 2)$. How should the probability be split over these two points? Note that splitting the probability essentially amounts to deciding the relative probability of being at time 1 and time 2. Nothing in the problem description gives us any indication of how to determine this. HT avoid making this

determination by making the singleton sets $\{(r_2, 1)\}$ and $\{(r_2, 2)\}$ nonmeasurable. Since they are not in the domain of the probability measure, there is no need to give them a probability. The only measurable sets in this space would then be $\emptyset$, $\{(r_1, 1)\}$, $\{(r_2, 1), (r_2, 2)\}$, and $\{(r_1, 1), (r_2, 1), (r_2, 2)\}$, which get probability 0, β, $1 - \beta$, and 1, respectively. An alternative is to apply the Principle of Indifference and take times 1 and 2 to be equally likely. In this case the probability of the set $\{(r_2, 1), (r_2, 2)\}$ is split over $(r_2, 1)$ and $(r_2, 2)$, and they each get probability $(1 - \beta)/2$. When $\beta = 1/2$, this gives Elga's first solution. Although it is reasonable to assume that times 1 and 2 are equally likely, the technical results that I prove hold no matter how the probability is split between times 1 and 2.

The second approach, which I call the *Elga approach* (since it turns out to generalize what Elga does), is to require that for any pair of points (r, m) and (r', m') on different runs, the relative probability of these points is the same as the relative probability of r and r'. This property is easily seen to hold for the HT approach in the synchronous case. With this approach, the ratio of the probability of $(r_1, 1)$ and $(r_2, 1)$ is $\beta : 1 - \beta$, as is the ratio of the probability of $(r_1, 1)$ and $(r_2, 2)$. This forces the probability of $(r_1, 1)$ to be $\beta/(2 - \beta)$, and the probability of each of $(r_1, 1)$ and $(r_2, 2)$ to be $(1 - \beta)/(2 - \beta)$. Note that, according to the Elga approach, if Pr is the probability on the runs of $\mathcal{R}_1$, $\beta = 1/2$, so that $\Pr(r_1) = \Pr(r_2) = 1/2$, and $\Pr'$ is the probability that the agent assigns to the three points in the information set, then

$$\begin{aligned}
&\Pr'((r_1, 1) \mid \{(r_1, 1), (r_2, 1)\}) \\
&= \Pr'((r_1, 1) \mid \{(r_1, 1), (r_2, 2)\}) \\
&= \Pr(r_1 \mid \{r_1, r_2\}) \\
&= 1/2.
\end{aligned}$$

Thus, we must have $\Pr'((r_1, 1)) = \Pr'((r_2, 1)) = \Pr'((r_2, 2))$, so each of the three points has probability 1/3, which is Elga's second solution. Moreover, note that

$$\Pr'((r_1, 1) \mid \{(r_1, 1), (r_2, 1)\}) = \Pr'((r_2, 1) \mid \{(r_1, 1), (r_2, 2)\}) = 1/2.$$

This is one way of formalizing the first step of Elga's argument; that is, that $\Pr'$ should have the property that, conditional on learning it is Monday, you should consider "it is now Monday and the coin landed heads" and "it is now Monday and the coin landed tails" equally likely.

The second step of Elga's argument used the Principle of Indifference to conclude that, if the coin landed tails, then all days were equally likely. That use of the Principle of Indifference is implicit in the assumption that the relative probability of (r_1, m) and (r_2, m) is the same for $m = 1$ and $m = 2$.

To summarize, the HT approach assigns probability among points in an information set I by dividing the probability of a run r among the points in I that lie on r (and then normalizing so that the sum is one), while the Elga approach proceeds by giving each and every point in I that is on run r the same probability as that of r, and then normalizing.

For future reference, I now give a somewhat more precise formalization of the HT and Elga approaches. To do so, it is helpful to have some notation that relates sets of runs to sets of points. If $\mathcal{S}$ is a set of runs and U is a set of points, let $\mathcal{S}(U)$ be the set of runs in $\mathcal{S}$ going through some point in U, and let $U(\mathcal{S})$ be the set of points in U that lie on some run in $\mathcal{S}$. That is,

$$\mathcal{S}(U) = \{r \in \mathcal{S}: (r, m) \in U \text{ for some } m\} \text{ and}$$
$$U(\mathcal{S}) = \{(r, m) \in U: r \in \mathcal{S}\}.$$

Note that, in particular, $\mathcal{K}_i(r, m)(r')$ is the set of points in the information set $\mathcal{K}_i(r, m)$ that are on the run r' and $\mathcal{R}(\mathcal{K}_i(r, m))$ is the set of runs in the system $\mathcal{R}$ that contain points in $\mathcal{K}_i(r, m)$. According to the HT approach, if $\Pr_i$ is agent i's probability on $\mathcal{R}$, the set of runs, then $\Pr^{HT}_{(r,m,i)}(\mathcal{K}_i(r, m)(r')) = \Pr_i(r' | \mathcal{R}(\mathcal{K}_i(r, m)))$. (Note that here I am using $\Pr^{HT}_{(i,r,m)}$ to denote agent i's probability at the point (r, m) calculated using the HT approach; I similarly will use $\Pr^{Elga}_{(i,r,m)}$ to denote agent i's probability calculated using the Elga approach.) That is, the probability that agent i assigns at the point (r, m) to the points in r' is just the probability of the run r' conditional on the probability of the runs going through the information set $\mathcal{K}_i(r, m)$. As I said earlier, Halpern and Tuttle do not try to assign a probability to individual points in $\mathcal{K}_i(r, m)(r')$ if there is more than one point on r' in $\mathcal{K}_i(r, m)$.

By way of contrast, the Elga approach is defined as follows:

$$\Pr^{Elga}_{(r,m,i)}(r', m') = \frac{\Pr_i(\{r'\} \cap \mathcal{R}(\mathcal{K}_i(r, m)))}{\Sigma_{r'' \in \mathcal{R}(\mathcal{K}_i(r,m))} \Pr_i(r'') \,|\, \mathcal{K}_i(r, m)(\{r''\})|}.$$

It is easy to check that $\Pr^{Elga}_{(r,m,i)}$ is the unique probability measure $\Pr'$ on $\mathcal{K}_i(r, m)$ such that $\Pr'((r_1, m_1))/\Pr'((r_2, m_2)) = \Pr_i(r_1)/\Pr_i(r_2)$ if

$\Pr_i(r_2) > 0$. Note that $\Pr^{Elga}_{(r,m,i)}$ assigns equal probability to all points on a run r' in $\mathcal{K}_i\ (r,\ m)$. Even if $\Pr^{HT}_{(r,m,i)}$ is extended so that all points on a given run are taken to be equally likely, in general, $\Pr^{HT}_{(r,m,i)} \neq \Pr^{Elga}_{(r,m,i)}$. The following lemma characterizes exactly when the approaches give identical results.

Lemma 3.1: $\Pr^{Elga}_{(r,m,i)} = \Pr^{HT}_{(r,m,i)}$ iff $|\mathcal{K}_i(r,\ m)\ (\{r_1\})| = |\mathcal{K}_i(r,\ m)\ (\{r_2\})|$ for all runs $r_1,\ r_2 \in \mathcal{R}(\mathcal{K}_i(r,\ m))$ such that $\Pr_i(r_j) \neq 0$ for $j = 1,\ 2$.

Note that, in the synchronous case, $|\mathcal{K}_i(r,\ m)\ (\{r'\})| = 1$ for all runs $r' \in \mathcal{R}(\mathcal{K}_i(r,\ m))$, so the two approaches are guaranteed to give the same answers.

4. COMPARING THE APPROACHES

I have formalized two approaches for ascribing probability in asynchronous settings, both of which generalize the relatively noncontroversial approach used in the synchronous case. Which is the most appropriate? I examine a number of arguments here.

4.1. Elga's Argument

Elga argued for the Elga approach, using the argument that if you discover or learn that it is Monday, then you should consider heads and tails equally likely. As I suggested above, I do not find this a compelling argument for the Elga approach. I agree that if you learn that it is Monday, you should consider heads and tails equally likely. On the other hand, Sleeping Beauty does not actually learn that it is Monday. Elga is identifying the probability of heads conditional on learning that it is Monday with the probability of heads given that it is Monday. While these probabilities could be equal, they certainly do not have to be. An example of Thomason makes the point nicely:[6] If I think my wife is much more clever than I, then I might be convinced that I will never learn of her infidelity should she be unfaithful. So, my conditional probability for *Y*, ''I will learn that my wife is cheating on

[6] Thanks to Jim Joyce for pointing out this example.

me'', given X, ''She will cheat on me'', is very low. Yet, the probability of Y if I actually learn X is clearly 1.[7]

In any case, in asynchronous systems, the two probabilities may be unequal for reasons beyond those that arise in the synchronous case. This is perhaps best seen by considering a system where the agent might actually learn that it is Monday. The system $\mathcal{R}_2$ described in Figure 5.2 is one such system. Note that in $\mathcal{R}_2$, even if the HT approach is used, if you discover it is Monday in run r_1^* or r_2^*, then you do indeed ascribe probability 1/2 to heads. On the other hand, in r_1 and r_2, where you do *not* discover it is Monday, you also ascribe probability 1/2 to heads when you are woken up, but conditional on it being Monday, you consider the probability of heads to be 2/3. Thus, using the HT approach, $\mathcal{R}_2$ gives an example of a system where the probability of heads given that it is Monday is different from the probability of heads conditional on learning that it is Monday.

Although $\mathcal{R}_2$ shows that Elga's argument for the 1/3–2/3 answer is suspect, it does not follow that 1/3–2/3 is incorrect. In the remainder of this section, I examine other considerations to see if they shed light on what should be the appropriate answer.

4.2. The Frequency Interpretation

One standard interpretation of probability is in terms of frequency. If the probability of a coin landing heads is 1/2, then if we repeatedly toss the coin, it will land heads in roughly half the trials; it will also land heads roughly half the time. In the synchronous case, ''half the trials'' and ''half the time'' are the same. But now consider the Sleeping Beauty problem. What counts as a ''trial''? If a ''trial'' is an experiment, then the coin clearly lands heads in half of the trials. But it is equally clear that the coin lands heads 1/3 of the times that the agent is woken up. Considering ''times'' and ''trials'' leads to different answers in asynchronous systems; in the case of the Sleeping Beauty problem, these

[7] There are other reasons why the probability of Y given X might be different from the probability of Y given that you learn or observe X. In the latter case, you must take into account how you came to learn that X is the case. Without taking this into account, you run into difficulties with, say, the Monty Hall problem. See Grünwald and Halpern (2003) for a discussion of this point in the synchronous setting. I ignore this issue here, since it is orthogonal to the issues that arise in the Sleeping Beauty problem.

different answers are precisely the natural 1/2–1/2 and 1/3–2/3 answers. I return to this issue in the next subsection.

4.3. Betting Games

Another standard approach to determining subjective probability, which goes back to Ramsey (1931) and De Finetti (1931), is in terms of betting behavior. For example, one way of determining the subjective probability that an agent ascribes to a coin toss landing heads is to compare the odds at which he would accept a bet on heads to one at which he would accept a bet on tails. While this seems quite straightforward, in the asynchronous case it is not. This issue was considered in detail in the context of the absented-minded driver paradox by Grove and Halpern (1997). Much the same comments hold here, so I just do a brief review.

Suppose that Sleeping Beauty is offered a $1 bet on whether the coin landed heads or the coin landed tails every time she is woken up. If the bet pays off every time she answers the question correctly, then clearly she should say ''tails''. Her expected gain by always saying tails is $1 (since, with probability 1/2, the coin will land tails and she will get $1 both times she is asked), while her expected gain by always saying heads is only 1/2. Indeed, a risk- neutral agent should be willing to pay to take this bet. Thus, even though she considers heads and tails equally likely and ascribes probabilities using the HT approach, this betting game would have her act as if she considered tails twice as likely as heads: she would be indifferent between saying ''heads'' and ''tails'' only if the payoff for heads was $2, twice the payoff for tails.

In this betting game, the payoff occurs at every time step. Now consider a second betting game, where the payoff is only once per trial (so that if the coin lands tails, the agent gets $1 if she says tails both times, and $0.50 if she says tails only once). If the payoff is per trial, then the agent should be indifferent between saying ''heads'' and ''tails''; the situation is analogous to the discussion in the frequency interpretation.

There is yet a third alternative. The agent could be offered a bet at only one point in the information-set. If the coin lands heads, she must be offered the bet at $(r_1, 1)$. If the coin lands heads, an adversary must somehow choose if the bet will be offered at $(r_2, 1)$ or $(r_2, 2)$. The third betting game is perhaps more in keeping with the second story told for $\mathcal{R}_1$, where the agent is not aware of time passing and must assign a

probability to heads and tails in the information set. It may seem that the first betting game, where the payoff occurs at each step, is more appropriate for the Sleeping Beauty problem—after all, the agent is woken up twice if the coin lands tails. Of course, if the goal of the problem is to maximize the expected number of correct answers (which is what this betting game amounts to), then there is no question that ''tails'' is the right thing to say. On the other hand, if the goal is to get the right answer ''now'', whenever now is, perhaps because this is the only time that the bet will be offered, then the third game is more appropriate. My main point here is that the question of the right betting game, while noncontroversial in the synchronous case, is less clear in the asynchronous case.

It is interesting to see how these issues play out in the context of Hitchcock's (2004) Dutch Book analysis of the Sleeping Beauty problem. As Hitchcock points out, there is a collection of bets that form a Dutch book, which can be offered by a bookie who knows no more than Sleeping Beauty provided Sleeping Beauty ascribes probability 1/2 to heads when she wakes up:[8]

- Before the experiment starts, Sleeping Beauty is offered a bet that pays off $30 if the coin lands tails and 0 otherwise, and costs $15. Since heads and tails are viewed as equally likely before the experiment starts, this is a fair bet from her point of view.
- Each time Sleeping Beauty is woken up, she is offered a bet that pays off $20 if the coin lands heads and 0 otherwise, and costs $10. Again, if Sleeping Beauty views heads and tails as equally likely when she is woken up, this bet is fair from her point of view.

Note that, if the coin lands heads, Sleeping Beauty is only woken up once, so she loses $15 on the first bet and has a net gain of $10 on the second bet, for an overall loss of $5. On the other hand, if the coin lands heads, Sleeping Beauty has a net gain of $15 on the first bet, but the second bet is offered twice and she has a loss of $10 each time it is offered. Thus, she again has a net loss of $5.

This Dutch Book argument is essentially dealing with bets that pay off at each time step, since if the coin lands tails, Sleeping Beauty loses

[8] The importance of taking the knowledge of the bookie into account, which is stressed by Hitchcock, is also one of the key points made by Halpern and Tuttle (1993). Indeed, it is argued by Halpern and Tuttle that probability does not make sense without taking the knowledge of the adversary (the bookie in this case) into account.

$10 each time she is woken up. By way of contrast, consider the following sequence of bets:

- Before the experiment starts, Sleeping Beauty is offered a bet that pays off $30 if the coin lands heads and 0 otherwise, and costs $15.
- Each time Sleeping Beauty is woken up, she is offered a bet that pays off $30 if the coin lands tails and 0 otherwise, and costs $20, with the understanding that the bet pays off only once in each trial. In particular, if the coin in fact lands tails, and Sleeping Beauty takes the bet both times she is woken up, she gets the $30 payoff only once (and, of course, only has to pay $20 for the bet once). The accounting is done at the end of the trial.

Note that the first bet is fair just as in the first Dutch Book, and the second bet is fair to an agent who ascribes probability 2/3 to tails when woken up, even though the payoff only happens once if the coin lands tails. Moreover, although the second bet is somewhat nonstandard, there is clearly no difficulty deciding when it applies and how to make payoffs. And, again, an agent who accepts all these bets will lose $5 no matter what happens.

4.4. Conditioning and the Reflection Principle

To what extent is it the case that the agent's probability over time can be viewed as changing via conditioning? It turns out that the answer to this question is closely related to the question of when the Reflection Principle holds, and gives further support to using the HT approach to ascribing probabilities in the asynchronous case.

There is a trivial sense in which updating is never done by conditioning. At the point (r, m), agent i puts probability on the space $\mathcal{K}_i(r, m)$; at the point $(r, m+1)$, agent i puts probability on the space $\mathcal{K}_i(r, m+1)$. These spaces are either disjoint or identical (since the indistinguishability relation that determines $\mathcal{K}_i(r, m)$ and $\mathcal{K}_i(r, m+1)$ is an equivalence relation). Certainly, if they are disjoint, agent i cannot be updating by conditioning, since the conditional probability space is identical to the original probability space. And if the spaces are identical, it is easy to see that the agent is not doing any updating at all; her probabilities do not change.

To focus on the most significant issues, it is best to factor out time by considering only the probability ascribed to runs. Technically, this

amounts to considering *run-based events*, that is sets U of points with the property that if $(r, m) \in U$, then $(r, m') \in U$ for all times m'. In other words, U contains all the points in a given run or none of them. Intuitively, we can identify U with the set of runs that have points in U. To avoid problems of how to assign probability in asynchronous systems, I start by considering synchronous systems. Given a set V of points, let $V^- = \{(r, m) : (r, m+1) \in V\}$; that is V^- consists of all the points immediately preceding points in V. The following result, whose straightforward proof is left to the reader, shows that in synchronous systems where the agents have perfect recall, the agents do essentially update by conditioning. The probability that the agent ascribes to an event U at time $m+1$ is obtained by conditioning the probability he ascribes to U at time m on the set of points immediately preceding those he considers possible at time $m+1$.

Theorem 4.1 (Halpern 2003): Let U be a run-based event and let $\mathcal{R}$ be a synchronous system where the agents have perfect recall. Then

$$\Pr_{r,m+1,i}(U) = \Pr_{r,m,i}(U \mid \mathcal{K}_i(r, m+1)^-).$$

Theorem 4.1 does not hold without assuming perfect recall. For example, suppose that an agent tosses a fair coin and observes at time 1 that the outcome is heads. Then at time 2 he forgets the outcome (but remembers that the coin was tossed, and knows the time). Thus, at time 2, because the outcome is forgotten, the agent ascribes probability 1/2 to each of heads and tails. Clearly, her time 2 probabilities are not the result of applying conditioning to her time 1 probabilities.

A more interesting question is whether Theorem 4.1 holds if we assume perfect recall and do not assume synchrony. Properly interpreted, it does, as I show below. But, as stated, it does not, even with the HT approach to assigning probabilities. The problem is the use of $\mathcal{K}_i(r, m+1)^-$ in the statement of the theorem. In an asynchronous system, some of the points in $\mathcal{K}_i(r, m+1)^-$ may still be in $\mathcal{K}_i(r, m+1)$, since the agent may not be aware of time passing. Intuitively, at time (r, m), we want to condition on the set of points in $\mathcal{K}_i(r, m)$ that are on runs that the agent considers possible at $(r, m+1)$. But this set is not necessarily $\mathcal{K}_i(r, m+1)^-$.

Let $\mathcal{K}_i(r, m+1)^{(r,m)} = \{(r', k) \in \mathcal{K}_i(r, m) : \exists m'((r, m+1) \sim_i (r', m'))\}$. Note that $\mathcal{K}_i(r, m+1)^{(r,m)}$ consists precisely of those points that agent

considers possible at (r, m) that are on runs that the agent still considers possible at $(r, m+1)$. In synchronous systems with perfect recall, $\mathcal{K}_i(r, m+1)^{(r,m)} = \mathcal{K}_i(r, m+1)^-$ since, as observed above, if $(r, m+1) \sim_i (r', m+1)$ then $(r, m) \sim_i (r', m)$. In general, however, the two sets are distinct. Using $\mathcal{K}_i(r, m+1)^{(r,m)}$ instead of $\mathcal{K}^-_{r,m+1}$ gives an appropriate generalization of Theorem 4.1.

Theorem 4.2 (Halpern 2003): Let U be a run-based event and let $\mathcal{R}$ be a system where the agents have perfect recall. Then,

$$\Pr^{HT}_{r,m+1,i}(U) = \Pr^{HT}_{r,m,i}(U \mid \mathcal{K}_i(r, m+1)^{(r,m)}).$$

Thus, in systems with perfect recall, using the HT approach to assigning probabilities, updating proceeds by conditioning. Note that since the theorem considers only run-based events, it holds no matter how the probability among points on a run is distributed. For example, in the Sleeping Beauty problem, this result holds even if $(r_2, 1)$ and $(r_2, 2)$ are not taken to be equally likely.

The analogue of Theorem 4.2 does not hold in general for the Elga approach. This can already be seen in the Sleeping Beauty problem. Consider the system of Figure 5.1. At time 0 (in either r_1 or r_2), the event heads (which consists of all the points in r_1) is ascribed probability 1/2. At time 1, it is ascribed probability 1/3. Since $\mathcal{K}_{SB}(r_1, 1)^{(r_1, 0)} = \{(r_1, 0),(r_2, 0)\}$, we have

$$1/3 = \Pr^{Elga}_{r_1,1,SB}(heads) \neq \Pr^{Elga}_{r_1,0,SB}(heads) | \mathcal{K}_{SB}(r_1, 1)^{(r_1,0)}) = 1/2.$$

The last equality captures the intuition that if Sleeping Beauty gets no additional information, then her probabilities should not change using conditioning.

Van Fraassen's (1995) Reflection Principle is a coherence condition connecting an agent's future beliefs and his current beliefs. Note that what an agent believes in the future will depend in part on what the agent learns. The Generalized Reflection Principle says that an agent's current belief about an event U should lie in the span of the agent's possible beliefs about U at some later time m. That is, if Pr describes the agent's current beliefs, and $\Pr_1, \ldots, \Pr_k$ describe the agent's possible beliefs at time m, then for each event U, $\Pr(U)$ should lie between $\min_j \Pr_j(U)$ and $\max_j \Pr_j(U)$. Savage's (1954) Sure-Thing Principle is

essentially a special case of the Generalized Reflection Principle. It says that if the probability of A is α no matter what is learnt at time m, then the probability of A should be α right now. This certainly seems like a reasonable criterion.

Van Fraassen (1995) in fact claims that if an agent changes his opinion by conditioning on evidence, that is, if $\Pr_j = \Pr(\cdot \mid E(j, m))$ for $j = 1, \ldots, k$, then the Generalized Reflection Principle must hold. The intuition is that the pieces of evidence $E(1, m), \ldots, E(k, m)$ must form a partition of underlying space (in each state, exactly one piece of evidence will be obtained), so that it becomes a straightforward application of elementary probability theory to show that if $\alpha_j = \Pr(E(j, t))$ for $j = 1, \ldots, k$, then $\Pr = \alpha_1 \Pr_1 + \cdots + \alpha_k \Pr_k$.

Van Fraassen was assuming that the agent has a fixed set W of possible worlds, and his probability on W changed by conditioning on new evidence. Moreover, he was assuming that the evidence was a subset of W. In the multiagent systems framework, the agent is not putting probability on a fixed set of worlds. Rather, at each time k, he puts probability on the set of worlds (i.e. points) that he considers possible at time k. The agent's evidence is an information set—a set of points. If we restrict attention to run-based events, we can instead focus on the agent's probabilities on runs. That is, we can take W to be the set of runs, and consider how the agent's probability on runs changes over time. Unfortunately, agent i's evidence at a point (r, m) is not a set of runs, but a set of points, namely $\mathcal{K}_i(r, m)$. We can associate with $\mathcal{K}_i(r, m)$ the set of runs going through the points in $\mathcal{K}_i(r, m)$, namely, in the notation of §3.2, $\mathcal{R}(\mathcal{K}_i(r, m))$

In the synchronous case, for each time m, the possible information sets at time m correspond to the possible pieces of evidence that the agent has at time m. These information sets form a partition of the time-m points, and induce a partition on runs. In this case, van Fraassen's argument is correct. More precisely, if, for simplicity, ''now'' is taken to be time 0, and we consider some future time $m > 0$, the possible pieces of evidence that agent i could get at time m are all sets of the form $\mathcal{K}_i(r, m)$, for $r \in \mathcal{R}$. With this translation of terms, it is an immediate consequence of van Fraassen's observation and Theorem 4.1 that the Generalized Reflection Principle holds in synchronous systems with perfect recall. But note that the assumption of perfect recall is critical here. Consider an agent that tosses a coin and observes that it lands heads at time 0. Thus, at time 0, she assigns probability 1 to the

event of that coin toss landing heads. But she knows that one year later she will have forgotten the outcome of the coin toss, and will assign that event probability 1/2 (even though she will know the time). Clearly Reflection does not hold.

What about the asynchronous case? Here it is not straightforward to even formulate an appropriate analogue of the Reflection Principle. The first question to consider is what pieces of evidence to consider at time m. While we can consider all the information sets of form $\mathcal{K}_i(r, m)$, where m is fixed and r ranges over the runs, these sets, as we observed earlier, contain points other than time-m points. While it is true that either $\mathcal{K}_i(r, m)$ is identical to $\mathcal{K}_i(r', m)$ or disjoint from $\mathcal{K}_i(r', m)$, these sets do *not* induce a partition on the runs. It is quite possible that, even though the set of points $\mathcal{K}_i(r, m)$ and $\mathcal{K}_i(r', m)$ are disjoint, there may be a run r'' and times m_1 and m_2 such that $(r'', m_1) \in \mathcal{K}_i(r, m)$ and $(r'', m_2) \in \mathcal{K}_i(r', m)$. For example, in Figure 5.4, if the runs from left to right are r_1–r_5, then $\mathcal{K}_{SB}(r_5, 1) = \{r_1, \ldots, r_5\}$ and $\mathcal{K}_{SB}(r_1, 1) = \{r_1, r_2, r_3\}$. However, under the assumption of perfect recall, it can be shown that for any two information sets $\mathcal{K}_i(r_1, m)$ and $\mathcal{K}_i(r_2, m)$, either (a) $\mathcal{R}(\mathcal{K}_i(r_1, m)) \cap \mathcal{R}(\mathcal{K}_i(r_2, m)) = \emptyset$, (b) $\mathcal{R}(\mathcal{K}_i(r_1, m)) \subseteq \mathcal{R}(\mathcal{K}_i(r_2, m))$, or (c) $\mathcal{R}(\mathcal{K}_i(r_2, m)) \subseteq \mathcal{R}(\mathcal{K}_i(r_1, m))$. From this it follows that there exists a collection $\mathcal{R}'$ of runs such that the sets $\mathcal{R}(\mathcal{K}_i(r', m))$ for $r' \in \mathcal{R}'$ are disjoint and the union of $\mathcal{R}(\mathcal{K}_i(r', m))$ taken over the runs $r' \in \mathcal{R}'$ consists of all runs in $\mathcal{R}$. Then the same argument as in the synchronous case gives the following result.

Theorem 4.3: If $\mathcal{R}$ is a (synchronous or asynchronous) system with perfect recall and $\mathcal{K}_i(r_1, m), \ldots, \mathcal{K}_i(r_k, m)$ are the distinct information sets of the form $\mathcal{K}_i(r', m)$ for $r' \in \mathcal{R}(\mathcal{K}_i(r, 0))$, then there exist $\alpha_1, \ldots, \alpha_k$ such that

$$\Pr_i(\cdot \mid \mathcal{R}(\mathcal{K}_i(r, 0))) = \sum_{j=1}^{k} \alpha_j \Pr_i(\cdot \mid \mathcal{R}(\mathcal{K}_i(r_j, m))).$$

The following corollary is immediate from Theorem 4.3, given the definition of $\Pr^{HT}_{(i,r,m)}$.

Corollary 4.4: If $\mathcal{R}$ is a (synchronous or asynchronous) system with perfect recall and $\mathcal{K}_i(r_1, m), \ldots, \mathcal{K}_i(r_k, m)$ are the distinct infor-

mation sets of the form $\mathcal{K}_i(r', m)$ for $r' \in \mathcal{R}(\mathcal{K}_i(r, 0)$, then there exist $\alpha_1, \ldots, \alpha_k$ such that for all $\mathcal{R}' \subseteq \mathcal{R}$,

$$\Pr^{HT}_{(i,r,0)}(\mathcal{K}_i(r, 0)(\mathcal{R}')) = \sum_{j=1}^{k} \alpha_j \Pr^{HT}_{(i,r_j,m)}(\mathcal{K}_i(r_j, m))(\mathcal{R}')).$$

Corollary 4.4 makes precise the sense in which the Reflection Principle holds for the HT approach. Although the notation $\mathcal{K}_i(r, m)(\mathcal{R}')$ that converts sets of runs to sets of points makes the statement somewhat ugly, it plays an important role in emphasizing what I take to be an important distinction that has largely been ignored. An agent assigns probability to points, not runs. At both time 0 and time m we can consider the probability that the agent assigns to the points on the runs in $\mathcal{R}'$, but the agent is actually assigning probability to quite different (although related) events at time 0 and time m. It is important to note that I am not claiming here that $\alpha_j = \Pr(\mathcal{R}(\mathcal{K}_i(r_j, m))$ in Theorem 4.3. While this holds in the synchronous case, it does not hold in general. The reason we cannot expect this to hold in general is that, in the synchronous case, the sets $\mathcal{R}(\mathcal{K}_i(r_j, m))$ are disjoint, so $\sum_{j=1}^{n} \Pr(\mathcal{R}(\mathcal{K}_i(r_j, m)) = 1$. This is not in general true in the asynchronous case. I return to this issue shortly.

The obvious analogue to Corollary 4.4 does not hold for the Elga approach. Indeed, the same example that shows conditioning fails in the Sleeping Beauty problem shows that the Reflection Principle does not hold. This example also shows that the Sure-Thing Principle fails. Using the Elga approach, the probability of heads (i.e. the probability of the points on the run where the coin lands heads) changes from 1/2 to 1/3 between time 0 and time 1, no matter what.

Arntzenius (2003) gives a number of other examples where he claims the Reflection Principle does not hold. In all of these examples, the agent either has imperfect recall or the system is asynchronous and the Elga approach is being used to ascribe probabilities. Thus, his observation may not seem surprising, given the previous analysis. However, in one case, according to my definition, Reflection in fact does not fail. This is due to the fact that I interpret Reflection in a slightly different way from Arntzenius. Since this example is of independent interest, I now consider it more carefully.

The example, credited by Arntzenius to John Collins, is the following: A prisoner has in his cell two clocks, both of which run perfectly accurately. However, clock A initially reads 6p.m. and clock B initially reads 7p.m. The prisoner knows that exactly one of the clocks is accurate; he believes that with probability 1/2 the accurate clock is clock A and with probability 1/2 it is clock B. The prisoner also knows that a fair coin has been tossed to determine if the lights go out at midnight; if it lands heads, they do, and if it lands tails, they stay on. Since the coin is fair, the prisoner initially places probability 1/2 on it landing heads.

There are four runs in the system corresponding to this problem, each of which has probability 1/4:

- r_1, where A is the accurate clock and the coin landed heads;
- r_2, where A is the accurate clock and the coin landed tails;
- r_3, where B is the accurate clock and the coin landed heads;
- r_4, where B is the accurate clock and the coin landed tails.

We can assume that the environment state encodes the true time and the outcome of the coin toss, while the prisoner's state encodes the clock readings and whether the light is off or on. Thus, a typical global state might have the form ((11.30, H), (11.30, 12.30, 1)). In this global state, the true time is 11.30 and the coin landed heads, clock A reads 11.30 (and is correct), clock B reads 12.30, and the light is on (denoted by the component 1 in the tuple). Thus, this is the global state at the point (r_1, 11.30). The other points in the same information set as (r_1, 11.30) are (r_2, 11.30) and (r_4, 12.30). Call this information set I_1. At all the three points in I_1, the prisoner's local state is (11.30, 12.30, 1). For future reference, note that the only other information set that includes time 11.30 points is $I_2 = \{(r_1, 10.30), (r_2, 10.30), (r_3, 11.30), (r_4, 11.30)\}$. At all the points in I_2, the pair of clocks read (10.30, 11.30) and the light is on.

It is easy to check that every information set has at most one point per run. It follows from Lemma 3.1 that, at every point, the HT approach and the Elga approach agree. Thus, no matter which approach is used, Reflection in the sense of Corollary 4.4 must hold. Observe that the prisoner's degree of belief that the coin landed heads in information set I_1 is 2/3, while in I_2 it is 1/2. Thus, the prisoner's initial probability of heads (1/2) is a convex combination of his possible probabilities of heads at 11.30, but the combination has coefficients 0 and 1. Taking the

coefficients to be 0 and 1 might seem a little strange. After all, why should we prefer I_2 so strongly? But I claim that the ''strangeness'' here is a result of carrying over inappropriate intuitions from the synchronous case. In the synchronous case, the coefficients reflect the probability of the information sets. This makes sense in the synchronous case, because the information sets correspond to possible pieces of evidence that can be obtained at time m, and the sum of these probabilities of the pieces of evidence is 1. However, in the asynchronous case, we cannot relate the coefficient to probabilities of obtaining evidence. Indeed, the ''evidence'' in the case of information set I_2 is that the clock readings are (10.30, 11.30) and the light is on. This is evidence that the prisoner initially knows that he will certainly obtain at some point (although not necessarily at 11.30). Indeed, it falls out of the analysis of Theorem 4.3 that it does not make sense to relate the coefficients in the asynchronous case to the probabilities of obtaining the evidence.

Arntzenius points out another anomaly in this example. Taking P_t to denote the prisoner's probability at (real) time t, Arntzenius observes that

$$\Pr_{7.00}(\text{clock } B \text{ is correct} \mid \Pr_{11.30}(\text{clock } B \text{ is correct}) = 1/3 = 0.$$

For $\Pr_{11.30}(\text{clock } B \text{ is correct}) = 1/3$ holds only in runs r_1 and r_2, since at the points (r_1, 11.30) and (r_2, 11.30), the prisoner's probability that B is correct is 1/3, while at the points (r_3, 11.30) and (r_4, 11.30), the prisoner's probability that B is correct is 1/2. On the other hand, B is not correct in runs r_1 and r_2, so the conditional probability is 0.

Arntzenius suggests that this is a problem, since the prisoner does not trust his later beliefs. I would argue that the prisoner should trust all his later beliefs that he is aware of. The trouble is, the prisoner has no idea when he has the belief $\Pr_{11.30}(\text{clock } B \text{ is correct}) = 1/3$, since he has no idea when it is 11.30. (Essentially the same point is made by Schervish *et al*. (2004).) Of course, in a synchronous system, an agent does know when 11.30 is, so beliefs of the form $\Pr_{11.30}(U)$ are ones he should trust.

Note that if we modify the problem very slightly so that (*a*) clock A gives the true time, (*b*) the lights will be turned off when the jailer's clock reads midnight, and (*c*) one of A and B gives the jailer's time, but the prisoner does not know which and ascribes each probability 1/2, then we get a synchronous system that is identical to Arntzenius's in all essential details. However, now Reflection is completely unproblematic. At 11.30, if the light is still on, the prisoner ascribes probability 1/3

to heads; if the light is off, the prisoner ascribes probability 1 to heads. Initially, the prisoner ascribes probability 3/4 to the light being on at 11.30 and probability 1/4 to the light being off. Sure enough $1/2 = 3/4 \times 1/3 + 1/4 \times 1$.

This example emphasizes how strongly our intuitions are based on the synchronous case, and how our intuitions can lead us astray in the presence of asynchrony. The prisoner has perfect recall in this system, so the only issue here is synchrony vs. asynchrony.

5. CONCLUSION

In this paper, I have tried to take a close look at the problem of updating in the presence of asynchrony and imperfect recall. Let me summarize what I take to be the main points of this paper:

- It is important to have a good formal model that incorporates uncertainty, imperfect recall, and asynchrony in which probabilistic arguments can be examined. While the model I have presented here is certainly not the only one that can be used, it does have a number of attractive features. As I have shown elsewhere (Halpern 1997), it can also be used to deal with other problems involved with imperfect recall raised by Piccione and Rubinstein (1997).
- Whereas there seems to be only one reasonable approach to assigning (and hence updating) probabilities in the synchronous case, there are at least two such approaches in the asynchronous case. Both approaches can be supported using a frequency interpretation and a betting interpretation. However, only the HT approach supports the Reflection Principle in general. In particular, the two approaches lead to the two different answers in the Sleeping Beauty problem.
- We cannot necessarily identify the probability conditional on U with what the probability would be upon learning U. This identification is being made in Elga's argument; the structure $\mathcal{R}_2$ shows that they may be distinct.

One fact that seems obvious in light of all this discussion is that our intuitions regarding how to do updating in asynchronous systems are rather poor. This is clearly a topic that deserves further investigation.

ACKNOWLEDGMENTS

Thanks to Moshe Vardi for pointing out Elga's paper, to Teddy Seidenfeld for pointing out Arntzenius's paper, to Moshe, Teddy, Oliver Board, and Sergiu Hart for stimulating discussions on the topic, and to Oliver, Moshe, Adam Elga, Alan Hájek, James Joyce, Kevin O'Neill, and two anonymous reviewers for the Ninth Conference on Knowledge, Reasoning and Representation for a number of useful comments on an earlier draft of the paper.

REFERENCES

Arntzenius, F. (2003) 'Some Problems for Conditionalization and Reflection', *Journal of Philosophy*, 100: 356–70.

Billingsley, P. (1986) *Probability and Measure*, 3rd edn. (New York: Wiley).

De Finetti, B. (1931) 'Sul significato soggestivo del probabilità', *Fundamenta Mathematica*, 17: 298–329.

Dorr, C. (2002) 'Sleeping Beauty: In Defence of Elga', *Analysis*, 62: 292–6.

Elga, A. (2000) 'Self-Locating Belief and the Sleeping Beauty Problem', *Analysis*, 60(2) 143–7.

Fagin, R., J. Y. Halpern, Y. Moses, and M. Y. Vardi (1995) *Reasoning about Knowledge* (Cambridge, Mass.: MIT Press). A slightly revised paperback version was published in 2003.

Grove, A. J., and J. Y. Halpern (1997) 'On the Expected Value of Games with Absentmindedness', *Games and Economic Behavior*, 20: 51–65.

Grünwald, P. D., and J. Y. Halpern (2003) 'Updating Probabilities', *Journal of A.I. Research*, 19: 243–78.

Halpern, J. Y. (1997) 'On Ambiguities in the Interpretation of Game Trees', *Games and Economic Behavior*, 20: 66–96.

——(2003) *Reasoning about Uncertainty* (Cambridge, Mass.: MIT Press).

——and R. Fagin (1989) 'Modelling Knowledge and Action in Distributed Systems', *Distributed Computing*, 3(4): 159–79.

——and M. R. Tuttle (1993) 'Knowledge, Probability, and Adversaries', *Journal of the ACM*, 40(4): 917–62.

——and M. Y. Vardi (1989) 'The Complexity of Reasoning about Knowledge and Time, I: Lower Bounds', *Journal of Computer and System Sciences*, 38(1): 195–237.

Hitchcock, C. (2004) 'Beauty and the Bets', *Synthese*, 139: 405–20.

Lewis, D. (1979) 'Attitudes *de dicto* and *de se*', *Philosophical Review*, 88(4): 513–43.

Lewis, D. (2001) 'Sleeping Beauty: Reply to Elga', *Analysis*, 61: 171–6.
Manna, Z., and A. Pnueli (1992) *The Temporal Logic of Reactive and Concurrent Systems: Specification* (Berlin and New York: Springer-Verlag).
Monton, B. (2002) 'Sleeping Beauty and the Forgetful Bayesian', *Analysis*, 62: 47–53.
Piccione, M., and A. Rubinstein (1997) 'On the Interpretation of Decision Problems with Imperfect Recall', *Games and Economic Behavior*, 20(1): 3–24.
Ramsey, F. P. (1931) 'Truth and probability', in R. B. Braithwaite (ed.), *The Foundations of Mathematics and Other Logical Essays* (London: Routledge & Kegan Paul), 156–98.
Savage, L. J. (1954) *Foundations of Statistics* (New York: Wiley).
Schervish, M. J., T. Seidenfeld, and J. B. Kadane (2004) 'Stopping to Reflect', *Journal of Philosophy*, 101: 315–22.
van Fraassen, B. C. (1984) 'Belief and the Will', *Journal of Philosophy*, 81: 235–45.
——(1995) 'Belief and the Problem of Ulysses and the Sirens', *Philosophical Studies*, 77: 7–37.

6. Doubt, Deference, and Deliberation: Understanding and Using the Division of Cognitive Labor

Frank Keil

In 'The ''Meaning of Meaning'' ' Hilary Putnam (1975) famously suggested, as part of a more general defense of an externalist account of semantic content, that our referential practices are upheld by a 'division of linguistic labor'. A speaker who lacks the individual capacity to identify a term's referent may nonetheless use the term successfully, so long as she belongs to a linguistic community where some (group of) expert(s) have that capacity. Over the past thirty years, there has been a lively discussion about the implications of such a theory for questions about wide vs. narrow content, internalism vs. externalism about meaning, and the like. In the spirit of labor-division, I will leave discussion of these matters to others (see e.g. Burge 1979; Fodor 1998; Prinz 2002), focusing instead on a question that has received relatively little attention in either the philosophical or the psychological literature, namely how laypeople understand the nature and character of the division of cognitive labor.

In particular, I will consider how non-experts understand the ways in which knowledge might cluster in other minds. I will describe four distinct ways that people might think about the division of cognitive labor and say something about how those different ways are used to make sharply contrasting inferences about domains of expertise. Although there is evidence that all four ways are available quite early in cognitive development, there are also striking differences in how they are used at various ages. The kind of expertise that Putnam implied as guiding deference for the meanings of natural kind terms, namely that of the natural sciences, gradually comes to hold a privileged status during middle childhood. This pattern of developmental change in turn sheds light on the everyday value that attaches to having insight into the division of cognitive labor.

DIVISIONS OF LABOR

It is hardly news that cultures divide up chores in ways that create different areas of expertise. As cultures became less nomadic, crafts and skills emerged with distinctive experts in each. Economists and sociologists have long argued that divisions of labor are an essential part of increasing productivity in a culture (Smith 1776; Durkheim 1947; Hume 1739). In most human cases, divisions of physical labor carry with them implications for divisions of cognitive labor. A person who achieves greater skill in an area is likely to have distinctive cognitive capacities that support that skill. In addition, most divisions of cognitive labor in humans reflect different paths of learning, different experiences, and immersion in different local communities of knowledge. Given its pervasiveness across cultures, it is surprising that there has been relatively little work in the field of cognitive anthropology devoted to the cognitive bases of divisions of labor (Hutchins 1995).

Important psychological questions arise concerning the division of cognitive labor. How do most collective enterprises, such as the basic sciences, engineering, legal systems, and medicine, function when each individual only has a fraction of the necessary knowledge and understanding to make the whole enterprise work? In particular, how does one access a domain of knowledge in other minds when one is largely ignorant about that domain? If we know already that an individual has one piece of knowledge, how do we decide what else that person is likely to know? How do we decide which of two competing experts is more likely to be a source of correct information?

The answers to such questions open up several topics that overlap with the field of 'social epistemology' (Goldman 1999, 2001). For the most part, they are also beyond the scope of this paper, as are questions about how members of a scientific community divide up their labor (Kitcher 1990). Instead, the more narrow goal of this paper is to consider the psychological heuristics that people use to think about how knowledge might be clustered in other minds. What do we need to know outside our own areas of expertise to be able to expand on our knowledge in those unfamiliar areas?

There are several distinct ways of thinking of how knowledge might be clustered in other minds, ways that draw on different sorts of cognitive requirements and which can be explored through experimental studies. Since detailed descriptions of those studies are under way

elsewhere in journals with a more experimental focus (Danovitch and Keil 2004; Lutz and Keil 2002; Keil 2003*a*), the focus of this paper will be on elaborating four distinct ways of thinking about expertise, summarizing the main findings of the experimental studies conducted by my laboratory group with adults and children, and considering how our developing understanding of the division of cognitive labor might be used in everyday life.

I will consider four ways of thinking about expertise: by category association, by privileged access, by goal implementation, and by underlying causal structure. These four possibilities do not exhaust the set of ways of thinking about knowledge clusters but they are the four most commonly used by laypeople. Moreover, they each suggest quite different heuristics for figuring out who knows what.

CATEGORY ASSOCIATION DIVISIONS OF KNOWLEDGE

Expertise can be understood as about anything normally associated with a category, providing that the categories involved are at the basic level of categorization or below. The basic level of categorization is the highest level at which categories seem to bristle with correlated properties not found at the next level up (Rosch *et al*., 1976; Murphy 2003). These levels can vary somewhat across individuals and cultures, but normally would be at a level of chairs, tables and sofas and below that of furniture. Similarly, shirts, pants and sweaters are the basic level below that of clothing, and cars, trucks and motorcycles from a basic level below that of vehicles. The basic level is also the level of categorization at which children also tend to use their first words to pick out sets of things in the world (Mervis and Crisafi 1982).

The category association heuristic assumes that people have knowledge clusters consisting of all pieces of information normally associated with members of a low-level category. Thus, one might plausibly think of people who are chair, or motorcycle, or pants experts. Even more plausibly, one can think of experts at levels below the basic level, such as Hitchcock chair, off road motorcycle, and ski pants experts. The basic level is the highest level at which we might normally employ the category association strategy. It is less plausible, however, to think of thorough experts on all kinds of furniture, or vehicles, or clothing. The lower the level, the more one might plausibly think that a person knows

most anything associated with a category. Thus, a ski pants expert might be expected to know the history of ski pants design, the costs of ski pants, which celebrities and racers wear what kinds of pants, and so on. A clothing expert could hardly be expected to have comparable diversity and detail of knowledge about all clothing.

Clustering of knowledge by category association might employ a simple cognitive heuristic. One merely needs to think of all bits of information that are normally associated with most members of that category. If I want to know something more about off road motorcycles, I might look for a person who demonstrates detailed knowledge about a few aspects of motorcycles and assume all other details will be known as well. This knowledge is perhaps best captured by the idea of people who are 'fans' or 'fanatics'. Elvis fans might be thought to know everything about Elvis, ranging from his songs, to his personal life, to the places he lived. Train fanatics might know everything about the history of trains, the ways trains worked, and the economic factors associated with trains. At a sufficiently low level of categorization, we might think it plausible that expertise could consist in having exhaustive knowledge of members of the category.

Where does this heuristic come from? It may arise from a social motivational hypothesis that people develop intense likes and dislikes for some categories; and, as a result, are deeply interested in everything frequently associated with most members of that category. We infer a drive to know 'everything' about a category either because it is highly valued or because it is a source of morbid fascination. The category association heuristic may also arise from the apparent ease of using a related strategy involving common lexical items. If John knows that 'Poodles' are F_1, where F_1 is an unusually detailed fact about poodles, simply assume that John is likely to know that 'Poodles' are F_n for any fact about poodles. Without knowing anything more about John or poodles one can blindly use the strategy of assuming that John is likely to have greater than average knowledge of the truth of virtually any sentence that makes a statement about 'poodles'. It would also be trivial to implement this strategy in a simple computer program that is fed text strings the size of sentences. If the category is low enough, a person's knowledge can be considered as exhausting everything that is typically associated with members of those categories or mentioned in discourse about lexical items that refer to that category.

Category exhaustion is interesting because it seems to be the simplest and most straightforward way of figuring out who knows what. The seductive simplicity of this heuristic makes it especially attractive to young children and to adults in cognitively loaded tasks. Thus, if one puts individuals under powerful time pressures, has them do several things at once, or inserts salient distracters in a task, these cognitive 'loads' tend to cause people to abandon more difficult cognitive heuristics in favor of simpler ones. Though subjects may not reveal their reliance on these heuristics in less pressured settings, cognitive load tasks can help experimenters identify which simple heuristics play a role in their everyday cognitive processing.

The category association approach, however, can be seriously misleading for one straightforward reason. It is virtually never the case people have exhaustive knowledge of members of a category, no matter how low the level. Moreover, as seen shortly, this strategy fails to predict other sorts of important elements of knowledge that can be reliably inferred from a few things that a person knows.

'ACCESS-BASED' DIVISIONS OF KNOWLEDGE

The socio-economic or subcultural practices of a society can often be used to think of divisions of cognitive labor that are 'access-based'. Thus, we can assume that different groups of people have different forms of expertise because they have been in proximity to a particular form of information that others have not by virtue of their station in life. For example, one might infer that a person who knows more than average about fine wines, resort spas, and charter jets, has that knowledge by virtue of being wealthy and therefore one might also expect that person to have greater knowledge about designer clothing, plastic surgeons, and home security systems. A person who knows more than average about soup kitchens, friendly police precincts, and warm heating vents may have that knowledge by virtue of being homeless and therefore is expected to have greater knowledge about homeless shelters and places with low and high rates of pedestrian traffic.

An understanding of access-based knowledge requires some sense of how people cluster in stable or semi-stable groups in a culture and what bits of information might be distinctive to those groups. This knowledge

is not based on a category or category label but rather on an understanding of the distinctive environments of subgroups and of the experiences offered by those environments. It might be based on simple associations of activities with members of that group, or it might involve induction of totally novel forms of knowledge based on an understanding of the group and why it coheres as such. Thus, if one believes that the wealthy tend to pick activities that are exclusive by virtue of the expenses associated with engaging in those activities, one can induce that wealthy people are more likely to know about some novel but highly expensive product. In this way, an understanding of the division of cognitive labor on the basis of access can have a generative quality

This generative property helps illustrate why access-based models of expertise are not variants of the category exhaustion strategy applied to the special case of social categories. When one relies on beliefs about why and how a group of people choose activities, the ability to then induce a large set of new forms of expertise contrasts with a mere list of all facts associated with the members of the category. Moreover, access-based strategies also exclude some forms of knowledge that might be associated with a category but which do not follow from causal explanatory beliefs about a particular kind of access. For example, wealthy individuals in the United States are more likely to know about local Republican politicians because of a strong association between wealth and support of Republicans (Green *et al*. 2002); but the access-based heuristic of expertise described earlier for wealth relies on the notion of increased knowledge of expensive goods and activities and might not see the relevance of party affiliation.

'GOAL-CENTERED' DIVISIONS OF KNOWLEDGE

Different people have different relatively long-term goals. One person may want to play professional soccer, another to heal the sick, and another run a successful fish charter business. Knowledge of another's goals, plus some knowledge of how those goals are normally achieved, can also be used as a basis for inferring clusters of knowledge in other minds. Thus, a person whose goal is to run a successful fish charter business might be expected to know more than average about topics that would further the goal of having a large number of customers in a financially viable manner. That person is likely to know more than

average about fish seasonal migration and foraging patterns, about marine weather, about diesel engines, about marine navigation, and about financing and insurance for commercial boats.

Goal-centered ways of understanding the division of cognitive labor are far-ranging and potentially powerful. They tend to go far beyond mere association of bits of information with individuals to a kind of problem-solving about what it takes to be successful in an endeavor and how the structure of a situation, such as a boat in a marine environment with customers, imposes certain requirements that in turn call on specific forms of expertise. The more one knows about the environment and about human capacities in such environments the more one can generate inductions about likely bits of knowledge in that area. An account of goal-based heuristics requires both a first order analysis of the knowledge of the goal-directed agent (e.g. the fishing boat skipper) and a second order analysis of the knowledge that one might have of goal-directed agents and their likely knowledge.

Goal-based ways of clustering of knowledge would seem to be those most closely associated with how the division of physical labor evolved in societies. Weavers, potters, farmers, and healers all developed expertise that furthered their relatively straightforward, and usually very public, goals. To infer who knows what in the world, one needs to keep track of different goals of groups of individuals and note how those goals are normally achieved. Even knowledge of a completely novel goal can often yield quite fertile inductions about knowledge. There is, for example, a group of individuals known as 'disk recovery specialists', whose goal is to recover data from computer hard drives that have become inoperable. I had never heard of that group until quite recently, yet a simple knowledge of their goal allowed me to induce what those professionals are likely to know about: how hard drives work, a huge array of software and disk operating systems, market rates for data recovery, and legally binding contracts between specialists and clients.

Sometimes, the goal becomes subordinate to a causal understanding of a set of closely related phenomena associated with that goal. For example, suppose one's goal is to treat cancer. As one pursues that goal, the biology of cancer starts to loom larger than the goal itself, which depends largely on an in-depth understanding of the relevant biology. Indeed, many of the sciences as we know them today grew out of goal-based practices, in which a rich pattern of causal regularities became far more the focus of knowledge than the goal itself. The goal of

transforming base elements into gold or silver was unattainable but led to an increased understanding of chemistry. The goals of breeding better crops, livestock, and pets led to a greater understanding of the biology of genetics. Some goals, such as those of the fishing charter business, intrinsically draw on many domains at once and continue to do so in their most advanced and refined forms; but others bring into relief the causal patterns and regularities of a particular science, which leads to the last way of understanding the clustering of knowledge.

CAUSAL-CLUSTER, OR DISCIPLINE-BASED, DIVISIONS OF KNOWLEDGE

For many academics, especially those in the natural and social sciences, the most obvious ways of clustering knowledge is by academic disciplines, with the additional assumption that such disciplines arise because of distinctive patterns of causal regularities in the world. Departments of biology, chemistry, physics, and psychology are often said to exist because there are special causal patterns that are signatures of each of those areas. We tend to assume that there is a relatively small set of core principles that govern much of what happens in a domain and that, by virtue of knowledge of those principles, people have greater than average knowledge of the indefinitely large number of phenomena arising from such principles. The canonical case is knowledge of Newtonian mechanics. We assume that a person who knows Newton's laws of motion and a certain level of mathematics is likely to understand virtually any set of interactions between bounded physical objects. (We may mistakenly underestimate the complexity of some multibodied systems, but the assumption as described is quite common.) Many scientists similarly assume there are comparable sets of principles underlying chemistry, biology, and other disciplines with further subdivisions within that form hierarchies of subdisciplines.

Understanding knowledge clusters in terms of underlying causal patterns might seem to be a rarified way of thinking about the division of cognitive labor. Perhaps it is a recent cultural invention that is only within the strong grasp of scientists. Wouldn't one need relatively sophisticated exposure to those causal patterns to be able to appreciate how they might be used as a way of understanding of the organization of knowledge in other minds? A brief consideration of some different

versions of realism and how each might influence the division of cognitive labor helps one see why it might be otherwise and how psychological studies are relevant.

All forms of realism embrace the idea that there are enduring patterns of regularities in the world independent of human activities on that world. They differ, however, in the extent to which they see a world of fundamentally distinct sorts of regularities. Consider the contrast between the view that all of nature is reducible to an account couched in terms of the laws of physics and the view that there are distinct levels of explanation such that the laws of economics, for example, cannot be reduced to those physics (Fodor 1974). Antireductionist views would seem most naturally associated with a division of cognitive labor corresponding to each of the levels of explanation they embrace. Reductionist approaches, in contrast, need not make such commitments.

Even at the same level of explanation, realists can debate about the extent to which the world should be seen as a relatively homogeneous network of causal links between properties, or whether it should be seen as more heterogeneous, consisting of distinct causal patterns with their own architectural principles. Should the world be seen as 'dappled' with different clusters of regularities or as more consistently all of the same type (Cartwright 1999)?

A dappled world-view offers a natural way of explaining how different realms of expertise might emerge, especially one that endorses 'thick' causal relations in which the causal relations such as 'compress', 'support', 'allow', 'feed', and 'prevent' are thought to be intrinsically different from each other and not reducible to a generic notion of cause (Cartwright 2003). Different realms might have different clusters of thick causal relations typically associated with them as well as different ways of describing the interactions between those relations. Perhaps one domain, such as evolutionary theory, uses intrinsically statistical arguments while another, such as the mechanics of macroscopic bounded objects, does not. Expertise in one of these domains might therefore be compartmentalized and not easily generalized to another.

Realists can also debate the extent to which there is a privileged way of carving up the world as opposed to an indefinitely large number of alternative ways, each of which might be based in a different form of real world structure and process. For example, laypeople often assume that there are two distinct natural kinds corresponding to 'trees' and 'flowers'. In most of the biological sciences, however, the tree/flower

contrast is meaningless. Daisies and apple trees are much more similar to each other in terms of microstructural properties, evolutionary ancestors, and DNA structure than apple trees and pine trees. Pine trees, in turn, are more similar to ferns than they are to oak trees (Dupre 1981). At the same time, the layperson is picking up on a real physical difference between trees and flowers. Indeed, computer simulations of how ancient plants would solve the problem of growing taller to get more light all tend to converge on structural solutions similar to modern trees with stout reinforced trunks and root structures and certain overall shapes that maximize light exposure to surfaces (Niklas 1996). There are two different sets of causal regularities that give rise to different sets of stable kinds, each of which might be stable because of its own form of causal homeostasis (Boyd 1999).

One can take the tree/flower case as suggesting a 'promiscuous realism' in which there are indefinitely many realities that can be articulated over the same class of entities (Boyd 1999). This view can, in turn, devolve into a form of social constructionism in which real world structure becomes arbitrary and where human convention and invention fully explain the domains of scientific inquiry (Hacking 1999; Kukla 2000). A more nuanced view sees the sciences as akin to the making of maps (Kitcher 2001). Maps are correct, or true, by virtue of their correspondence with some set of relations in the world; but even given that strong commitment to realism, there are many such correspondences. (Just consider all the different kinds of maps one can have of a large city.) Thus, the map metaphor illustrates how the relationship between the causal structure of the world and domains of expertise, while quite varied, is not arbitrary. Intuitions about domains of expertise may also arise from social constructions or from innate biases about domains of inquiry; but there are versions of realism in which persistent causal regularities give rise to families of maps corresponding to domains of expertise. Studies on the psychological mechanisms people use to ascertain the division of cognitive labor therefore not only have the potential to inform how we access and rely on knowledge in other minds but also to shed some light on how our knowledge of the world might be related to the structure of that world. Moreover, if laypeople and children are able to pick up on those patterns of causal regularities, they might have some insight into domains of expertise roughly corresponding to the natural and social sciences without ever having direct exposure to those sciences.

INTUITIONS ABOUT WHO KNOWS WHAT

One way to explore intuitions about clusters of knowledge would be to simply ask people for their intuitions of what the scientists and other experts know; but such free-form questioning tends to yield a diverse and unstructured body of information about all activities associated with scientists. In our laboratory we have taken a more focused strategy, (Danovitch and Keil 2004; Lutz and Keil 2002). I describe phenomena that a person understands well and then ask what other phenomena that person also understands by virtue of understanding those initial phenomena. Most often this technique has been done as a triad task in which a person is described as knowing a great deal about a particular phenomenon and is then asked which of two other phenomena the person is also likely to know about. By presenting a forced choice between two alternatives it is possible to create various contrasts, or minimal pairs, that allow one to explore the relative 'pulls' of different dimensions. Thus, the format is typically of the form:

> John knows a great deal about why P1.
> What else is he likely to know about?
> Why P2?
> *or*
> Why P3?

For example:

> John knows a great deal about why water is transparent to light.
> What else is he likely to know about?
> Why gold conducts electricity so well?
> *or*
> Why gold prices rise in times of high inflation?

This sort of technique arguably reflects a common, real-life, way in which we attempt to rely on the division of cognitive labor. When trying to understand which of several possible people to approach so as to acquire a better understanding of a phenomenon, we will take as important data what each of those people already know, seeking out the relevant dimension of similarity between their known knowledge and the new phenomenon.

Several questions arise with respect to people's intuitions about triads of this sort. To what extent do people need explicit access to the causal

mechanism themselves to be able to make a judgment of knowledge clustering? Is the structure of scientific disciplines in the modern university related in any way to laypeople's intuitions about knowledge clusters? How successful are laypeople at using underlying causal principles to cluster knowledge as opposed to clustering by surface objects, access, or goals? Finally, how do such patterns of judgment vary across development and across cultures? A series of studies have begun to provide answers to these questions.

In several studies, we described eight domains: physical mechanics, chemistry, adaptive/ecological biology, physiological biology, cognitive psychology, social psychology, economic and political science.[1] The divisions we chose correspond to distinct departments in at least some universities, although the two subareas of biology and psychology are often collapsed together. This particular group of eight was chosen because it could be placed into a neat symmetrical hierarchy of the natural and social sciences which are then further divided into the physical and biological sciences and the psychological and 'social system' sciences. This hierarchy allowed us to ask if items that were 'closer' together at the bottom of the hierarchy, such as physics vs. chemistry, would be harder to distinguish as knowledge clusters than those that were 'further' apart, such as physics vs. psychology. This hierarchy does reflect some Procrustean distortion of the disciplines into a more neatly ordered structure than really exists, but if it captures some degree of real-world structure, it should be reflected in patterns of judgments.

Expert informants who generated the items were asked to list phenomena that could easily be recognized and understood as such by both adults and elementary school children and would not involve any technical terms or exotic relations. Thus, the path of bouncing of a ball would be a better item than the nature of precession in gyroscopes. From a large set of generated items, the experimenters then selected a set that seemed clearest and least ambiguous and most likely to be accessible to children as well as adults. The items were further edited to make sure that various lexical cues to clustering were unlikely to be useful. Thus, if one physical mechanics item asked about the bouncing

[1] We avoided the humanities as it is much less plausible that those domains are organized around a set of core processes or causal relations that generate surface phenomena. Thus, the areas of study of an English department are more likely to be organized around various periods of literature and particular authors or regions and not around mechanisms of irony or production of imagery.

of balls the other physical mechanics item that it might be pitted against would not include a reference to a ball or bouncing, but might instead refer to the speed at which a pendulum swung.

Most of our adult subjects have been college students in North America, a limitation addressed partly by our developmental studies. These adults performed in a manner that was nearly 'error' free, meaning that they would cluster items that were in the same disciplines as more likely to be known in common. For example, if told that one person 'knew all about why a basketball bounces better on the sidewalk than on the grass', they would judge that the same person was much more likely to know 'why a big, heavy boat takes a really long time to stop' than 'why laundry soap cleans kids' dirty clothes'. The basketball and boat cases are both in the domain of mechanics while the soap case is in the domain of chemistry. Because their performance was so high, there was not a strong distance effect in which items further apart in the hierarchy were more easily seen as distinct. Nonetheless, there was a modest effect along these lines. An equally important finding was that many adult subjects were unable to actually explain the phenomena that they clustered together. For example, an adult might judge that a person who knew a great deal about 'why people sometimes fight more when they are tired' would be more likely to know 'why people smile at their friends when they see them' than 'why salt on people's icy driveways makes the ice melt sooner'. In many cases, adults would report that they had no idea of why the phenomenon occurred but were highly confident of their clustering judgment. Similarly, most adults easily judged that a person who knows a great deal about 'why sugar gives us energy to run around and do things' is more likely to know 'why bug spray in the water hurts fish' than 'why policemen can't put people in jail without a reason'—yet many of those same adults were unable to provide even the simplest explanations for those phenomena.

The coupling of a strong confidence in judgments with a frequent inability to explain the basis for such judgments suggested developmental studies. Children might also have a sense of the division of cognitive labor based on discipline-like principles even if they were unable to articulate those principles. Several studies with children ranging in age between 5 and 10 years have now shown that quite young children do have intuitions about the division of cognitive labor that map roughly onto those corresponding to the academic disciplines. There is also evidence that the 'distance' between the disciplines, as represented by

their hierarchical relations, influences performance. Thus, even 5 year olds were at above chance levels on contrasts such as physics vs. cognitive psychology or economics vs. adaptive/ecological biology, approaching almost 70 per cent correct response rates. By contrast 5 year olds were unable to distinguish cognitive from social psychology as in the following example:

> This expert knows all about why some people act like leaders and some people act like followers.
> Do they know more about why people forget things when they get interrupted by the telephone ringing?
> *or*
> Do they know more about why people help each other when they're in trouble?

Nine year olds, on the other hand, immediately saw the contrast and clustered like adults.

Thus, by 5 years of age, children are showing some ability to cluster knowledge in a manner that seems to correspond to the ways in which phenomena are generated by common underlying sets of causal relations. Although the children rarely mentioned such causal relations directly, they do seem to have some implicit sense of broad patterns of causation distinctive to different domains of the natural and social sciences. These might include notions that mechanics is a domain with immediate causal consequences between objects that are monotonically related to the causal force of the first object. By contrast, in the social psychological realm, interactions are often non-monotonic and can occur with considerable delays.

Because the younger children so rarely explained their answers we had to use more indirect methods to assess what causal schemas they might be using. In one follow-up study, we tested the presence of such simple causal schemas by using cases that were technically in a domain such as mechanics but which did not contain a salient causal schema and others that were not in mechanics but had a component that was similar to a causal schema in the domain of mechanics. For example, it appeared that young children saw a coherent domain consisting of bounded objects in motion where consequences were monotonically related to the speed of the initial object mentioned. It was quite easy for them to cluster together these cases. However, when asked about problems of static mechanics, such as bridge structures, the children were less sure

about how to cluster that piece of knowledge. Conversely, when presented with a phenomenon in psychology that involved a salient bounded object in motion ('John knows why you cannot see a bullet moving by you'), some younger children erroneously clustered that knowledge with mechanics.

To what extent could children be solving these problems by simply noting word co-occurrence patterns in roughly paragraph-sized chunks of text? Perhaps children don't need any sense of the causal patterns that exist in the world; they merely need to keep mental tabulations of how often terms such as ' ball', 'bounce', 'fall', and ' hit' co-occur. Then, they cluster phenomena based on their mental tabulations of how much the words in two phenomena have been noted to co-occur in bodies of text. The more powerful co-occurrence methods also tabulate how often words co-occur with an intervening word as a measure of conceptual similarity (Landauer and Dumais 1997). Thus, if 'ball' and 'bounce' co-occur frequently and 'bounce' and 'spring' co-occur frequently, even if 'ball' and 'spring' rarely co-occur, 'ball' and 'spring' will be judged as more similar because of the intervening relationship with 'bounce'. This procedure has been automated and strings of words can be put into programs based on large bodies of text, which then calculate conceptual relatedness.

Such frequency-based cues may help see knowledge clusters of various sorts, but they cannot be the sole basis. In the studies with children just described, the sentences describing the phenomena were fed into a popular frequency-based computer program (Landauer and Dumais 1997). As the sentences were constructed with an eye towards minimizing influences of frequency, it was expected that the program could not cluster the phenomena on discipline-based grounds. Indeed, it was at chance. Even in a study where preschoolers engaged at above chance levels of sorting, word frequency cues were ruled out (Lutz and Keil 2002). Another possible cue might be the clustering together of certain phenomena in instructional curricula. This alternative is more difficult to definitively test, but a look at elementary school curricula in the natural sciences (there is virtually none in the social sciences) suggests that very little information is imparted that would convey the appropriate clusterings.

In short, it appears that, by the age of 5, and possibly even in the later preschool years, when children are asked to cluster bits of causal explanatory knowledge (i.e. knowledge of 'why' for various

phenomena) their judgments appear to be based on inferences about what kinds of causal patterns give rise to those phenomena. They seem to assume that a person who understood one phenomenon well must have done so by virtue of a good grasp of the causal principles that gave rise to that phenomena and therefore is likely to understand other phenomena arising from the same causal principles. On the few occasions where children did attempt to justify their responses, they often talked about the underlying basis for the phenomena and not about the experts or the knowledge itself. For example, one child clustered together two economics items because they both involved 'selling' (even though selling was never explicitly mentioned in the examples). That child said nothing about the experts themselves. Through their intuitions about knowledge clustering, these children are reflecting some of the divisions of knowledge that correspond roughly to natural and social science departments in the modern university. They see these clusters even though most of them have never heard of such departments.

FRAGILITY OF DISCIPLINE-BASED KNOWLEDGE CLUSTERS IN CHILDREN AND A CONTINUING TENSION

Even though young children do cluster knowledge in a manner that accords roughly with some academic disciplines, this ability is fragile when it is faced with competing ways of clustering knowledge. Thus, if a child is presented with a phenomenon that is caused by a certain set of causal relations but also has a salient goal, the goal may well dominate clustering decisions with other phenomena. For example, if a child is told that a person knows all about how marbles bounce off each other in the game of marbles and can use that to win a lot, the child might think the person is more likely to be an expert on another non-mechanics phenomena associated with winning at marbles (e.g. 'why different colored marbles help you keep track of who is winning?') than on a phenomenon that is mechanics but is unrelated to marbles (e.g. 'why yo-yos come back up?'). When goal-based clusters are pitted against discipline-based ones, the goal-based ways of clustering tend to dominate in younger children, with a dramatic shift occurring during the elementary school years such that discipline-based choices start to dominate by the age of 10 years (Danovitch and Keil 2004). When domains such as mechanics and psychology were pitted against salient

goals, the goals won out in almost all 5 year olds and many 7 year olds. Discipline-based ways of clustering knowledge, although available to young children when presented with no competition from goal-based or category association heuristics, are not particularly salient or privileged early on. Instead, goal-based ways of clustering knowledge are more appealing to younger children.

Between roughly the ages of 5 and 10 years, however, a view develops in which underlying causal principles come to be seen as especially powerful ways of understanding the division of cognitive labor. We are currently exploring several mechanisms that might be helping to bring about this shift. One important influence may be the use of higher and higher level category labels with increasing age. We have recently shown that even kindergarteners are more likely to think that an 'animal' expert would have animal knowledge clustered on the basis of biological principles while a 'duck' expert might well be understood as having knowledge organized around goals or category labels (e.g. knowing everything and anything about ducks). The higher the category, the more implausible it is that knowledge would be clustered by goal or topic. For example, when told that a person knew a lot about ducks and asked if she would know more about 'why ducks need to sleep' or about 'why ducks are in a lot of cartoons' children chose roughly equally between these two alternatives. But when told that a person knew a lot about animals, children of all ages made the discipline-based choice ('why ducks need to sleep') by a huge margin. Since children's language reveals an increasing use of higher-order terms with age (Mervis and Crisafi 1982), it may well be that use of such terms helps reveal the special nature of discipline-based clusters.

The tension between discipline-based clusters and other forms remains in adolescence and on into adulthood. If one increases the cognitive load of the knowledge-clustering tasks, people may start to be influenced by topics or goals. For example, if instead of presenting people with triads, they are presented with a large set of say, forty-eight file cards with different phenomena on each and asked to cluster them into like kinds, roughly 35 per cent of adults will cluster them by category labels as opposed to underlying causal discipline (Keil and Rozenblit 1997).

In short, there are clear signs of a sensitivity to causal structure in very young children, a sensitivity that can be used as a way of thinking about the division of cognitive labor. This way of clustering knowledge,

however, is just one of many for young children and seems to be cognitively more challenging than alternatives such as goals and surface topics. During the elementary school years there is a profound shift in which clustering knowledge by underlying causal structure comes to have a privileged status, at least in simple triad tasks. We are currently exploring more fully the basis for this shift and how it relates to other changes in how children understand the nature of knowledge and its distribution in other minds. We are interested in how changes in various patterns of language use might provide clues to the special status of knowledge clustered on the basis of causal principles. In addition, we are interested in whether richer understanding of underlying causal mechanisms in one domain can act as a kind of model that triggers a bias for that way of clustering knowledge in all domains.

FOCUSING THE LENS ON UNDERLYING CAUSAL STRUCTURE

Not all ways of asking about what others know shine an equally bright spotlight on underlying causal structure. Through a series of studies we have been able to show that certain factors highlight discipline-like relations.

The actual form of posing such questions makes quite a difference. For example, the 'why' part of the questions and the division of labor framing may collectively have a strong influence on judgments of clusters. In the tasks described earlier, the framing has usually been of the form:

X knows why P1
What else is X more likely to know?
Why P2?
or
Why P3?

Consider now a triad that strips away both the 'why framing' and the question about expertise and simply presents the phenomena:

P1
Which is more similar to P1?
P2
or
P3?

This second triad would seem to be simpler, and yet in tasks with both adults and children the tendency to cluster on disciplinary grounds drops considerably as other ways of clustering knowledge such as by goals or surface topics become more prominent. There are, of course, many different dimensions of similarity along which phenomena can be compared and when the raw phenomena are presented the discipline-based dimension is not especially salient. One can cluster on surface perceptual similarity of phenomena, on the basis of common lexical items or on the basis of any number of other dimensions. Embedding phenomena in frames that ask about people's 'why knowledge' tends to highlight the underlying causal principles. For example, if adults are presented with the following triad in stripped away form, they may be close to chance levels in clustering either P2 or P3 with P1. By contrast, when the same three phenomena are embedded in a 'X knows all about why' context, there is a strong preference to cluster P3 with P1. Knowing why a phenomenon occurs highlights the core causal processes responsible for that phenomenon in ways that most other contexts do not.

(P1) *A big, heavy boat takes a really long time to stop*
(P2) You can't understand two friends talking at the same time
(P3) You can bounce a basketball better on the street than on grass

Other factors can also enhance a focus on underlying causal processes. There is an advantage in posing the question as one of information-seeking, as in 'You want to know more about why P1: who would be a better person to ask, a person who knows why P2 or a person who knows why P3?' That way of framing the question, which seems to make it more immediately relevant to a participant, shifts children to even higher levels of discipline-based sortings (Danovitch and Keil 2004). As mentioned earlier, posing the question about higher-level categories, such as animals as opposed to ducks, also shifts participants more towards discipline-based clusters.

Thus, asking about the division of cognitive labor with a special focus on why-questions, using more high-level categories, and posing the questions in terms of personal information-seeking, all tend to focus the lens of similarity on the dimension corresponding to underlying causal relations. All these factors enhance performance in children at least as young as 5 years of age. Moreover, young children find it very

natural to make judgments about who knows what based on an initial piece of knowledge. Many facets of meta-cognitive awareness, such as about the limits of one's memory and attentional processes, develop quite late; but a sense that knowledge is clustered into different domains in other minds emerges early and is robust.

TO WHAT END?

Why should young children be so adept at thinking about the division of cognitive labor and why should they show some ability to detect underlying causal relations and use them as a basis for thinking about expertise? Put differently, to what end do they use their sense of the division of cognitive labor? We do not yet know the full answer to this question; but there are some indications of potential uses that help us understand why children are sensitive to the different forms of expertise.

One use may be in evaluating the quality of potential experts. A series of studies in progress is exploring the idea that when children seek out new information, they use their notions of the division of cognitive labor to decide which individuals or sources to approach for new information. A child is told about two self-proclaimed experts. One claims to know a great deal about three phenomena, one from physics, one from economics, and one from psychology, while another claims to know a great deal about three phenomena from physics. Very preliminary evidence suggests that quite young children may know that the first 'expert' is much less plausible than the second. Thus, even young children may have doubts about the likely expertise of a 'Renaissance person'.

A second more direct use of divisions of cognitive labor is to know who to ask for further information or help on a topic. Even preschoolers may seek out different teachers for different problems, even when the problems are novel and don't simply match old ones that certain teachers have solved on prior occasions. When faced with several different adults to approach for information or for a problem solution, it can be very helpful to consider what proven areas of knowledge each of those adults already have. As we have seen, younger children might use different and sometimes misleading heuristics for seeking out the best experts, but in many cases they will do far better than chance. In a similar vein, when children hear bits of conflicting information from

different adults they may use their sense of the legitimate division of cognitive labor to weigh the quality of the information that they hear. Thus, if a series of statements from one individual does not cohere as a natural domain of knowledge, a particular fact in that series may be discounted more than the same fact embedded in a series that is more coherent.

There may be a more important and subtler use, however, that is seen in groups at all ages. A sense of the division of cognitive labor provides confidence about one's current knowledge. The vast causal complexity of the natural and artificial worlds makes it impossible for any one person to have much more than the shallowest grasp of causal structure in a domain (Wilson and Keil 1998). Although there is evidence that people delude themselves in thinking that they understand such causal relations in far more detail than they really do (Rozenblit and Keil 2002), they are nonetheless also aware of at least some of the gaps in their knowledge.[2] A grasp of the division of cognitive labor enables them to feel that their knowledge is well grounded to the extent that there are legitimate experts who, collectively, could provide additional supporting information that could fill in the gaps. This form of support is closely related to how we might rely on the division of linguistic labor. If I believe that a panda bear is a particular kind of bear and label it as such, I may have considerable confidence about that belief because I have heard biologists state that DNA analyses show a clear pattern of commonality with other bears as opposed to other species.

My confidence arises from my sense of how knowledge in the science is distributed, a sense of the modern discipline of biology, and of the central role of microstructural properties such as DNA to understanding species. This idea of experts in biology is not restricted to those who encountered such concepts late in high school or college. It is accessible in a rough manner to surprisingly young children. Across a wide range

[2] There is a tendency to grossly overestimate one's causal explanatory understanding of both devices and natural phenomena. Whether it is everyday objects as simple as a zipper or a flush toilet or more complex objects such as a helicopter, adults and children alike think they have far more detailed understandings of the mechanism than they really do. People's initial ratings of what they know drop sharply after they are asked to actually provide explanations. This 'illusion of explanatory depth' is specific to estimates of how well one understands how things work. In contrast, people tend to be quite well calibrated in their estimates of how well they know facts, procedures, or narratives (Rozenblit and Keil 2002).

of ages, it may guide the strength of our beliefs and the extent to which we are willing to revise those beliefs and be persuaded by others.

Cross-cultural investigations of people's notions of the division of cognitive labor are just beginning and will be an important way of examining the extent to which the causal structure of the world drives intuitions about who knows what. The developmental studies suggest that there may be a striking universality of intuitions about clustering of why knowledge of everyday phenomena. Thus, even in traditional societies that have never had any exposure to the Western sciences, there may be a shared sensitivity to clusters of causal patterns that are used to infer clusters of knowledge. The causal patterns are relatively invariant across cultures; and if they are an important source of information for intuitions about expertise, they should cause a convergence on beliefs about relevant experts. Clustering of knowledge on the basis of category association, access, and goals, however, may show far more cultural variation. All three of those factors can be heavily influenced by culture and language. Discipline-based ways of thinking about expertise may therefore be the most robust and constant across cultures. This prediction poses a challenge to views that the domains of inquiry of the natural and social sciences are largely socially constructed.

In short, in all cultures, we come to depend on the knowledge of others. The division of cognitive labor is an essential infrastructure that allows us to transcend the very limited understandings that exist in the mind of any one individual. To benefit from the division of cognitive labor, however, we need ways of thinking about domains of expertise that can be used to tap into that expertise when needed. There are several distinct heuristics that can be used to figure out who knows what. Although there are major developmental changes in which heuristics are preferred, very young children are sensitive to many of these heuristics, including one that refers to the underlying causal patterns responsible for large classes of phenomena. At all ages, these heuristics provide a rudimentary sense of domains of expertise that can be used to evaluate the quality of new information. Thus, an important basis for doubt lies in our patterns of deference to others, patterns that heavily influence our deliberations throughout much of our development.[3]

[3] Much of the research reported on in this article was supported by NIH Grant R37-HD23922 to Frank Keil. Many thanks to Tamar Gendler for extensive comments on earlier drafts of this paper.

REFERENCES

Boyd, Richard (1999) 'Homeostasis, Species, and Higher Taxa', in R. Wilson (ed.), *Species: New Interdisciplinary Studies* (Cambridge), 141–85.

Burge, Tyler (1979) 'Individualism and the Mental', in Peter French (ed.), *Midwest Studies in Philosophy*, iv. *Studies in Metaphysics* (Minneapolis), 73–122.

Cartwright, Nancy (1999) *The Dappled World* (Cambridge).

—— (2003) 'Causation: One Word; Many Things', *Causality: Metaphysics and Methods Technical Report*. Centre for the Philosophy of Natural and Social Science, London School of Economics (London).

Danovitch, Judith, and Frank Keil (2004) 'Should you ask a Fisherman or a Biologist? Developmental Shifts in Ways of Clustering Knowledge', *Child Development*, 75: 918–31.

Dupre, John (1981) 'Natural Kinds and Biological Taxa', *Philosophical Review*, 90: 66–90.

Durkheim, Emile (1947) *The Division of Labor in Society*, tr. George Simpson (New York).

Fodor, Jerry (1974) 'Special Sciences', *Synthese*, 28: 97–115

—— (1998) *Concepts: Where Cognitive Science went Wrong* (New York).

Goldman, Alvin (1999) *Knowledge in a Social World* (Oxford).

—— (2001) 'Social Epistemology', *Stanford Encyclopedia of Philosophy* (http://plato.stanford.edu/entries/ epistemology-social/).

Green, Donald, Bradley Palmquist, and Eric Schickler (2002) *Partisan Hearts and Minds: Political Parties and the Social Identities of Voters* (New Haven).

Hacking, Ian (1999) *The Social Construction of What?* (Cambridge).

Hume, David (1739) *A Treatise of Human Nature* (Oxford).

Hutchins, Edwin (1995) *Cognition in the Wild* (Cambridge).

Keil, Frank (2003*a*) 'Folkscience: Coarse Interpretations of a Complex Reality', *Trends in Cognitive Sciences*, 7: 368–73.

—— (2003*b*) 'That's Life: Coming to Understand Biology', *Human Development*, 46: 369–77.

——and Leonid Rozenblit (1997) 'Knowing Who Knows What', presented at the 1997 Meeting of the Psychonomics Society, Philadelphia.

Kitcher, Philip (1990) 'The Division of Cognitive Labor', *Journal of Philosophy*, 87: 5–22.

—— (2001) *Science, Truth and Democracy* (New York).

Kukla, Andre (2000) *Social Construction and the Philosophy of Science* (London).

Landauer, Thomas, and Susan Dumais (1997) 'A Solution to Plato's Problem: The Latent Semantic Analysis Theory of Acquisition, Induction, and Representation of Knowledge', *Psychological Review*, 104: 211–40.

Lutz, Donna, and Frank Keil (2002) 'Early Understanding of the Division of Cognitive Labor', *Child Development*, 73: 1073–84.
Mervis, C. B., and M. A. Crisafi (1982) 'Order of Acquisition of Subordinate, Basic, and Superordinate Categories', *Child Development*, 53: 258–66.
Murphy, Gregory (2003) *The Big Book of Concepts* (Cambridge).
Niklas, Karl (1996) 'How to Build a Tree', *Natural History*, 105: 48–52.
Prinz, Jesse (2002) *Furnishing the Mind: Concepts and their Perceptual Basis* (Cambridge).
Putnam, Hilary (1975) 'The Meaning of ''Meaning'' ', in Keith Gunderson (ed.), *Language, Mind, and Knowledge*, ii (Minneapolis), 131–93.
Rosch, E., C. Mervis, W. Gray, D. Johnson, and P. Boyes-Braem (1976) 'Basic Objects in Natural Categories', *Cognitive Psychology*, 8: 382–439.
Rozenblit, Leonid, and Frank Keil (2002) 'The Misunderstood Limits of Folk Science: an Illusion of Explanatory Depth', *Cognitive Science*, 26: 521–62.
Smith, Adam (1976) *An Inquiry into the Nature and Causes of the Wealth of Nations*, ed. R. H. Campbell and A. S. Skinner (Oxford).
Wilson, Robert, and Frank Keil (1998) 'The Shadows and Shallows of Explanation', *Minds and Machines*, 8: 137–59.

7. The Epistemic Significance of Disagreement

Thomas Kelly

> Looking back on it, it seems almost incredible that so many equally educated, equally sincere compatriots and contemporaries, all drawing from the same limited stock of evidence, should have reached so many totally different conclusions—and always with complete certainty.
>
> (John Michell, *Who Wrote Shakespeare?*)

1. INTRODUCTION

Consider the following issues, each of which is the object of considerable controversy:

(1) the extent to which a desire to intimidate the Soviet Union played a role in Harry Truman's decision to drop the atomic bomb on Japan in 1945
(2) whether Truman's decision to do so was morally justified
(3) whether there are in fact any truths of the kind that Immanuel Kant called ''synthetic a priori''

I have a belief about each of these issues, a belief that I hold with some degree of conviction. Moreover, I ordinarily take my beliefs about each of these matters to be rational—I think of myself as having good reasons for holding them, if pressed to defend my position I would

For helpful discussion and correspondence, I am grateful to Adam Elga, David Chalmers, David Christensen, Richard Feldman, Anil Gupta, Peter van Inwagen, Derek Parfit, Jim Pryor, Pamela Hieronymi, Michael Rescorla, Kerian Setiya, Jonathan Vogel, Ralph Wedgwood, and Roger White. Earlier versions of this paper were read at the University of Notre Dame, Harvard University, the University of Pittsburgh and at a Pacific Division meeting of the American Philosophical Association; I am grateful to the audiences present on those occasions.

cite those reasons, and so on.[1] On the other hand, I am very much aware of the fact that, with respect to each issue, there are many others who not only do not share my belief, but in fact, take a diametrically opposed position. Of course, the mere fact of disagreement need not be problematic: if, for example, I was convinced that all of those who disagreed with me were simply being foolish, or hadn't bothered to think about the matter carefully enough, or were unfamiliar with evidence that I happen to possess (evidence which, if presented to them, would result in a change in their views), then I might simply shrug off this disagreement. But in fact, I believe no such thing: I acknowledge that on many controversial issues with respect to which I have a firmly held belief, there are some who disagree with me whose judgement cannot be simply discounted by appeal to considerations of intelligence, thoughtfulness, or ignorance of the relevant data.

Can one rationally hold a belief while knowing that that belief is not shared (and indeed, is explicitly rejected) by individuals over whom one possesses no discernible epistemic advantage? If so, what assumptions must one be making about oneself and about those with whom one disagrees? In deciding what to believe about some question, how (if at all) should one take into account the considered views of one's **epistemic peers**?[2]

My aim in this paper is to explore the *epistemic significance of disagreement*. A central concern is whether the practice of retaining beliefs that are rejected by individuals over whom one claims no epistemic advantage is a defensible one. It is, of course, far from clear that the relevant practice *is* defensible. For it is natural to suppose that

[1] Of course, not all rational beliefs are rationalized by supporting reasons: my belief that $2 + 2 = 4$ is (I assume) a rational belief, but it is not rationalized in virtue of standing in a certain relation to supporting considerations, in the way that my rational belief that *communist economies tend to be inefficient* is. In this paper, however, I will ignore the case of beliefs whose rationality consists in their status as 'properly basic' (to borrow a phrase from Alvin Plantinga). Indeed, I suspect that beliefs of this kind would require a very different treatment than the one offered here.

[2] I owe the term 'epistemic peer' to Gutting (1982). Gutting uses the term to refer to those who are alike with respect to 'intelligence, perspicacity, honesty, thoroughness, and other relevant epistemic virtues' (p. 83). I will use the term in a somewhat extended sense. As I will use the term, the class of epistemic peers with respect to a given question are equals, not only with respect to their possession of the sort of general epistemic virtues enumerated by Gutting, but also with respect to their exposure to evidence and arguments which bear on the question at issue. I discuss this notion further in § 2.3 below.

persistent disagreement among epistemic peers should undermine the confidence of each of the parties in his or her own view. This natural intuition was voiced by Henry Sidgwick in a memorable passage in *The Methods of Ethics*:

> the denial by another of a proposition that I have affirmed has a tendency to impair my confidence in its validity . . . And it will be easily seen that the absence of such disagreement must remain an indispensable negative condition of the certainty of our beliefs. For if I find any of my judgements, intuitive or inferential, in direct conflict with a judgement of some other mind, there must be error somewhere: and if I have no more reason to suspect error in the other mind than in my own, reflective comparison between the two judgements necessarily reduces me . . . to a state of neutrality. (p. 342)

Sidgwick's idea—that reflection on the relevant sort of disagreement should reduce one to 'a state of neutrality'—has been endorsed by thinkers both early and late. The idea played a prominent role in ancient skepticism as one of the 'modes of Pyrrhonism' designed to rationally induce suspension of judgement. Here is the characterization offered by Sextus Empiricus:

> According to the mode deriving from dispute, we find that undecidable dissension about the matter proposed has come about both in ordinary life and among the philosophers. Because of this we are not able either to choose or to rule out anything, and we are driven to suspend judgement. (I. 165)[3]

Indeed, in his own presentation of the case for skepticism, Sextus seems to indicate that the existence of such disagreement is ultimately the most fundamental consideration of all.[4] The same idea is a recurrent theme in Montaigne's case for skepticism as presented in his *Essays*. More recently, Keith Lehrer (1976) has claimed that there is simply no room for rational disagreement among those who share the same information and have even a minimal level of respect for each other as judges: in such circumstances, each party to the dispute should revise his or her own judgement until consensus is achieved. In economics, a substantial body of literature similarly seems to suggest that the

[3] As reported by Sextus, the argument from disagreement was one of the Ten Modes of Aenesidemus as well of one of the Five Modes of Agrippa; it thus played a part in both early and late Pyrrhonhism.

[4] See I. 178–9, where the standard modes seem to ultimately depend upon the existence of 'interminable controversy among the philosophers'.

uniquely rational response to known disagreement is to revise one's original beliefs so as to bring about consensus.[5]

Despite its attractiveness, this line of thought is, I believe, mistaken. Disagreement does not provide a good reason for skepticism or to change one's original view. In what follows, I will argue for the following thesis: once I have thoroughly scrutinized the available evidence and arguments that bear on some question, the mere fact that an epistemic peer strongly disagrees with me about how that question should be answered does not itself tend to undermine the rationality of my continuing to believe as I do. Even if I confidently retain my original view in the face of such disagreement, my doing so need not constitute a failure of rationality. Indeed, confidently retaining my original belief might very well be the *uniquely* reasonable response in such circumstances.

According to the view that I will defend then, disagreement does not have the kind of significance that has often been claimed for it. However, it would be a mistake to conclude that disagreement is therefore without epistemic significance. I will thus also attempt to clarify the nature of the significance that disagreement does have in those cases in which it *is* of significance.

The discussion which follows is as much exploratory as it is polemical. A primary concern is to make fully explicit the substantive commitments and assumptions about rationality of one who defends the views that I defend. I do not pretend that the relevant commitments are costless. I myself do not find the costs unacceptably high. But this, of course, is itself something about which others might very well disagree.

2. SOME PRELIMINARY DISTINCTIONS

2.1.

I begin by locating the question that I want to pursue relative to certain other, closely related questions. Here, the most straightforward distinction to be drawn is that between descriptive and normative questions. There is a considerable amount of empirical evidence which suggests that an awareness of disagreement tends to lead us to significantly

[5] Here I have in mind the tradition of research which dates from Aumann's seminal paper 'Agreeing to Disagree' (1976). I discuss the import of this literature in § 3 below.

moderate our opinions. That is, within isolated groups, there are strong psychological pressures that tend to lead to the formation of consensus, or at least, to the formation of a dissensus that is less polarized than the one which would otherwise have obtained. Questions about the pervasiveness and scope of such phenomena have been fruitfully explored by social psychologists.[6]

In contrast to descriptive questions about how an awareness of disagreement in fact affects our beliefs, the question that I want to pursue belongs to the class of normative questions—questions about how an awareness of disagreement *should* affect our beliefs. Answering these normative questions could, in principle, lead us to revise our actual practice, to alter our characteristic responses to disagreement. Alternatively, if it is beyond our power to revise our actual practice—say, because our actual responses to disagreement are psychologically fixed[7]—how we answer these normative questions might affect our attitudes towards our unalterable responses. Thus, suppose that, as a matter of fact, an awareness of disagreement tends to more or less inevitably lead us to revise our views in the direction of greater consensus. If we conclude that it is epistemically appropriate to give a great deal of weight to the judgements of others in revising our own beliefs, then we might view this unavoidable psychological tendency with relative equanimity, or even with pride, as symptomatic of our natural and reflexive rationality. If, on the other hand, we conclude that doing so is not the epistemically appropriate response, then we might view our inevitable tendency to respond in this way in a less favorable light: perhaps as symptomatic of a somewhat craven desire to adhere to orthodoxy for orthodoxy's sake.

2.2.

As I have emphasized, it is at least somewhat natural to suppose that when one discovers that others explicitly reject some view that one holds, this discovery ought to make one more skeptical of that view. It is important, however, to distinguish carefully between two quite

[6] The classic studies in this tradition were conducted by Solomon Asch (1952, 1956).

[7] A prominent theme in recent epistemology is that much of the epistemological tradition seems to presuppose that we possess a degree of control over our beliefs that we do not in fact possess. See e.g. Alston (1988) and Plantinga (1993: esp. ch. 1).

distinct kinds of skepticism that such a discovery might be thought to warrant. The first kind of skepticism is skepticism about whether there is, after all, a fact of the matter about the disputed question. That is, it might be thought that persistent disagreement with respect to a given domain warrants some kind of non-factualism or error theory about that domain. Thus, in moral philosophy the existence of disagreement with respect to fundamental ethical questions is often claimed to strengthen the case for non-factualism or some variety of error theory on the grounds that there being no fact of the matter is the best explanation of our inability to reach agreement.[8] Similarly, the phenomenon of persistent disagreement among theorists concerning the correct solution to various decision problems is sometimes thought to bolster the case for expressivist accounts of discourse about practical rationality. Although in contemporary philosophy this move is most often made with respect to normative domains, it has in the past often been made with respect to non-normative domains as well. Thus, the logical positivists frequently insisted that the seemingly interminable controversies among theologians and metaphysicians are due to the fact that the relevant bodies of discourse are not truth-apt but rather 'cognitively meaningless'. Here again, the driving idea is that the best explanation of why we cannot agree about what the facts are is simply that there are no facts upon which we might agree.

Questions about the circumstances in which disagreement warrants some variety of non-factualism or error theory about a given domain are interesting ones, but they will not be pursued here. Instead, I want to examine cases in which we are confident that there *is* a genuine fact of the matter—*despite* the existence of disagreement—in order to inquire as to how an awareness of that disagreement should affect our beliefs in such cases. I assume that there are some domains with respect to which we occupy this position. Consider, for example, history. There is, I assume, a fact of the matter about whether a desire to intimidate the Soviet Union played a role in Harry Truman's decision to drop the atomic bomb—however much knowledgeable and highly qualified historians might disagree about what that fact of the matter is. Of course, certain postmodernists and anti-realists about the past might question this. But here it is fair to say, I think, that our commitment to a robust

[8] Mackie (1977) is a classic attempt to motivate an error theory by appeal to facts about ethical disagreement.

factualism about historical discourse is stronger than any argument that such thinkers have yet provided.

Compared to questions about whether disagreement should undermine our commitment to factualism about various domains, questions about the extent to which disagreement poses a distinctively *epistemic* challenge have been relatively underexplored. In fact, much of what little discussion this question has received has taken place within the philosophy of religion: philosophers of religion have debated the extent to which an awareness of the great diversity of (sometimes) incompatible religious traditions ought to make a theist more skeptical about the distinctive claims of her own tradition.[9] It is unclear, however, whether there is any *special* problem about religious claims in particular. For, as Peter van Inwagen (1996) has emphasized, everyone, or almost everyone, would seem to be in the position of the theist with respect to at least some questions. That is, virtually everyone has at least some beliefs that are explicitly rejected by individuals over whom he or she possesses no discernible epistemic advantage. This phenomenon, while no doubt familiar enough from everyday life, is perhaps especially salient for philosophers. For philosophy is notable for the extent to which disagreements with respect to even the most basic questions persist among its most able practitioners, despite the fact that the arguments thought relevant to the disputed questions are typically well-known to all parties to the dispute. (It is not, after all, as though Compatibilists about free will think themselves privy to some secret master argument, such that if this argument were presented to the Incompatibilists, the Incompatibilists would see fit to abandon their view.)

2.3.

It is uncontroversial that there are some circumstances in which one should give considerable weight to the judgements of another party in deciding what to believe about a given question. Paradigmatic examples consist of cases in which it is clear that the other party enjoys some epistemic advantage with respect to the question at issue. The list of possible advantages which one party might enjoy over another seems to divide naturally into two general classes. First, there are advantages that

[9] See, for example, Gutting (1982), Plantinga (2000), and the essays collected in Quinn and Meeker (2000).

involve a superior familiarity with or exposure to evidence and arguments that bear on the question at issue. Thus, suppose that I know that you possess not only all of the evidence which I possess but also some relevant evidence which I lack. (That is, my total evidence is a proper subset of your total evidence.) In these circumstances, it makes sense for me to treat your beliefs as indicators of the actual state of the evidence since I have no independent access to the character of that evidence. More subtly: it might be that although we have both been exposed to the same body of evidence, you have carefully scrutinized that evidence while I have considered it only hastily or in a cursory manner. Here again, it is your superior familiarity with the evidence which makes a certain measure of deference on my part the appropriate course.

A second class of epistemic advantages which one might potentially enjoy consists in superiority with respect to general epistemic virtues such as intelligence, thoughtfulness, freedom from bias, and so on. Thus, if I know that I have great difficulty being objective when it comes to assessing the quality of my work but that you labor under no such handicap, then I have a reason to defer to your judgements about my work, all else being equal.[10]

Any plausible view, I take it, will allow for the fact that I should give considerable weight to your judgements when I have reason to believe that your epistemic position is superior to my own in either of these ways (at least, provided that I do not claim some compensating advantage). Because some measure of deference seems clearly appropriate in such circumstances, the question that I want to pursue concerns the normative significance of disagreement in cases in which *neither* of the parties enjoys such an advantage.

Let us say that two individuals are **epistemic peers** with respect to some question if and only if they satisfy the following two conditions:

[10] As this last example makes clear, it is no doubt overly simple to attribute to an individual some particular level of (e.g.) objectivity or thoughtfulness irrespective of a particular subject matter: the extent to which an individual possesses such qualities might very well (and in the usual case, will) vary significantly from domain to domain. Attributions of a given level of objectivity or thoughtfulness should thus be relativized to particular domains. (It is an empirical question, I take it, how the relevant domains should be demarcated.) In what follows, the need for such relativization should be taken as understood; for expository purposes, I will avoid repeated mentions of this need, and write simply of an individual's objectivity (etc.) rather than her objectivity-with-respect-to-domain-A.

(i) they are equals with respect to their familiarity with the evidence and arguments which bear on that question, and
(ii) they are equals with respect to general epistemic virtues such as intelligence, thoughtfulness, and freedom from bias.[11]

The question at issue, then, is whether known disagreement with those who are one's epistemic peers in this sense must inevitably undermine the rationality of maintaining one's own views.

3. NO AGREEING TO DISAGREE?

Why might one think that it is unreasonable to steadfastly maintain one's views in the face of such disagreement? In economics, there is a substantial body of literature which purports to show the irrationality of 'agreeing to disagree' in various circumstances. The first to develop general results along these lines was Robert Aumann (1976). In a classic paper, Aumann showed that if two or more individuals (i) update their beliefs by Bayesian conditionalization, (ii) have common prior probabilities, and (iii) have common knowledge of each other's opinions, then (iv) those individuals will not knowingly disagree on the answer to any question: rather, they will continuously revise their beliefs until consensus is reached. Subsequent work has shown that Aumann's 'no

[11] It is a familiar fact that, outside of a purely mathematical context, the standards which must be met in order for two things to count as *equal* along some dimension are highly context-sensitive. Thus, inasmuch as classes of epistemic peers with respect to a given issue consist of individuals who are 'epistemic equals' with respect to that issue, whether two individuals count as epistemic peers will depend on how liberal the standards for epistemic peerhood are within a given context. That is, whether two individuals count as epistemic peers will depend on *how much* of a difference something must be in order to count as a *genuine* difference, according to the operative standards. In the same way, whether two individuals count as 'the same height' will depend on the precision of the standards of measurement that are in play. (Lewis 1979 is a classic discussion of the relevant kind of context-sensitivity.) Of course, given sufficiently demanding standards for epistemic peerhood, it might be that no two individuals *ever* qualify as epistemic peers with respect to any question. (Perhaps there is always at least some *slight* difference in intelligence, or thoughtfulness, or familiarity with a relevant argument.) Similarly, it might be that no two individuals count as the same height given sufficiently demanding standards of equality. My sense is that, often enough, the standards that we employ in assessing intelligence or thoughtfulness (like the standards that we employ when measuring height) are sufficiently liberal to allow individuals to qualify as equal along the relevant dimensions.

agreeing to disagree' result survives various weakenings of his original assumptions.[12]

Contrary to what one might naturally assume, however, this tradition of research does not in fact support the conclusion that known disagreement among epistemic peers provides each of the peers with a good reason to revise his or her view. Indeed, close examination reveals that the technical results which have been established thus far do not bear on the case of disagreement among epistemic peers at all. As noted, Aumann's original proof depends on the assumption of common prior probabilities. This assumption is tantamount to assuming that there is a prior agreement as to the normative import of any piece of evidence which might be encountered. In effect, Aumann's 'no agreeing to disagree' result holds only for individuals *who would hold identical views given the same evidence*. And although subsequent work in this tradition has shown that Aumann's result can survive certain weakenings in his original assumptions, the assumption of common prior probabilities has not proven dispensable. Now, by definition, individuals who are epistemic peers with respect to a given question have been exposed to the same evidence which bears on that question. Disagreement among epistemic peers then, is disagreement among those who disagree *despite* having been exposed to the same evidence. Thus, our question concerns a case which stands outside the range of cases for which Aumann's result holds.

The guiding idea behind the 'no agreeing to disagree' literature is that, in many circumstances, the discovery that another person holds a view that one is inclined to reject constitutes evidence that the other person has access to relevant evidence which one does not possess. By giving some weight to the view of the other person, one is able to take into account the import of that evidence to which one would otherwise lack access. Thus, one does not have to posses the evidence for oneself in order to take its epistemic import into account. This guiding idea represents a genuine insight.[13] Indeed, as emphasized above (§ 2.3), *any* plausible epistemological view will allow for the fact that I should give considerable weight to your beliefs when I have reason to think

[12] Geanakoplos (1994) provides a basic exposition of Aumann-like results through 1994.

[13] For further exploration of this theme, as well as an attempt to specify normative principles which should guide our attempts to take account of evidence that we do not possess, see Kelly (forthcoming *b*).

that the fact that you believe as you do is attributable to your possession of some relevant evidence which I do not possess. However, our present question is not how one should respond to the beliefs of others when one lacks access to the evidence on which those beliefs are based. The question, rather, is how one should respond when one *does* have access to the relevant evidence.

The technical results of the 'no agreeing to disagree' literature then, do not bear directly on the question at issue. It might be thought, however, that one who appreciates the guiding idea which underlies these technical results (viz. that one takes into account evidence which one does not possess by taking into account the views of those with whom one disagrees) will naturally embrace the view that disagreement provides a good reason for skepticism. For suppose that I know that you are significantly better informed than I am with respect to some question. In these circumstances, it makes sense for me to defer to your better informed judgement in deciding what to believe about that question. In reasoning in this way, I presuppose that you are a competent evaluator of that evidence which is available to you but not to me. (If I knew that you were *not* a competent evaluator of this evidence, then it would be illegitimate for me to draw inferences about the character of your evidence from the content of your beliefs.) Suppose that at some later time, our epistemic positions are equalized: I gain access to that evidence which was previously available only to you. I am now in a position to make my own judgement about the probative force of the evidence. Still, it might be thought that consistency requires that I continue to give considerable weight to your judgement about what our (now common) total evidence supports. After all, even if I'm strongly inclined to disagree with you as to the overall import of the evidence in a given case, shouldn't I give considerable weight to your judgement given my readiness to defer to you when I am otherwise ignorant of that evidence? Indeed, unless I have some positive reason to think that one of us is more likely to do a better job with respect to assessing the relevant evidence than the other, shouldn't I give *equal* weight to our considered judgements? Recall Sidgwick's remark: 'if I find any of my judgements, intuitive or inferential, in direct conflict with a judgement of some other mind, there must be error somewhere: and if I have no more reason to suspect error in the other mind than in my own, reflective comparison between the two judgements necessarily reduces me . . . to a state of neutrality'.

Ultimately, the argument that it is unreasonable to maintain one's views in the face of such disagreement depends on considerations of symmetry. According to this line of thought, the only thing that would justify one in maintaining views that are rejected by one's epistemic peers would be if one had some positive reason to privilege one's own views over the views of those with whom one disagrees. But *ex hypothesi*, no such reason is available in such cases. In the next section, I take up this argument and attempt to show how it might be resisted.

4. THE APPEAL TO SYMMETRY

Suppose that two epistemic peers—let's call them 'you' and 'I'—are each deliberating about what attitude to take towards a given hypothesis H in the light of the available evidence. Suppose further that, as a result of my assessment of the evidence, I come to believe H, while as a result of your assessment of the evidence, you come to believe not-H. If we subsequently become aware of our disagreement, how if at all should we revise our respective views? Again, in these circumstances, considerations of symmetry would seem to dictate that suspension of judgement is the uniquely reasonable response on both of our parts: that is, each of us should abandon his or her prior conviction and retreat towards an attitude of agnosticism with respect to H. For how could either of us defend doing otherwise? Consider how the situation appears from my perspective. *Ex hypothesi*, I admit that there are no objective criteria that make it antecedently more probable that I am more likely than you are to be correct on this particular occasion—I do not claim to be any smarter, a better reasoner, or to possess some relevant evidence which you lack. Given the acknowledged, perfect symmetry of our positions, how can I possibly justify *not* giving equal weight to your considered judgement? After all, wouldn't this be the most rational course for some objective, on-looking third party who knew nothing about our dispute other than the fact that it is two judges of equal competence and qualifications who disagree? Given this, wouldn't my failure to give equal weight to your judgement amount to a kind of epistemic arbitrariness on my part, an indefensible privileging of my own position for no other reason than the fact that it is my own?

However, the claim that things are perfectly symmetrical in such cases deserves further scrutiny. Indeed, to uncritically assume that

things are perfectly symmetrical with respect to all of the epistemically relevant considerations in such cases is, I think, to subtly beg the question in favor of the skeptical view. For consider: I am no smarter than you are, no better at reasoning, no better informed, and (hence) no more fit to pronounce upon the issue at hand. So far, it is uncontroversial that things are perfectly symmetrical between us. Then a body of evidence is introduced, and we are asked to make a judgement about how strongly that body of evidence confirms or disconfirms a certain hypothesis. Suppose that, as it turns out, you and I disagree. From my perspective, of course, this means that you have misjudged the probative force of the evidence. The question then is this: why shouldn't I take *this* difference between us as a relevant difference, one which effectively breaks the otherwise perfect symmetry?

After all, the question of how well someone has evaluated the evidence with respect to a given question is certainly the kind of consideration that is relevant to deciding whether his or her judgement ought to be credited with respect to that question. That is, it is exactly the sort of consideration that is capable of producing the kind of asymmetry that would justify privileging one of the two parties to the dispute over the other party. And from my vantage point—as one of the parties *within* the dispute, as opposed to some on-looking third party—it is just this undeniably relevant difference that divides us on this particular occasion.[14]

One might wonder: is my assessment that you have misjudged the probative force of the evidence consistent with my continuing to regard you as a genuine epistemic peer? Yes, it is. Of course, if I came to believe that I am, in general, a better evaluator of evidence than you are, then this would be a good reason for me to demote you from the ranks of those to whom I accord the status of epistemic peer. But a revision in my assessment of our relative levels of competence is in no way mandated by the judgement that one of us has proven superior with respect to the exercise of that competence on a given occasion. Two chess players of equal skill do not always play to a draw; sometimes one or the other wins, perhaps even decisively.

[14] In cautioning against the tendency to think that the correct way to view such disputes is from a purely external, third-person point of view, I echo Richard Foley. As he puts the point: 'It is deeply misleading to think about such conflicts in terms of a model of neutral arbitration between conflicting parties' (2001: 79). Cf. Foley (1994: 65–6) and Plantinga (1995: 182).

At the outset of the paper, I asked what I must be assuming about myself and about others who have been exposed to the same evidence when I continue to hold a belief that they reject. My answer to this question is: perhaps not very much. In particular, I need not assume that I was better qualified to pass judgement on the question than they were, or that they are likely to make similar mistakes in the future, or even more likely to make such mistakes than I am. All I need to assume is that *on this particular occasion* I have done a better job with respect to weighing the evidence and competing considerations than they have.

Of course, there is still the question of whether I am *correct* in thinking that I have done a better job with respect to evaluating the evidence and arguments than those with whom I find myself in disagreement. Suppose that they reason in a parallel way and conclude that *I'm* the one who has misjudged the evidence. On the present view, the rationality of the parties engaged in such a dispute will typically depend on who has *in fact* correctly evaluated the available evidence and who has not. If you and I have access to the same body of evidence but draw different conclusions, which one of us is being more reasonable (if either) will typically depend on which of the different conclusions (if either) is in fact better supported by that body of evidence. No doubt, especially in the kinds of cases at issue, it will often be a non-trivial, substantive intellectual task to determine what the totality of relevant evidence supports. Therefore, the question of which one of us is doing a better job evaluating the evidence will often be a non-trivial, substantive intellectual question. But here as elsewhere, life is difficult. On any plausible conception of evidence, we will be extremely fallible with respect to questions about what our evidence supports.[15] The amount of disagreement that we find among well-informed individuals simply makes this fact more salient than would otherwise be the case.

On the present view, the rationality of one's believing as one does is not threatened by the fact that there are those who believe otherwise. Rather, any threat to the rationality of one's believing as one does depends on whether those who believe otherwise have good reasons for believing as they do—reasons that one has failed to accurately

[15] Indeed, if Williamson (2000) is correct, then our ability to fully appreciate our evidence is subject to *in principle* limitations. However, even if one finds Williamson unconvincing on this point, one should still admit that we are in fact extremely fallible when it comes to evaluating large and diverse bodies of evidence. I discuss this fallibility further in § 6 below.

appreciate in arriving at one's own view. I explore this theme further in the following section.

5. RATIONALITY AND MERELY POSSIBLE DISAGREEMENT

Consider the circumstances in which we are apt to find disagreement intellectually threatening. Here, the first point to appreciate is the following: it is extremely implausible that *actual* disagreement is always more epistemically significant than certain kinds of *merely possible* disagreement. After all, whether there is any actual disagreement with respect to some question as opposed to merely possible disagreement might, in a particular case, be an extremely contingent and fragile matter. In particular, whether there is any actual disagreement might very well depend on factors that everyone will immediately recognize as irrelevant to the truth of the question at issue. (Suppose, for example, that there would be considerable disagreement with respect to some issue, but that all of the would-be dissenters have been put to death by an evil and intolerant tyrant.)

The existence of actual disagreement, then, need be no more intellectually threatening than certain kinds of merely possible disagreement. However, not *every* kind of merely possible disagreement will be intellectually threatening: the possibility that individuals who are insane or who are otherwise clearly irrational might disagree with some view that we hold would presumably *not* provide us with a good reason to doubt that view. The question, then, is this: under what circumstances should we find the possibility of disagreement intellectually threatening? Whether we find the possibility of disagreement intellectually threatening, I suggest, will and should ultimately depend on our considered judgements about *how rational* the merely possible dissenters might be in so dissenting. And our assessment of whether rational dissent is possible with respect to some question (or our assessment of the extent to which such dissent might be rational) will depend in turn on our assessment of the strength of the evidence and arguments that might be put forward on behalf of such dissent. But if this is correct, then the extent to which merely possible dissent should be seen as intellectually threatening effectively reduces to questions about the strength of the reasons that might be put forward on behalf of such

dissent. Now, there might be cases in which we judge that the arguments and evidence that could be brought forth on behalf of a hypothetical dissent are truly formidable, and this might justifiably make us doubt our own beliefs. But in that case, the reasons that we have for skepticism are provided by the state of the evidence itself, and our own judgements about the probative force of that evidence. The role of disagreement, whether possible or actual, ultimately proves superfluous or inessential with respect to the case for such skepticism.

Suppose that those members of the philosophical community who have both (i) thought seriously about Newcomb's Problem and (ii) are familiar with the main arguments on both sides are approximately evenly divided between One-Boxers and Two-Boxers.[16] We can imagine various ways in which this state of disagreement gives way to a consensus. Here is one way: someone thinks of an ingenious argument that convinces all of the One-Boxers (or, alternatively, all of the Two-Boxers) that they have been in error up until now. Here is a second way: an evil and intolerant tyrant, bent on eliminating the scourge of One-Boxing once and for all, seizes power and initiates a systematic and ultimately wholly successful campaign of persecution against the One-Boxers. (Again, in these circumstances, I assume that the mere absence of disagreement is of no epistemic significance at all.) These cases, clearly, lie at opposite ends of a certain spectrum. Consider finally a third possible world in which disagreement about Newcomb's Problem is absent. In this possible world, there is no evil tyrant, nor is there any ingenious argument that inspires rational conviction in all of those who consider it. The only known arguments that are thought relevant to Newcomb's Problem are exactly those arguments that *we* presently possess. The only difference between this possible world and our own world is the following. In this possible world, everyone who has studied Newcomb's Problem happens to be a One-Boxer, because everyone who

[16] In his original presentation of Newcomb's Problem, Robert Nozick wrote: 'I have put this problem to a large number of people . . . To almost everyone it is perfectly clear and obvious what should be done. The difficulty is that these people seem to divide almost evenly on the problem, with large numbers thinking that the opposing half is just being silly' (1969: 48). My sense is that the by-now over three decades worth of sustained debate on Newcomb's Problem has resulted in a significant shift in the original distribution of opinion in favor of a policy of Two-Boxing. But I will abstract away from this fact in what follows: what is crucial for my purposes is simply that there are some actual defenders of One-Boxing as well as some actual defenders of Two-Boxing. (In what follows, one might consider the actual world as it stood *circa* 1969, as reported by Nozick.)

has studied Newcomb's Problem has been convinced by the very arguments that convince One-Boxers in our world. There just are, as a matter of contingent fact, no actual defenders of Two-Boxing—although the known arguments for Two-Boxing are just as strong as the known arguments for Two-Boxing in our world. (Indeed, they are the same arguments.) Moreover, there is no *deep* explanation of *why* this is so—it is not as though the brain chemistry of the inhabitants of this world differs from ours, in a way that makes them peculiarly susceptible to the allures of One-Boxing. It is just that in this possible world, everyone who has thought about it up until now finds the case for One-Boxing more compelling, and there is thus complete consensus that One-Boxing is the uniquely rational strategy.

Do these empirical and contingent facts about the state of opinion make a difference about what it is rational to believe about Newcomb's Problem? Imagine an intelligent student who sets out to study Newcomb's Problem. She scrupulously exposes herself to all of the arguments and intuition pumps that favor One-Boxing and to all of the arguments and intuition pumps that favor Two-Boxing. In the process of thinking about the problem, she increasingly comes into contact with others who have thought about the problem, and she naturally begins to take note of their views. In our world, the student finds that roughly half of those she meets are One-Boxers and half Two-Boxers. In the other possible world, she finds that everyone she meets is a One-Boxer. Having thoroughly investigated the issue, she thus resolves to make up her own mind about Newcomb's Problem. Should she take a different view about Newcomb's Problem in the other, unanimous world than she does in our fragmented and divided world? *Despite* the fact that she has access to exactly the same arguments in both worlds? This seems extremely dubious—after all, can't the student in the unanimous possible world simply look over at our own fragmented world, and realize that *here* she has epistemic peers who extol Two-Boxing? But to judge that there are close possible worlds in which individuals can rationally take certain considerations as warranting a given belief is just to make a judgement about the probative force of those considerations themselves.

Interestingly, there are philosophical questions with respect to which the state of informed philosophical opinion in our world seems to be unanimous, or very close to unanimous, in much the way that the state of informed philosophical opinion about Newcomb's Problem is unanimous in my imaginary world. Consider the way in which radical forms

of skepticism—about the external world, or about other minds, for example—have traditionally been treated within epistemology. In view of how seriously skepticism has been taken throughout the history of philosophy, it is a striking fact that there have been relatively few genuine skeptics. I am unaware of a single contemporary philosopher, for example, who genuinely believes that she does not know whether anyone besides herself has a mind.[17]

There are, of course, various possibilities here. One possibility is that I am just factually wrong—there are, in fact, a significant number of philosophers who believe that they don't know whether anyone else has a mind, but, for understandable reasons, they don't bother announcing this belief to the rest of the world. (As Bertrand Russell once observed, there is no doubt something extremely odd about a genuine skeptic about other minds who makes a point of *professing* this belief to others.) Another, perhaps more important possibility is the following. It might be that there are a considerable number of individuals who would be genuine skeptics, but that it is simply *psychologically impossible* (or very nearly so) to believe the conclusion of a skeptical argument except at the moment when one is attending to the argument, if then. (Here I have in mind the kind of epistemic weakness of the will that Hume so famously made vivid in book 1 of the *Treatise*.) However, the fact that there have been few if any genuine skeptics about other minds is not, I think, primarily due to the fact that individuals find themselves simply psychologically unable to believe the conclusions of skeptical arguments. Rather, there have been, I think, very few individuals who have believed that there is some sound argument for skepticism about other minds. Of course, many philosophers have defended skeptical arguments by attempting to show that particular objections to their soundness are misguided, or even that all extant objections are misguided. Some philosophers no doubt believe that we have yet to produce good objections to skeptical arguments, or even that we can reasonably hope to find good objections to skeptical arguments in the future. But all of these broadly sympathetic stances *vis-à-vis* skepticism are much weaker than genuine skepticism, in the sense of believing that there is some sound argument that has as its conclusion 'I cannot know that

[17] Peter Unger seems to have been an exception at the time of his (1975) but later writings reveal that his attitudes towards skepticism have evolved considerably; see e.g. his (1984: ch. 3).

there is any mind other than my own'. And it is in this sense, I think, that there have been few if any genuine skeptics about other minds.[18]

Nevertheless, the relative absence of genuine skeptics has not been taken to be a significant fact in the assessment of skepticism itself. That is, in assessing the case for skepticism, the discussion has been about the probative force of skeptical arguments. The contingent fact (assuming that it is a fact) that there are few if any philosophers who actually endorse some skeptical argument as a sound argument has not been taken to be relevant. To put it in another way: it has *not* been considered a good *objection* to skepticism to simply note that there are few if any genuine skeptics. We can, of course, easily imagine that things are otherwise—that is, we can imagine that philosophical opinion about the truth of skepticism about other minds is more or less evenly divided in our world between genuine skeptics and non-skeptics, in much the way that philosophical opinion is genuinely divided between One-Boxers and Two-Boxers. Would the case for skepticism about other minds be any stronger if it were so? Given that the best arguments offered by the genuine skeptics are simply *our* best arguments? In general, it has been thought—correctly, I believe—that the case for skepticism stands or falls with the probative force of skeptical arguments and does not depend on contingent and empirical facts concerning the actual existence or nonexistence of skeptics. As Laurence BonJour has written: 'the need to consider skepticism does not depend in any crucial way . . . on whether or not serious proponents of skepticism are actually to be found; if skeptics did not exist, one might reasonably say, the serious epistemologist would have to invent them' (1985: 14–15).

6. The Views of One's Peers as Higher-Order Evidence

It is a presupposition of the issue under discussion that we are fallible with respect to our ability to correctly appreciate our evidence. Of course, reasonable individuals are disposed to respond correctly to their evidence. But even generally reasonable individuals are susceptible

[18] The relatively recent advent of skeptic-friendly varieties of contextualism (e.g. Lewis 1996) might cause some difficulties for this (admittedly rough) construal of what counts as 'genuine skepticism'. But not, I think, in a way that materially affects the point at issue.

to making mistakes on particular occasions. The possibility of error makes the following question salient: how do we know what our evidence supports? Could one have evidence which is relevant to the question of what one's evidence supports? If so, what would such evidence consist of?

Given that reasonable individuals are disposed to respond correctly to their evidence, the fact that a reasonable individual responds to her evidence in one way rather than another is itself evidence: it is evidence about her evidence. That is, the fact that a (generally) reasonable individual believes hypothesis H on the basis of evidence E is some evidence that it is reasonable to believe H on the basis of E. The beliefs of a reasonable individual will thus constitute *higher-order* evidence, evidence about the character of her first-order evidence. Of course, such higher-order evidence, like most other evidence, is not conclusive evidence: it does not follow from the fact that a generally reasonable individual believes H on the basis of E that it is reasonable to believe H on the basis of E. In a case in which E does not adequately support H but a generally reasonable individual mistakenly believes H on the basis of E, the fact that the individual believes as she does constitutes *misleading* evidence about the character of the evidence E. But misleading evidence is evidence nonetheless. In general, then, the fact that a reasonable person believes H on the basis of E constitutes evidence about the character of E.

Given the general reasonableness of one's epistemic peers, what they believe on the basis of one's shared evidence will thus constitute evidence about what it is reasonable to believe on the basis of that evidence. There are subtle questions, I think, about how one should integrate such higher-order considerations into one's own deliberations and what difference such considerations make to what it is reasonable for one to believe. At the very least, evidence about one's evidence will make a difference to what it is reasonable for one to believe *about one's evidence*. Will such higher-order evidence also make a difference to what it is reasonable for one to believe about propositions that are *not* about one's evidence? Let E represent our shared total evidence with respect to H. Consider the epistemic proposition that

(1) E is good evidence that H is true

On the present view, if I discover that you believe that H on the basis of E, I should treat this discovery as confirming evidence for (1). Should

I also treat it as confirming evidence for H itself? If I discover instead that you believe that not-H on the basis of E, this discovery would constitute disconfirming evidence for (1). Would it also constitute evidence against H?

Here is a reason for thinking that I should *not* treat the evidence for or against (1) that is afforded by your believing as you do as evidence for or against H itself. Imagine that I have yet to make up my mind about H: that is, I am in the process of actively deliberating about what attitude I should take up towards the hypothesis. Suppose further that I find that you believe H on the basis of our shared first-order evidence E. If I treat the fact that you believe as you do as an additional piece of evidence which bears on the truth of H, then, when I enumerate the considerations which tend to confirm H, I will list not only the various first-order considerations that speak in favor of H, but also the fact that you believe that H is true. That I treat your belief in this way might seem to involve a certain admirable modesty or humility on my part. But notice that, when *you* enumerate the reasons why *you* believe that H is true, you will list the various first-order considerations that speak in favor of H—but presumably, *not* the fact that you yourself believe that H is true. From your perspective, the fact that you believe as you do is the *result* of your assessment of the probative force of the first-order evidence: it is not one more piece of evidence to be placed alongside the rest. That is, you do not treat the fact that you believe H as a further reason to believe that H, above and beyond the first-order considerations that you take to rationalize your belief. (If you subsequently changed your mind and came to doubt that the first-order evidence was sufficient to rationalize your believing H, you would not treat the fact that you believe that H as a reason to continue believing it. Similarly, when you first came to believe that H on the basis of your initial consideration of the first-order evidence E, you did not then proceed to treat the fact that *I believe that H is true* as a reason to *increase* your confidence that H is true. Rather, you arrived at that level of confidence which you thought appropriate given the first-order evidence E.) I am thus in the awkward position of treating your belief that H as a reason to believe that H, despite the fact that you do not treat this as an epistemically relevant consideration. Again, it might make sense for me to treat your belief in this way if I lacked access to your first-order evidence: in that case, your belief stands in as a sort of proxy for the evidence on which it is based (cf. § 3 above). But when I do have access to your

first-order evidence for H, and I continue to treat the belief that you have formed in response to that evidence as a further reason to believe that H, aren't I essentially engaged in a kind of double-counting with respect to the relevant evidence?[19]

Perhaps the relevance of my knowing that you believe as you do with respect to a given question is much like the relevance of an insurance company's knowledge that some particular driver happens to be a teenager. Because teenage drivers are, taken as a group, more reckless than other drivers, it makes sense for an insurance company to give a considerable amount of weight to this fact in particular cases. But if the insurance company had direct access to the underlying facts about the actual recklessness of a particular teenager, then this person's age would be rendered an irrelevant piece of information, and continuing to give weight to it would be to engage in a kind of illegitimate double-counting. In the language of the statisticians: access to the underlying facts about the actual recklessness of the driver 'screens off' knowledge of the driver's age, rendering the latter probabilistically irrelevant. Similarly, it might be that my having access to all of the evidence on which you base your belief screens off the fact that you believe as you do.[20]

[19] I have assumed that, when you enumerate the considerations that you take to bear on the truth of the hypothesis H, you will not include your own belief that H is true among those considerations. Consider, however, the view known as *epistemic conservatism*. According to epistemic conservatism, the mere fact that one presently believes that H makes it normatively appropriate to go on believing H, in the absence of positive reasons for abandoning that belief. Suppose that epistemic conservatism is, in fact, a correct view about belief revision. In that case—it might be argued—you *ought* to treat the fact that you believe that H as a reason to believe that H.

But this suggestion misunderstands the nature of epistemic conservatism. Adherents of epistemic conservatism typically do not present their view as implying that one possesses a reason to believe a proposition in virtue of believing that proposition. Rather, the view is that one does not *need* a reason for it to be normatively appropriate to continue believing a proposition that one already believes. (Beliefs are 'innocent until proven guilty', as opposed to the more traditional view that they are 'guilty until proven innocent'.)

For endorsements of epistemic conservatism, see Sklar (1975), Harman (1986), and Quine and Ullian (1978). For criticism, see Foley (1987), Vogel (1992), and Christensen (1994).

[20] Compare also the legal norm of 'Best Evidence'. If an original document is unavailable, a transcription of the original might be admitted as evidence of its author's intentions. But if the original document *is* available, then the transcription is considered inadmissible. The underlying thought, of course, is that while the transcription might have significant evidential value in the absence of the original, it is rendered irrelevant by the original's presence, since whatever evidential value it does have is exhausted by its (perhaps imperfect) reflection of the contents of the original. Similarly, one might think that, since the evidential value of the belief of some other party ultimately depends on the

At the very least then, there seems to be a certain awkwardness in my giving additional weight to your belief that H is true when I have already taken into account all of those considerations on which your belief is based, considerations which you take to exhaust the case for H. I do not think that this line of thought is decisive, however. Issues about how one's higher-order evidence does or does not interact with one's first-order evidence when that first-order evidence is itself available are, I think, extremely complicated. I will not attempt to resolve these issues here.[21] Rather, in the remainder of this section, I will argue that even if we *do* treat the higher-order evidence that is provided by the views of our epistemic peers as further evidence that bears on the disputed questions themselves, it does not follow that skepticism or agnosticism is the reasonable response to disagreements of the relevant kind.

Again, let E represent our total evidence with respect to H at time t0. In order to avoid premature complications, let's suppose that each of us is ignorant of the other's existence at this point.[22] Let's further stipulate that E is such as to rationalize the belief that H. Recognizing this fact, you form the reasonable belief that H at time t1, an instant later. Unfortunately, however, I badly misjudge the probative force of the evidence E at time t1 and take up the unreasonable belief that not-H.

At time t1 then, prior to our learning about the other person, the situation stands as follows. You hold the reasonable belief that H on the basis of your total evidence E while I hold the unreasonable belief that not-H on the basis of my total evidence E. The asymmetry in the epistemic statuses of our respective beliefs is due simply to the fact that E *really does* support H and does not support not-H.

fact that is a (perhaps imperfect) reflection of some more fundamental piece of evidence on which it is based, the belief is rendered irrelevant by the presence of the more fundamental piece of evidence (even if the same belief would be highly relevant in the absence of the more fundamental piece of evidence).

[21] For further discussion, see Kelly (forthcoming *a*).

[22] Of course, it might be that the most typical way for two individuals to have the same evidence is for them to have shared their evidence with one another—or at least, for both of them to be members of some community which shares its evidence (think of the Compatibilists and the Incompatibilists here). But it is, I assume, at least logically possible for two individuals to have arrived at the same evidence independently of one another. I want to begin, then, by considering what's true in a case in which you and I have the same evidence, but where both of us are ignorant of the fact that there is someone else who has exactly that evidence.

Suppose that we become aware of our disagreement at time t2. According to the view in question, our total evidence with respect to H has now changed. Let's call our new total evidence at time t2 E′. What does E′ include? E′ will include the following:

E′ = (i) the original, first-order evidence E,
(ii) the fact that you believe H on the basis of E, and
(iii) the fact that I believe not-H on the basis of E.

The crucial fact here is the following: there is no reason to think that the new evidence E′ will invariably mandate an attitude of abstention or agnosticism with respect to the hypothesis H. In particular, there is no reason to think that your continuing to believe H is unreasonable on evidence E′ given that it was reasonable when your total evidence consisted of E. For in the usual case, the character of the new evidence E′ will depend a great deal on the character of the original evidence E. Indeed, if we give equal weight to (ii) and (iii), then H will be more probable than not-H on the new evidence E′, given that it was more probable on the original evidence E. Our original evidence E does not simply vanish or become irrelevant once we learn what the other person believes on the basis of that evidence: rather, it continues to play a role as an important subset of the new total evidence E′. In general, what one is and is not justified in believing on the basis of E′ will depend a great deal on the character of the first-order evidence E.

Thus, even if one treats the higher-order evidence which is provided by the beliefs of one's epistemic peers as evidence which bears on the disputed theses, it does not follow that agnosticism or suspension of judgement is the correct response to such disputes.

7. Actual Disagreement Reconsidered

I have argued that disagreement does not have the kind of epistemic significance that has sometimes been claimed for it. Still, it would be a mistake to think that disagreement is therefore epistemically *in*significant. What epistemic role or roles are left for disagreement, on the view that I have defended? Of course, an awareness of disagreement can serve to call one's attention to arguments that one might never have considered or to evidence of which one might have been unaware. However, even when all parties to a dispute have access to the same evidence and

arguments, disagreement can still play an epistemically salutary role. In the last section, I noted that the views of one's epistemic peers provide higher-order evidence. In this section, I want to highlight two other important roles that disagreement can play in cases of shared evidence.

First, it might be that the presence of disagreement with respect to some question at earlier times tends to produce a better pool of evidence bearing on that question at later times. That is, over time, the goals of inquiry might be best promoted when there is a diversity of opinions among inquirers. This theme has been endorsed and developed by a distinguished tradition of thinkers, a tradition which includes John Stuart Mill, Frederick Hayek, and Paul Feyerabend.[23]

In addition, there is, I want to suggest, a more subtle way in which disagreement can prove epistemically beneficial. My suggestion is that the role of actual disagreement among epistemic peers is analogous to the role that actuality sometimes plays in falsifying modal claims that are mistakenly thought to be justified a priori.

Taken as a class, philosophers are somewhat notorious for making claims, ostensibly justified a priori, about *what must be the case*, or *what could not be otherwise*, that are subsequently falsified by empirical discoveries.[24] Not only does a putatively a priori necessary truth fail to hold in all possible worlds, it does not even hold in our own, actual world. (The logical positivists often accused Kant of making this mistake.) Presumably, there is a sense in which these empirical discoveries were not essential to falsifying the modal claim in question: someone with sufficient imaginative abilities would not need actual, empirically discovered counterexamples in order to see that the modal claim is false. But because human beings not infrequently suffer from persistent blindspots or failures of imagination, actuality occasionally plays a key role in falsifying such modal claims. (Once the modal claim is seen to be false, it can then come to seem *obviously* false; additional counterexamples are easy to come by, and it can seem almost embarrassing that we needed an empirical discovery in order to perceive its falsity.)

I suggest that something analogous is true of the role of actual dissenters. In principle, we ought to be able to give due weight to the

[23] Mill; Hayek (1960); Feyerabend (1975). A contemporary philosopher who has further developed this general theme is Philip Kitcher; see especially his (1993).

[24] For a recent excoriation of philosophers on this score, see Nozick (2001: esp. ch. 3).

available reasons that support a given view, even in the absence of actual defenders of the view who take those reasons as compelling. But in practice, the case for a view is apt to get short shrift in the absence of any actual defenders. The existence of actual defenders can serve to overcome our blindspots by forcefully reminding us of just how formidable the case is for the thesis that they defend, just as actual counterexamples are sometimes needed to overcome our blindspots concerning modality. But the case for a given view itself is no stronger in virtue of the fact that that view has actual defenders—just as a genuine counterexample to a modal claim is no stronger in virtue of being an actual, empirically discovered counterexample.

8. CONCLUSION: EPISTEMIC EGOISM WITHOUT APOLOGY

I have argued that disagreements of the relevant kind do not provide a compelling basis for skepticism. The mere fact that others whom I acknowledge to be my equal with respect to intelligence, thoughtfulness, and acquaintance with the relevant data disagree with me about some issue does not undermine the rationality of my maintaining my own view. I admit to finding this conclusion somewhat unsettling. Among my reasons for finding it unsettling is the following: many of those whom I take to be my epistemic peers disagree with me about this issue. Disappointingly, even some of those whom I would expect to be most sympathetic to my view given their own practice tend to argue against it as a matter of theory.

That I find it unsettling that many people I know and respect disagree with me about the epistemic significance of disagreement is perhaps unsurprising. There are, after all, psychological studies that suggest that we are highly disposed to being greatly influenced by the views of others, and I have no reason to think that I am exceptional with respect to this particular issue. It is, of course, a different question whether the fact that many others disagree with my thesis provides a good reason for me to doubt that thesis. And my answer to this question, as might be expected, is 'No': because I accept the general thesis that known disagreement is not a good reason for skepticism, I do not, in particular, regard the fact that people disagree with me about this general thesis as

a reason for being skeptical of *it*. Although I tend to find it somewhat unsettling that many disagree with my view, I am inclined to regard this psychological tendency as one that I would lack if I were more rational than I in fact am. In contrast to my psychological ambivalence, my considered, reflective judgement is that the fact that many people disagree with me about the thesis that disagreement is not a good reason for skepticism is not itself a good reason to be skeptical of the thesis that disagreement is not a good reason for skepticism.

The fact that I both endorse this thesis and refuse to take the fact that others disagree with me as a compelling reason for doubting its truth means that my views have a certain kind of internal coherence. This kind of internal coherence is not trivial: all combinations of views do not have it. However, I am not inclined to put too much weight on this kind of internal coherence, for this particular virtue proves surprisingly robust. Suppose, for example, that despite my considered judgement I one day do give in to the psychological pressure occasioned by the fact that so many of those who I know and respect disagree with me, and I abandon my thesis. (In the question-and-answer session following a talk at which I present these ideas, all of the questioners make it clear that they think that my thesis is clearly false. It is not that anyone provides a rationally compelling argument for this conclusion; rather, I am simply overwhelmed by my ever-increasing ideological isolation.) From my present vantage point, the envisaged change in my beliefs seems to be a craven (if understandable and all too predictable) capitulation to brute psychological pressure. *After* I have changed my mind about the epistemic significance of disagreement, however, it is of course open to me to look upon my recent conversion in a much more charitable light. I have changed my mind, after all, because I am influenced by the fact that others disagree with me, and this—according to the view that I will *then* hold—is the epistemically rational response to an epistemically relevant consideration. My later self might then say: my fundamental epistemic rationality—that is, the responsiveness of my beliefs to considerations that are in fact epistemically relevant—won out, in the end, over my misguided adherence to a mistaken philosophical thesis that would have permitted me to treat these epistemically relevant considerations as though they were irrelevant. So it looks as though, either way, a certain amount of self-congratulation will seem to be in order in the future.

REFERENCES

Alston, William (1988) 'The Deontological Conception of Epistemic Justification', in James Tomberlin (eds.), *Philosophical Perspectives 20: Epistemology* (Atascadero, Calif.: Ridgeview), 257–99.

Asch, Solomon (1952) *Social Psychology* (Englewood Cliffs, NJ: Prentice-Hall).

—— (1956) *Studies of Independence and Conformity: A Minority of One against a Unanimous Majority*, Psychological Monographs 70(9).

Aumann, Robert J. (1976) 'Agreeing to Disagree', *Annals of Statistics*, 4(6): 1236–9.

BonJour, Laurence (1985) *The Structure of Empirical Knowledge* (Cambridge, Mass.: Harvard University Press).

Christensen, David (1994) 'Conservatism in Epistemology', *Nous*, 28(1): 69–89.

Feyerabend, Paul (1975) *Against Method* (London: Verso).

Field, Hartry (2000) 'Apriority as an Evaluative Notion', in Boghossian and Peacocke (eds.), *New Essays on the A Priori* (Oxford: Oxford University Press), 117–49.

Foley, Richard (1987) *The Theory of Epistemic Rationality* (Cambridge, Mass.: Harvard University Press).

—— (1994) 'Egoism in Epistemology', in F. Schmitt (ed.), *Socializing Epistemology* (New York: Rowman & Littlefield), 53–73.

—— (2001) *Intellectual Trust in Oneself and Others* (Cambridge: Cambridge University Press).

Geanakoplos, John (1994) 'Common Knowledge', in *Handbook of Game Theory*, ii, ed. R. J. Aumann and S. Hart (Amsterdam: Elsevier Science), 1437–96.

Gutting, Gary (1982) *Religious Belief and Religious Skepticism* (Notre Dame: University of Notre Dame Press).

Harman, Gilbert (1986) *Change in View* (Cambridge, Mass.: MIT Press).

Hayek, Friederich (1960) *The Constitution of Liberty* (Chicago: University of Chicago Press).

Kelly, Thomas (forthcoming *a*). 'Confidence and Belief Revision', in John Hawthorne and Dean Zimmerman (eds.), *Philosophical Perspectives 20: Epistemology* (Oxford: Blackwell).

—— (forthcoming *b*) 'Reasoning about Evidence one does Not Possess'.

Kitcher, Philip (1983) *The Advancement of Science* (New York: Oxford University Press).

Leher, Keith (1976) 'When Rational Disagreement is Impossible', *Noûs*, 10: 327–32.

Lewis, David (1979) 'Scorekeeping in a Language Game', *Journal of Philosophical Logic*, 8: 339–59. Reprinted in his *Philosophical Papers*, i (Oxford: Oxford University Press, 1983).

—— (1996) 'Elusive Knowledge'. Reprinted in his *Papers in Metaphysics and Epistemology* (Cambridge: Cambridge University Press, 1999).

Mackie, J. L. (1977) *Ethics: Inventing Right and Wrong* (New York: Penguin).

Mill, John Stuart (1859) *On Liberty* (many editions).

Nozick, Robert (1969) 'Newcomb's Problem and Two Principles of Choice'. First appeared in N. Rescher *et al.* (eds), *Essays in Honor of Carl G. Hempel*. Reprinted in Nozick's collected papers, *Socratic Puzzles* (Cambridge, Mass.: Harvard University Press, 1997), 44–73. Page references are to the reprinted version.

—— (2001) *Invariances: The Structure of the Objective World* (Cambridge, Mass.: Harvard University Press).

Plantinga, Alvin (1993) *Warrant: The Current Debate* (Oxford: Oxford University Press).

—— (1995) 'Pluralism: A Defense of Religious Exclusivism', in Thomas Senor (ed.), *The Rationality of Belief and the Plurality of Faith* (Ithaca, NY: Cornell University Press).

—— (2000) *Warranted Christian Belief* (Oxford: Oxford University Press).

Quine, W.V., and J. Ullian (1978) *The Web of Belief* (New York: McGraw-Hill).

Quinn, P., and K. Meeker (eds.) (2000) *The Philosophical Challenge of Religious Diversity* (New York: Oxford University Press).

Sextus Empiricus (2000) *Outlines of Skepticism*, ed. Julia Annas and Jonthan Barnes (Cambridge: Cambridge University Press).

Sidgwick, Henry (1981) *The Methods of Ethics* (Cambridge: Hackett).

Sklar, Lawrence (1975) 'Methodological Conservatism', *Philosophical Review*, 74: 374–401.

Unger, Peter (1975) *Ignorance: A Case for Skepticism* (Oxford: Oxford University Press).

—— (1984) *Philosophical Relativity* (Minnesota: University of Minnesota Press).

van Inwagen, Peter (1996) ' ''It is Wrong, Always, Everywhere, and for Anyone, to Believe Anything, Upon Insufficient Evidence'' ', in J. Jordan and D. Howard-Snyder (eds.) *Faith, Freedom, and Rationality* (Lanham, Md.: Rowman & Littlefield), 137–54.

Vogel, Jonathan (1992) 'Sklar on Methodological Conservatism', *Philosophy and Phenomenological Research*, 52: 125–31.
Williamson, Timothy (2000) *Knowledge and its Limits* (Oxford: Oxford University Press).

8. The Assessment Sensitivity of Knowledge Attributions

John MacFarlane

Recent years have seen an explosion of interest in the semantics of knowledge-attributing sentences, not just among epistemologists but among philosophers of language seeking a general understanding of linguistic context sensitivity. Despite all this critical attention, however, we are as far from consensus as ever. If we have learned anything, it is that each of the standard views—invariantism, contextualism, and sensitive invariantism—has its Achilles heel: a residuum of facts about our use of knowledge attributions that it can explain only with special pleading. This is not surprising if, as I will argue, there is a grain of truth in each of these views.

In this paper, I propose a semantics for ''know'' that combines the explanatory virtues of contextualism and invariantism. Like the contextualist, I take the extension of ''know'' to be sensitive to contextually determined epistemic standards. But where the contextualist takes the relevant standards to be those in play at the context of *use*, I take them to be those in play at the context of *assessment*: the context in which one is assessing a particular use of a sentence for truth or falsity. Thus, I can agree with the invariantist that ''know'' is not sensitive to the epistemic standards in play at the context of use, while still acknowledging a kind of contextual sensitivity to epistemic standards. The proposed semantics

I presented versions of this paper to the Stanford Philosophy Department on 17 Oct. 2003, to the Themes in Philosophy of Language conference at Yale on 8 Nov. 2003, and to the Department of Logic and Philosophy of Science at UC Irvine on 5 Dec. 2003. I am grateful to audiences at all three talks for stimulating discussions, and especially to Keith DeRose, who commented on my paper at Yale. I would also like to thank Kent Bach, Gilbert Harman, Ram Neta, Jonathan Schaffer, Lionel Shapiro, and Matt Weiner for useful correspondence and discussion. This work was made possible in part by an ACLS/Andrew W. Mellon Fellowship for Junior Faculty and a UC Berkeley Humanities Research Fellowship.

for ''know'' is contextualist along one dimension (contexts of assessment) and invariantist along another (contexts of use).[1]

In the first part of the paper, I motivate my proposal by considering three facts about our use of ''know'' (§2) that collectively cause trouble for *all* of the standard views about the semantics of ''know'' (taxonomized in §1). I argue that the usual attempts to explain away the anomalies by appeal to pragmatics or to speaker error are unpersuasive (§3). In §4, I show how standard semantic frameworks must be modified to make room for my ''relativist'' semantics, and I show how the proposed semantics makes sense of the features of our use of ''know'' that proved puzzling on the standard views. Finally, in §5, I respond to worries about the coherence of relativist semantics by describing the role assessment-relative truth plays in a normative account of assertion.

1. A TAXONOMY

For our purposes, the standard views about the semantics of ''know'' can be divided into three main classes. *Strict invariantists* hold that ''know'' is associated with a fixed epistemic standard, in much the same way as ''six feet apart'' is associated with a fixed standard of distance. A person and a fact satisfy ''*x* knows *y*'' just in case the person's epistemic position with respect to the fact is strong enough to meet this fixed epistemic standard. *Sensitive invariantists* allow the epistemic standard to vary with the subject and the *circumstances of evaluation* (in the sense of Kaplan 1989), in much the same way as the standard of distance expressed by ''as far apart as Mars and Jupiter'' varies with the circumstances (for instance, the time) of evaluation. And *contextualists* allow the epistemic standard to vary with the *context of use*, like the standard of distance expressed by ''as far apart as my two hands are right now.'' The differences are summed up in Figure 8.1.

This is of course only one way of carving up the range of positions that have been taken, and it lumps together positions that may seem

[1] For a kindred view, developed rather differently, see Richard 2004. I learned of Richard's work too late to take account of it in this paper. There are also some affinities between the present proposal and the ''perspectival'' view of knowledge attributions defended in Rosenberg 2002: ch. 5 (see esp. 148–9, 163–4), though Rosenberg does not develop his proposal in a truth-conditional framework.

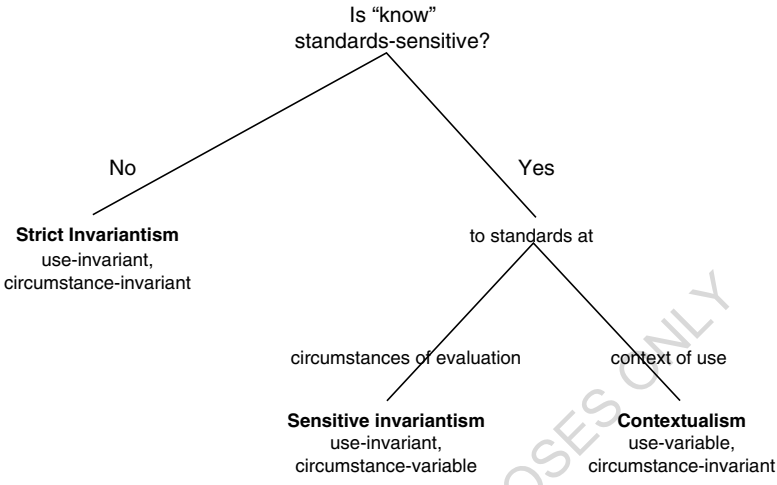

Figure 8.1. Standard taxonomy of positions on the semantics of ''know''

very different, even from a semantic point of view.[2] The advantage of this taxonomy is that it will allow us to see in a perspicuous way what is wrong with *all* of the views it encompasses. Because a ''formal'' taxonomy will be enough for our purposes, I leave it completely open here what an epistemic position is, how an epistemic standard might be specified, and what features determine which epistemic standard is relevant in a given context or circumstance. In particular, although I will sometimes talk of ''high'' and ''low'' standards, I wish to leave it open whether standards vary on a linear scale—from ''low'' to ''high''—or in a more complex and qualitative way, as on ''relevant alternatives'' theories. Different views in each of our formal categories will cash out these notions in different ways. The arguments that follow abstract from these details.

The differences between contextualism and sensitive invariantism tend to be obscured when we consider first-person, present-tense knowledge attributions. For in these cases the epistemic standards in play at

[2] e.g. it classes the view advocated in Kompa 2002 as a form of contextualism, even though on Kompa's view ''know'' always expresses the same, ''unspecific'' relation, and so would be counted a form of invariantism on some criteria. Kompa's view will be discussed in §4.2.

the context of use will coincide with those in play for the subject at the circumstances of evaluation. To see the differences, we need to vary the context of use while keeping the circumstances of evaluation constant—say, by considering ''On Tuesday, Joe knew that whales are mammals'' as uttered by Sally on Wednesday and by Fred on Thursday—and vary the circumstances of evaluation while keeping the context of use constant—say, by considering both ''On Tuesday, Joe knew that whales are mammals'' and ''On Wednesday, Joe knew that whales are mammals'' as uttered by Fred on Thursday. Contextualism predicts that the epistemic standard one must meet in order to count as ''knowing'' should shift as we shift the context of use (even if the circumstances of evaluation are kept fixed), while sensitive invariantism predicts that it should shift as we shift the circumstances of evaluation (even if the context of use is kept fixed). Thus we may aptly describe a contextualist semantics for ''know'' as *use-variable* and a sensitive invariantist semantics as *circumstance-variable*. Standard versions of contextualism are *circumstance-invariant*, and standard versions of sensitive invariantism are *use-invariant*—though of course one might also have a hybrid view that was both use-variable and circumstance-variable. Strict invariantism is both use-invariant and circumstance-invariant.

2. SOME FACTS ABOUT OUR USE OF ''KNOW''

I now want to look at three facts about our use of knowledge attributions that have figured prominently in discussions of the semantics of ''know''.

2.1. Variability of Standards

Normally, I am perfectly happy to say that I know that my car is parked in my driveway. I will say this even when I'm at work, several miles away. But if someone asks me how I know that my car has not been stolen (and driven away), I will admit that I do not know this. And then I will have to concede that I do not know that my car is in my driveway: after all, if I knew this, then I would be able to deduce, and so come to know, that it has not been stolen.

How should we think of my shift from claiming to know to denying that I know? It doesn't seem right to describe me as having learned

something, or as correcting a mistake. If I have learned something, what exactly have I learned? It's not as if I was unaware of the existence of car thieves when I made my original knowledge claim. Besides, the next day I will go right back to saying that I know that my car is in my driveway. Am I so dense as never to learn from my mistakes?

Nor does it seem right to say that, when I claimed to know, I didn't mean it literally. I would have said the same thing in a forum where non-literal speech is discouraged, like a courtroom. And I would have said the same thing if I had been instructed to say just what I meant, without exaggeration, artifice, or innuendo. Indeed, I would have said the same thing in a crowd of epistemologists, so it was not just a matter of ''speaking with the vulgar''.

Perhaps my mistake lies in conceding that I don't know that the car is in the driveway. Perhaps the fact that I cannot rule out auto theft is actually irrelevant to whether I know. But what, then, *should* I say when the possibility is floated? Should I ignore it or dismiss it as irrelevant? That might be the right response to certain far-fetched sceptical worries—say, ''how do you know that the matter in your car has not spontaneously reorganized to form a giant lizard?''—but it hardly seems appropriate in response to a perfectly mundane worry about thieves. Should I say ''Although I know that the car is in my driveway, there's always a chance that it has been stolen and is not in my driveway''? This sounds close to contradictory. Or should I say ''Since I know that the car is in my driveway, I know that it hasn't been stolen''? That too seems wrong. I am not in a position to know that the car has not been stolen. If I am making a mistake, it is not one that ordinary speakers recognize as a mistake.

If I was speaking literally both times and didn't make a mistake, then presumably the standards I must meet in order to count as ''knowing'' must have changed. I met the laxer standards that were in play at the time of my first knowledge claim, but not the stricter ones that came into play after the mention of car thieves.

Examples like this can easily be multiplied. They form the basis of standard arguments for contextualism and sensitive invariantism.

2.2. *Embedded Occurrences of ''Know''*

Temporal and modal operators shift the circumstances of evaluation. But we seem to use the same epistemic standard in evaluating ''know''

when it is embedded in the scope of temporal or modal operators as we do when it occurs unembedded. We don't seem to *mix* different standards at a single context of use, even when we're considering putative knowers in very different circumstances.

To take up the example from the last section, when I concede that I don't know that my car is parked in the driveway, I won't insist that I *did* know this two minutes ago, before the bothersome question raised the standards. I will say that I did not know it then either. In deciding whether I knew it then, I use the standards in play now, not the standards that were in play then.

Relatedly, we do not say things like ''Before the possibility that he might win the lottery became relevant, John knew that he would not be able to afford health insurance, but now he does not know this (though he still believes it)'', or ''John knows that he won't be able to afford health insurance, but if he were discussing the possibility that he might win the lottery, he would not know this.'' If the judge asks Doris whether she knew on 13 January that her car was in the driveway, it would be positively bizarre for her to answer ''I don't know: I can't remember whether I was worried about car thieves that day'' or ''Remind me: what epistemic standards were in play at that time?'' All this suggests that at any given context of use, we hold the standards that one must meet in order to count as ''knowing'' constant over all circumstances of evaluation. Observations such as these form the basis of standard arguments against sensitive invariantism (see DeRose 2000 and 2004*b*).

2.3. *Truth Ascriptions and Retraction*

When standards have been raised, I will say not only that I don't know that my car is in my driveway, and that I *didn't* know this earlier, but that my earlier assertion of ''I know that my car is in the driveway'' was *false* (noticed by Feldman 2001: 77; Rosenberg 2002: 164; Hawthorne 2004: 163; among others). In part, this is because we tend to report knowledge claims homophonically, even when they were made in very different epistemic contexts (see Hawthorne 2004: §2.7). Thus, I will report myself as having asserted that I knew that my car was in the driveway. Since I now take myself not to have known this, I must reckon my earlier assertion false.

I won't just *say* that it was false; I will *treat* it as false. If challenged, I will *retract* my earlier claim, rather than reformulating it in a way that

shows it to be consistent with my current claim—for example, by saying, ''What I asserted was merely that I met the standard for 'knowing' that was in place *when I was making the claim*.''[3] I will have correlative expectations about when others ought to retract their knowledge claims. If yesterday Sally asserted ''I know that the bus will be on time,'' and today she admits that she didn't know yesterday that the bus would be on time, I will expect her to retract her earlier assertion. I will find it exceedingly bizarre if she replies by saying that her assertion was true, even if she adds ''by the standards that were in place yesterday.''

In these respects ''know'' functions very differently from ordinary indexicals like ''here'' and from other expressions generally regarded as context-sensitive, like ''flat'' and ''tall''.[4] Suppose I'm on a moving train. At 3.30 we pass some big factories and tenement houses, and I say ''It's very urban here.'' By 3.31 we have passed into suburbs, and I say ''It's not very urban here.'' I *won't* retract my earlier claim. If it is challenged, I'll say: ''When I said a minute ago 'It's very urban here,' what I said was true, and I stand by that, even though it's not very urban *here*.'' To avoid confusion, I may reformulate my earlier claim: ''What I asserted was that it was very urban where we were a minute ago.'' Similarly, if I find myself in a scientific context where tiny bumps and ridges are important, I might assert ''The table is not flat'', but I would not regard this as any reason to withdraw my assertion, made earlier in an everyday context, of ''The table is flat''. If pressed, I would say: ''I only committed myself to the table's being flat by everyday standards.''

If we are correct in ascribing truth and falsity to our earlier knowledge claims in light of present standards, and retracting or standing by them accordingly, then it seems that we do not take the epistemic standards one must meet in order to count as ''knowing'' to vary across contexts of use. This fact forms the basis of standard arguments against contextualism.

[3] As Stephen Schiffer notes, ''no ordinary person who utters 'I know that *p*,' however articulate, would dream of telling you that what he meant and was implicitly stating was that he knew that *p* relative to such-and-such standard'' (1996: 326–7). See also Feldman 2001: 74, 78–9; Hawthorne 2004: §2.7.

[4] See Stanley 2004 for a detailed discussion of differences between ''know'' and various kinds of context-sensitive expressions.

3. ASSESSING THE STANDARD VIEWS

Let's assemble the upshots of these observations. The apparent *variability of standards* suggests that the truth of sentences containing ''know'' depends somehow on varying epistemic standards. That would rule out strict invariantism. The facts about *embedded occurrences* suggest that the semantics of ''know'' is circumstance-invariant. That would rule out sensitive invariantism. And the facts about *truth ascriptions and retraction* suggest that the semantics of ''know'' is use-invariant. That would rule out contextualism. Taken at face value, then, our three facts about use seem to rule out all three standard views about the semantics of ''know''.

What should we conclude? I think we have three basic options:

1. We can argue that one of our three facts about use is a misleading guide to the semantics of ''know,'' either
   (*a*) because it can be explained pragmatically, in terms of our broader communicative purposes, or
   (*b*) because it can be attributed to systematic and widespread error on the part of ordinary speakers.
2. We can argue that our practice in using ''know'' is so confused and incoherent that knowledge-attributing sentences cannot be assigned definite truth conditions. Instead of doing semantics, we can advocate reform, perhaps through the introduction of new, unconfused terms of epistemic assessment.
3. We can try to make conceptual space for a semantics for ''know'' that is use-invariant and circumstance-invariant, but still somehow sensitive to changing epistemic standards.

My aim in this paper is to explore the last of these options, which I will take up in §4, below. But first I want to say a bit about why I find the other options unpromising.

3.1. Pragmatic Explanations of the Data

One of the most important lessons of philosophy of language in the 1960s was that the connection between meaning and use is indirect (see Grice 1989; Searle 1969: ch. 6). Even if we restrict ourselves to sincere, knowledgeable informants, the most we can discern directly from their use of sentences are the conditions in which they find it reasonable to

use these sentences to make assertions. And these are not the same as the truth conditions. It is often reasonable to make assertions using sentences one knows to be literally false—not just because it is sometimes reasonable to lie, but because it is often reasonable to engage in hyperbole, harmless simplication, irony, and metaphor. Conversely, it is often reasonable to *refrain* from asserting something that is true, germane to the topic, and potentially informative. For example, one might refrain from asserting that Harvard's university library is one of the fifty largest in the world—though this is true—because doing so would encourage certain audiences to infer that Harvard is closer to number fifty than to number one.

Thus the facts about use catalogued in the previous section do not *by themselves* rule out any proposal about the semantics of ''know''. These facts may tell us something about when people find it reasonable to use certain sentences containing ''know'' to make assertions, but they do not directly tell us anything about the truth conditions of these sentences. To get from use to truth conditions, we must rule out the possibility that it is reasonable to use these sentences *despite* their falsity, or to refrain from using them *despite* their truth. I know of no fully general way of doing this: all we can do is examine putative explanations one by one and show how they fail. Because I will consider the possibility of speaker error in §3.2, I will assume in this section that speakers are under no relevant substantive or semantic misapprehensions: when they utter false sentences, they know that they are false, and when they refrain from uttering true sentences, they know that they are true.

3.1.1. Variability of standards

Variability is primarily a problem for strict invariantists. Strict invariantists come in two varieties. *Sceptical* invariantists hold that the fixed epistemic standards are very stringent, perhaps so stringent that human beings never meet them (at least with respect to empirical facts). *Moderate* invariantists hold that the standards are meetably lax. The two kinds of invariantists face different challenges in giving a pragmatic explanation of the variability data, so I will consider them separately.

(a) Fixed high standards If standards are fixed and high, we need to explain why speakers should so frequently find it reasonable to claim

to know things they are fully aware they don't know. (Remember, we are saving the possibility that speakers are unaware of their own ignorance for later.) One possible explanation is that they are trying not to mislead others who do not realize that the standards for knowledge are very high, and who would conclude from a denial of knowledge that the speaker was in a much poorer epistemic position than is actually the case. But this explanation applies only to the discourse of an enlightened sceptic talking to the unenlightened masses—surely a very special case. To explain the masses' own low-standards attributions of knowledge, an error theory would be needed.

Another possibility is that speakers are prone to hyperbole. Just as I might say ''I could eat a horse!'' instead of saying, more accurately, ''I could eat ten pancakes and a four-egg omelette'', so I might say that I *know* my car is in my driveway instead of saying merely that I have pretty good reason to believe this. If this kind of hyperbole were systematic and widespread, it might explain why we often claim to know things even when our grounds fall short of being conclusive (see Schaffer 2004). But I find the prospects of such an explanation dim. Hyperbole must be deliberate: if I really believed that I could eat a horse, I would not be exaggerating in saying that I could. However, ordinary speakers don't seem to regard their ordinary knowledge claims as exaggerations. Nor do they mark any distinction between what they literally know and what they only hyperbolically ''know''. When their knowledge claims are challenged, they don't say ''I was speaking hyperbolically'', the way I would if you replied to my horse-eating boast by saying, ''Not even a grizzly bear can consume an entire horse in one sitting.''

In defense of the hyperbole view, Jonathan Schaffer notes that hyperbole can be ''non-obvious,'' particularly when it is highly formulaic (2004: n. 3). We are so accustomed to the trope ''I'm dying of thirst'' that we no longer pause to consider its literal significance; instead, we jump directly to the intended meaning. Schaffer concludes that ''the fact that 'I know that I have hands' is not obviously hyperbolic is no objection''. But my point is not about obviousness. Even if speakers do not realize at first that in saying ''I'm dying of thirst'' they are speaking hyperbolically, they will immediately concede this when it is pointed out to them. ''Of course I'm not literally *dying*,'' they will say, ''and I never meant to suggest that I was.'' In contrast, those who say ''I know that I have hands'' will not, in general, concede that they were speaking

hyperbolically, even when confronted with sceptical counterpossibilities. No one reacts to the sceptic by saying, ''I never meant to suggest that I literally *knew* that I had hands!''

A third approach would appeal to the *inconvenience* of adding all the pedantic hedges and qualifications that would be needed to make our ordinary knowledge claims strictly true. As long as no one is likely to be misled, it may be more efficient to assert (falsely) that one knows that *p* than to assert (truly, but cumbersomely) that one knows that probably *p*, unless of course *q*; or that one has ruled out possibilities *X*, *Y*, and *Z*, but not *W*. For the same reason, one might say ''My tank holds 15 gallons'' when it really holds 14.5. As the potential misleadingness of unqualified and strictly false knowledge claims varies with the conversational context, so does our willingness to make them.

Like the hyperbole view, however, this approach fails to explain how we actually react when our ordinary knowledge claims are challenged. If I say ''My tank holds 15 gallons'' and someone calls me on it—''But the manual says it holds 14.5!''—I will say, ''I was speaking loosely: what I meant was that it holds *about* 15 gallons.'' But if I say ''I know that my car is in my driveway'' and someone calls me on it—''How can you rule out the possibility that it has been stolen?''—I will *not* say, ''I was speaking loosely: what I meant was that I know that my car is *most likely* in my driveway,'' or ''What I meant was that I know that my car is in my driveway, provided it has not been stolen or moved in some other abnormal way.'' In this respect I believe I am representative of ordinary speakers: otherwise, sceptical arguments would be greeted with shrugs, not surprise.

(b) Fixed low standards If standards are fixed and low, then what needs explaining is why we sometimes *deny* that people know, even when they clearly meet these standards. Patrick Rysiew has suggested that we sometimes deny that we know because we do not want to implicate that we can rule out certain salient but irrelevant counterpossibilities (2001: 492, 499). In asserting that *p*, one ordinarily represents oneself as knowing that *p*. If I make this implicit knowledge claim explicit by saying ''I *know* that my car is parked in my driveway'', my choice of words will be noticed. My hearers may well wonder why I did not simply say ''My car is parked in my driveway,'' and they may assume I meant to imply that I could rule out the conversationally salient possibility that my car had been stolen. Even

if I do not need to rule out this possibility in order to count as knowing, I do not want to be taken to be implying that I can rule it out. So, Rysiew argues, I have reason to disavow knowledge.

This is an ingenious explanation, but it fails on two counts. First, although worries about misleading implicatures may be good reasons to refrain from asserting something, they aren't good reasons to assert its *negation*. Before Cal has played any games, I will refrain from asserting (truly) that Cal has won all of its games so far this season, because my doing so would misleadingly imply that Cal has played at least one game already. But these considerations do not give me any reason to assert that Cal has *not* won all of its games so far this season. Similarly, even if Rysiew's story can explain why it would be rational for me to refrain from saying that I know, it cannot explain why I should say that I *don't* know.

Second, even if Rysiew's explanation worked in the first-person case, it could not be extended to third-person knowledge attributions. It is essential to Rysiew's explanation that the question arises, ''Why did the speaker say that he *knows* that *p* rather than just that *p*?'' The question does not arise in the same way in third-person cases. In saying that *p*, one does not ordinarily implicate that someone else, *X*, knows that *p*. So an assertion that *X* knows that *p* does not call attention to itself in the same way as a first-person knowledge ascription. Thus, Rysiew's explanation does not generalize to third-person knowledge attributions. But the phenomenon it seeks to explain does extend to third-person attributions. So the explanation fails.

3.1.2. Embedded ''know''

There is an easy pragmatic explanation for the infelicity of asserting ''I knew that *p* earlier, but now that standards have gone up, I don't know that *p*''.[5] In asserting that I knew that *p* earlier, I represent myself as knowing that I knew that *p*. But in representing myself as knowing that I knew that *p*, I also represent myself as knowing that *p*, since it is common knowledge that knowledge is factive. Thus there is a clash between what I commit myself to in asserting ''I don't know that

[5] For a slightly different version of this explanation, directed at third-person knowledge ascriptions rather than past-tensed ones, see Hawthorne 2004: 160.

p now'' and what I represent myself as knowing in asserting ''I knew that *p* earlier''.

But this explanation only takes us so far. It explains why we do not assert ''I don't know now that *p*, but I knew then that *p*''. But it does not explain our tendency to *deny* that we knew then that *p* (see DeRose 2002: §3). Nor does it explain why it is infelicitous to assert ''If *p* is true, then I knew that *p* before standards went up, though I don't know that *p* now'' (Hawthorne 2004: 166), or ''Joe doesn't know now that *p*, but he knew then that *p*'', or ''I know now that *p*, but I didn't know then that *p*'', when all that has changed are the standards. Here, it seems, a defender of circumstance-variable semantics must resort to an error theory.

3.1.3. Truth ascriptions and retraction

It might be suggested that the *inconvenience* of reformulating knowledge claims in a way that reflects their dependence on past standards sometimes makes it reasonable to treat them as if they had been made in light of current standards—even if this means saying that they were false when we know that they were true. The differences in usage between ''know'' and ordinary indexicals might then be attributed to the comparative ease of reformulating claims made using ordinary indexicals when the relevant contextual factors have changed. If I say ''I am tired now'' at 3.30 p.m. today, others can easily re-express the content of my claim tomorrow by using the sentence ''he was tired at 3.30 p.m. yesterday''. But when it comes to ''know''—supposing that ''know'' is context-sensitive—things are messier. How can we re-express a knowledge claim made in one context in another, where standards are different? I might say something like this: ''I asserted that I knew, by the relatively low standards for knowing in place at the time, that my car was in my driveway.'' Or perhaps: ''I said something that is true just in case I met the standards in place at the time for knowing that my car was in my driveway.'' But these reformulations are cumbersome and not very informative.[6] Even if they are correct, it may seldom be

[6] More informative reformulations would require a way of specifying epistemic standards directly, rather than as the standards in play at such-and-such a context. We do not consider speakers masters of the indexicals ''here'' and ''now'' unless they are in command of coordinate systems for specifying places and times independently of utterance events (''in Berkeley, California'', ''at 3.30 p.m. GMT on 14 October 2003''), which they can use

worth the trouble to use them; in many cases, it may be more efficient simply to withdraw the earlier knowledge claim. In this way, a contextualist might attempt to explain away the data about truth ascriptions and retraction that suggest a use-invariant semantics for ''know''.

But if this is the explanation of our retraction behavior, there ought to be *some* cases in which the disadvantages of retracting outweigh the inconvenience of reformulating. Suppose Sam is in the courtroom:

Judge: Did you know on December 10 that your car was in your driveway?
Sam: Yes, your honor. I knew this.
Judge: Were you in a position to rule out the possibility that your car had been stolen?
Sam: No, I wasn't.
Judge: So you didn't know that your car was in the driveway, did you?
Sam: No, I suppose I didn't, your honor.
Judge: But you just said you did. Didn't you swear an oath to tell the whole truth, and nothing but the truth?

However inconvenient it would be for Sam to reply,

My claim was that on December 10 I knew, by the standards for knowledge that were in play before you mentioned car thieves, that my car was in my driveway. That was true, your honor, so I did not speak falsely,

it would surely be more inconvenient for him to be charged with perjury. Nonetheless, I think that Sam, if he is like most ordinary speakers, will concede that his previous assertion was false and promise to be more careful in his future answers. This suggests that the calculus of inconvenience alone cannot explain why speakers tend to abandon their earlier knowledge claims when they are shown to be false in light of present standards.

3.2. Error Theories

A sincere speaker who wants to speak the literal truth and avoid literal falsity may fail to do so if she has false beliefs, either about the facts or about the literal meanings of the words she uses.[7] If I believe (as I once

to reiterate claims made using these indexicals in other contexts. Ordinary speakers possess no comparable coordinate system for specifying epistemic standards.

[7] Although I doubt that a clean distinction between semantic and substantive error can be made, a rough and ready distinction will suffice for our purposes here. Note that it is

did) that ''gravy'' is the name of a vitamin-deficiency disease, I will refrain from asserting ''I like gravy'', even if I do like meaty sauce. And if I believe that whales are fish, I may assert ''Whales are fish'', even though this is false. Before we make any inferences from facts about ordinary use to truth conditions, then, we must rule out the possibility that ordinary speakers are systematically mistaken in certain ways. As before, I'll consider our three facts about use in turn.

3.2.1. Variability of standards

To explain the variability data, moderate strict invariantists must argue that speakers often underestimate their success in meeting the standards for knowledge and as a result disavow knowledge that they actually possess. Sceptical invariantists, by contrast, must argue that speakers systematically *over*estimate their success in meeting the standards for knowledge and as a result claim to know when in fact they do not.

The sceptical version of the error theory is sometimes rejected on the grounds that it rules expressions of paradigm cases of knowledge, like ''I know that I have hands'' false. But the paradigm case argument is not a good argument. A supposed paradigm case of *F*-ness can turn out not to be an *F* at all. Whales turned out not to be fish; glass turned out not to be a solid. This might even happen on a large scale. Suppose that in 1750, all the emeralds on earth had been replaced by synthetic duplicates indistinguishable by the technology of the time. Then *none* of the extant ''paradigm cases'' of emeralds would have been emeralds. The sceptic's claim that ordinary speakers are mistaken in nearly all of their knowledge claims cannot be rejected out of hand.

Nonetheless, it is fair to ask the sceptical invariantist for an *explanation* of the widespread and uniform error she attributes to speakers. Why do speakers so quickly revert to making everyday knowledge claims even after they have been led through sceptical arguments (cf. Hawthorne 2004: 131)? Human beings are educable; the fact that the lesson does not stick deserves special explanation. Moreover, the

ignorance about literal meaning that is at stake here, not ignorance about speaker's meaning. On many accounts of speaker's meaning, it is implausible to suppose that a speaker could be ignorant of what *she* means. Nonetheless, she can very well be ignorant of what *her words* mean, or of what she has literally *said*. See Rysiew 2001: 483, commenting on §IV of Schiffer 1996.

sceptic must explain how ''know'' comes to have the exacting meaning it has, despite the fact that looser use is the norm. (It would be difficult to argue that ''decimate'' still means just ''to kill one in every ten of'', when it is now routinely used for cases of larger scale destruction.) Here the sceptic will have to put great weight on certain widely accepted generalizations about knowledge (such as closure principles) that can be exploited in sceptical arguments. But it is not clear why these generalizations should have a better claim to be meaning-constituting than the ''paradigm cases'' the sceptic rejects. At the very least, the sceptic owes us a fancy story here.

The moderate strict invariantist does not face this problem, since she takes many of our ordinary knowledge claims to be true. But she must explain why speakers find the premises exploited in sceptical arguments so compelling, despite the implausibility of the conclusions to which they lead. If these premises are false, why do speakers not come to see their falsity and stop feeling the pull of sceptical arguments? Presumably a moderate strict invariantist will say that I can sometimes know that my car is in the driveway, even though I have been gone for fifteen minutes and cannot absolutely rule out the possibility of car theft in the interim. Why, then, does the closure-exploiting argument that I cannot know this seem so compelling? These are deep and difficult questions, to be sure. My point here is that, until she answers them satisfactorily, the moderate strict invariantist cannot explain away the apparent variability of standards in our knowledge attributions.

There is a further problem with both kinds of error theory, recently emphasized by Keith DeRose (2002) and John Hawthorne (2004: 132–5). Ordinary speakers accept many generalizations linking knowledge with other concepts. For example, one ought not assert something unless one knows it, one ought to decide what to do by reasoning from what one knows, and so on. The sceptical invariantist will have to hold that these generalizations, too, are in error, or else take the hard line that the vast majority of our assertions are improper and our decisions and actions irresponsible. The moderate strict invariantist will have trouble here, too, though less spectacularly, because in some situations (where much is at stake) we seem to require a very high standard of evidence before we will act on or assert a proposition. She must either say that our scruples here are unwarranted or reject the generalizations linking knowledge with assertion and action.

3.2.2. Embedded ''know''

According to sensitive invariantism, the fact that speakers use the same epistemic standards in evaluating embedded and non-embedded instances of ''know'' reflects some kind of systematic error. But what kind? There are two possibilities. First, speakers might take the standards required to count as ''knowing'' to be fixed, or to be determined entirely by the context of use. Alternatively, instead of being mistaken about the semantics of ''know'', speakers might systematically misjudge the standards in play at different circumstances of evaluation.

There is something a bit perverse about the first explanatory strategy. One of the best arguments *in favor of* a circumstance-variable, use-invariant semantics for ''know'' is that it promises to explain both the variability data and the data about truth ascriptions and retraction. But it cannot explain these data unless it plays some role in guiding speakers' linguistic behavior. Thus, if we explain away the data about embedded occurrences by arguing that speakers implicitly take ''know'' to be circumstance-invariant and use it accordingly, we undercut one of the best arguments in favor of sensitive invariantism.

Better, then, to argue that speakers systematically misjudge the standards relevant at alternative circumstances of evaluation. Along these lines, John Hawthorne argues that we tend to ''project'' the standards currently in play to other putative knowers, times, and circumstances:

> we do have some tendency to suppose that, as more and more possibilities of error become salient to us, we are reaching an ever more enlightened perspective. Thus when we consider someone who is not alive to these possibilities, we have a tendency to let our (putatively) more enlightened perspective trump his. This tendency, when left unchecked, leads to scepticism. (Hawthorne 2004: 164–5)

This kind of projection is not unprecedented: it is well known that those for whom a recent disaster is salient will overestimate risks in past, future, and counterfactual situations. In much the same way, Hawthorne urges (162–3):

> Once we have gotten ourselves into the frame of mind of thinking 'I do not in fact know whether or not I'll be able to afford the Safari,' as we frequently do when we use parity reasoning, we are not only unwilling to say 'However I used to know that;' we are positively willing to say 'I never did know that.'

This strategy is worth pursuing, but we should remind ourselves how heavy an explanatory burden it must bear. It *always* seems wrong to say that Joe knew before, but doesn't know now, when the only thing that has changed are the relevant standards. Projection might explain occasional or even frequent mistakes, but I doubt it can account for our universal unwillingness to shift standards across circumstances of evaluation.

Even if the projection strategy works, it is a double-edged sword. If it succeeds in explaining why we evaluate *embedded* occurrences of ''know'' in light of present standards, it should also explain why we evaluate occurrences of ''know'' at other *contexts of use* in light of present standards. That is, it should explain the data about truth ascriptions and retraction. Indeed, Hawthorne suggests as much himself, when he adds, immediately after the second passage quoted above: ''And, if pressed, we are willing, moreover, to say that 'I was mistaken in thinking that I did know that' '' (163). The problem is that one of the best arguments for an invariantist semantics for ''know'' is that it explains the data about truth ascriptions and retraction. If those data are explained instead by the story about projection, then the argument for preferring sensitive invariantism to contextualism is significantly weakened.

3.2.3. Truth ascriptions and retraction

The data about truth ascriptions and retraction are most straightforwardly explained by a use-invariant semantics for ''know''. A contextualist must explain these data in some other way. We have ruled out a pragmatic explanation (§3.1.3), so it seems that a contextualist must appeal to an error theory here. Many contextualists are explicit about this: for example, Stewart Cohen says that ''We *mistakenly* think that knowledge ascriptions we make in everyday contexts conflict with the skeptical judgements we make in stricter contexts'' (2001: 89; emphasis added).[8]

[8] Cohen argues that this error theory is innocuous, on the grounds that speakers make similar mistakes with gradable adjectives like ''flat'' (pp. 90–1). Richard 2004 concedes Cohen's analogy and rejects his error theories, plumping for a relativist treatment of both ''flat'' and ''know''. For my part, I am not convinced of the analogy: I think that a pragmatic explanation of our retraction and reporting behavior is much more plausible for ''flat'' than for ''know''. When standards change so that the surface imperfections on

As before, there are two options: the contextualist can suppose either that ordinary speakers are wrong about the semantics of ''know''—treating it as use-invariant when it is not—or that they make systematic errors about what standards are in play in contexts other than their own. The problem is that both forms of error theory threaten to undermine the positive case for contextualism. This is especially clear if the error is semantic in character. If ordinary speakers have a faulty grasp of the *meaning* of ''know'', then we cannot confidently appeal to variability in the standards they require someone to meet in order to count as ''knowing'' as support for a theory about the meaning of ''know''. Yet these data are the primary evidence in favor of contextualism.

What about the second option? It is undeniable that speakers often misjudge features of other contexts of use than their own, but if we are to explain the data, the error we posit must be *systematic*. We must explain why speakers *never* allow their previous day's assertion of ''I know that *p*'' to stand as true while asserting ''I did not know that *p* yesterday''. I doubt that our tendencies to project features of our present situations onto other situations are nearly strong or uniform enough to explain away the uniform data about truth ascriptions and retraction.

The ''double-edged sword'' point applies here, too. If the projection story works with contexts of use, it ought to work with circumstances of evaluation, too. So if it explains the data about truth ascriptions and retraction, it ought to explain the data about embedded occurrences of ''know'' as well. This would significantly weaken the contextualist's case against sensitive invariantism.

As should now be clear, a *general* problem with positing speaker error to explain away facts about use is that such explanations tend to undermine the evidential basis for the semantic theories they are intended to support. All of these semantic theories are justified indirectly on the basis of facts about speakers' use of sentences, and the more error we attribute to speakers, the less we can conclude from these facts. We have seen that the cost of defending sensitive invariantism in this way is that the case against contextualism is severely weakened, and conversely

pancakes count as ''bumps'' and ''holes'', a speaker might retract an earlier assertion of ''pancakes are flat'', but only to avoid pedantry, not because she thinks she's really contradicted herself. If enough were at stake, she would no doubt find an appropriate way to reiterate her earlier claim. (Contrast what is alleged about ''know'' in §3.1.3, above.)

that the cost of defending contextualism in this way is that the case against sensitive invariantism is compromised. It is possible that an error theory can be made to work—perhaps in conjunction with pragmatic explanations—but the prospects do not look good.

3.3. Eliminativism

So far we have looked at ways of showing that one of the standard views is in fact consistent with all of the facts about use we considered in §2. An alternative response would be to concede that no single account of the semantics of ''know'' accounts for all of these facts (see Schiffer 1996). Perhaps our talk of ''knowledge'' confuses several distinct notions, in much the same way that prescientific talk of ''warmer than'' confused *having a higher temperature than, having more heat energy than*, and *exchanging heat at a higher rate than*.[9] In that case there may be no fully coherent way to assign truth conditions to our knowledge-attributing sentences. The rational course of action would be to reform our thought and talk by introducing new, unconfused terms of epistemic assessment.

At the risk of use–mention confusion, we might call this approach ''eliminativism about knowledge''. Like other eliminativisms, it is radical and should not be accepted unless there is no other good alternative.

3.4. Expanding the Field of Options

Let us sum up our conclusions so far. Together, our three facts about use suggest that an adequate semantics for ''know'' must be sensitive to changing epistemic standards, but that it cannot be either use-variable or circumstance-variable. That rules out all three standard views: strict invariantism because it is not sensitive to changing epistemic standards at all, sensitive invariantism because it is circumstance-variable, and contextualism because it is use-variable. We might make room for one of these views by arguing that one of our three facts about use is a poor guide to truth conditions, but attempts to do this either pragmatically or by positing systematic error on the part of ordinary speakers have so far been unpersuasive. If there is no other option, then, it seems we are left with eliminativism.

[9] For a discussion of this example, see Churchland 1979: ch. 2.

But how *could* there be another option? How could there be a semantics for ''know'' that was use-invariant and circumstance-invariant, but still in some way sensitive to changing epistemic standards? What we would need is another dimension of variability. In the next section, I am going to open up room for just such a thing. This will make possible a semantics for ''know'' that neatly explains *all three* facts about use.

4. A RELATIVIST SEMANTICS FOR ''KNOW''

Here is my proposal. The epistemic standards relevant to determining the extension of ''know'' are not those in play at the context of use or those in play at the circumstance of evaluation, but those in play at the *context of assessment*.

4.1. Assessment Sensitivity

The notion of a context of assessment may be unfamiliar, but it is readily intelligible. Just as a context of use is a situation in which a sentence might be *used*, so a context of assessment is a situation in which a (past, present, or future, actual or merely possible) use of a sentence might be *assessed* for truth or falsity. I do not think that there should be any worries about the very *idea* of a context of assessment; even an arch anti-relativist ought to be able to accept it.

What is controversial is the suggestion that we relativize sentence truth not just to a context of use, but to a context of assessment as well. This is certainly a departure from semantic orthodoxy, and I will defend it shortly.[10] Here I want to focus on what we can do with it. By making sentence truth doubly context-relative, we open up a new way in which sentences can be context-sensitive. A sentence is context-sensitive in the usual way, or *use-sensitive*, if its truth value varies with the context of use (keeping the context of assessment fixed). A sentence is context-sensitive in the new way, or *assessment-sensitive*, if its truth value varies with the context of assessment (keeping the context of use fixed). Similarly, a subsentential expression is use-sensitive if it is partially

[10] The relativization of truth to a context of assessment should not be confused with the relativization of truth to a ''point of evaluation'' (e.g. a tuple of time, world, and variable assignment) that is standard in model-theoretic semantics. A point of evaluation is not a context, but a sequence of parameters that can be ''shifted'' by operators. For more on the difference, see Lewis 1980 and MacFarlane 2003: §V.

responsible for the use sensitivity of (at least some) sentences containing it, and assessment-sensitive if it is partially responsible for the assessment sensitivity of (at least some) sentences containing it.

My proposal is that ''know'' is sensitive to the epistemic standards in play at the context of assessment. It is a kind of contextualism, then, but not at all the usual kind. To avoid confusion, I will call it ''relativism'', reserving the term ''contextualism'' for the view that ''know'' is sensitive to the epistemic standards in play at the context of *use* (see Figure 8.2). Call a semantics for ''know'' *assessment-variable* just in case it allows the epistemic standard relevant for determining the extension of ''know'' to vary with the context of assessment, and *assessment-invariant* otherwise. If ''know'' is assessment-sensitive, then its semantics can be assessment-variable while being use- and circumstance-invariant, and in this way we can neatly explain all three facts about use:

4.1.1. Variability of standards

Why is it that I'll happily assert ''Joe knows that his car is parked in his driveway'' when standards are low, and ''Joe doesn't know that his car is

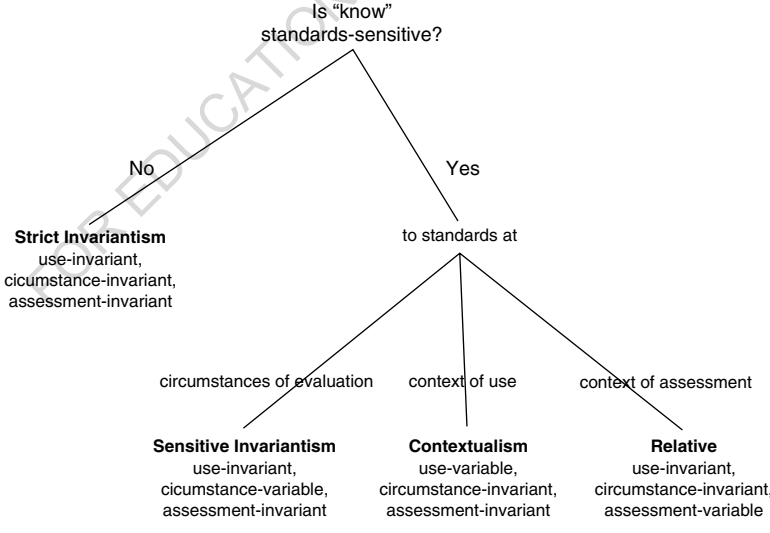

Figure 8.2. Expanded taxonomy of positions on the semantics of ''know''

parked in his driveway'' when standards are high? The relativist semantics affords a simple explanation: the former sentence is true as used and assessed in a context where standards are low, and the latter is true as used and assessed in a context where standards are high. Because I can properly assess each sentence as true at the context in which I utter it, there is no need to appeal to pragmatic explanations or error theories to explain the variability data. (Note that in the special case where the context of use and the context of assessment coincide, the relativist semantics yields exactly the same truth-value assignments as the standard contextualist semantics. So the relativist is in just as good a position to explain the variability data as the contextualist. This is significant, because one of the primary selling points of the contextualist account is its ability to explain these data.)

4.1.2. Embedded ''know''

Why is it that we don't say things like ''Before the standards went up, Harry knew that his car was in the driveway, but now he doesn't know this''? Or: ''Harry doesn't know that his car is in the driveway, but he would know this if the possibility of car theft weren't relevant''? The relativist semantics has a straightforward explanation: the semantics of ''know'' is not circumstance-variable. This is the same explanation that contextualism and strict invariantism offer.

4.1.3. Truth ascriptions and retraction

Why is it that when standards go up, leading us to assert ''Joe doesn't (and didn't) know that his car is in his driveway'', we expect Joe to retract his earlier assertion of ''I know that my car is in my driveway'', and to concede that what he asserted was false? That is, why do we tend to use the standards appropriate to the present context in assessing past utterances? Where contextualism is forced to appeal to an error theory, the relativist semantics offers an easy, semantic explanation. Namely: the present standards *are* the appropriate standards to use in assessing past assertions, even ones that were made when very different epistemic standards were in play. According to the relativist, knowledge claims are always properly assessed in light of the standards in play at the assessor's current context.[11]

[11] A full discussion of the truth ascription data would require giving a semantics (at least a naive semantics) for the monadic object-language predicate ''true''. I will not pause

4.2. Expressive Relativism and Propositional Relativism

Contextualists typically hold that ''know'' expresses different relations at different contexts of use, and that this indexicality is the source of its use sensitivity. Must the relativist, then, hold that ''know'' expresses different relations relative to different contexts of assessment? This would be an odd view. It would require us to give up the idea that ''knowledge'' attributers are making determinate claims. Assessors at different contexts could disagree about what was said, and they could all be right! According to this *expressive relativism*, there would be no non-relative fact of the matter about what proposition was expressed by the sentence used, at its context of use.[12]

I go back and forth about the coherence of expressive relativism. It sometimes seems to me that, with a bit of imagination, we can make sense of it. Even if we can make sense of it, however, it does not seem to be a very attractive view. It would require significant changes in orthodox theories of meaning. For example, we could no longer say, with Stalnaker 1978, that the effect of assertion is to add the proposition asserted to a ''common ground'' of presupposed propositions, for there may be no common fact of the matter about which proposition *was* asserted.[13] Moreover, although expressive relativism might help us understand *speech acts* made using ''know'', it would leave it rather mysterious what it is to *believe* that Joe knows that his car is in his driveway.

Fortunately, assessment sensitivity can be had without expressive relativism. Call a sentence *use-indexical* if it expresses different propositions at different contexts of use (keeping the context of assessment fixed), and *assessment-indexical* if it expresses different propositions relative to different contexts of assessment (keeping the context of use fixed). Call a subsentential expression *use-* or *assessment-indexical* if it is at least partially responsible for the use or assessment indexicality of

to do that here. It turns out, not surprisingly, that in a language containing assessment-sensitive expressions, object-language ''true'' must also be assessment-sensitive.

[12] Sometimes a single use of a sentence may express multiple propositions, as when a teacher says to a class of thirty: ''Of the three people sitting nearest to you, only two are likely to finish this class.'' I take it that in this case the teacher has asserted thirty singular propositions, not a single general one. This does not amount to expressive relativism, since all parties can agree about which propositions were expressed.

[13] This is pointed out by Egan *et al*. forthcoming, who also give other arguments against what they call ''content relativism''.

sentences containing it. According to expressive relativism, ''know'' is assessment-indexical, and that is why it is assessment-sensitive. But in fact a sentence or subsentential expression can be assessment-sensitive without being assessment-indexical.

Indeed, although this is often overlooked, a sentence can be *use*-sensitive without being use-indexical. Here is an example: ''The number of AIDS babies born in the United States in 2003 is greater than 1000.'' This sentence expresses the same proposition at every context of use, so it is not use-indexical. But it *is* use-sensitive, because its truth value varies with the world of the context of use. Uttered in a world in which there were no AIDS babies in 2003, it would express a falsehood; uttered in the actual world, it expresses a truth.[14]

To see how a sentence can be use-sensitive without being use-indexical, we need to be explicit about the relation between sentence truth at a context and proposition truth at a circumstance of evaluation. (For simplicity, let us forget for a moment about contexts of assessment.)

> *Sentence Truth and Proposition Truth I*: A sentence *S* is true at a context of use *C* just in case for some proposition *p*,
>
> (1) *S* expresses *p* at *C*, and
> (2) *p* is true when evaluated at the circumstance determined by *C*.[15]

Notice that the context of use plays two distinct roles here: (1) it determines what proposition is expressed by the sentence, and (2) it determines how that proposition is to be evaluated to yield a truth value for the sentence in context. Indexicality produces use sensitivity via role (1), while contingency produces use sensitivity via role (2).

[14] David Lewis put this point by saying that ''[c]ontingency is a kind of indexicality'' (1998: 25)—using ''indexicality'' for what I call ''use sensitivity''. Other writers use ''context sensitivity'' for what I call ''use indexicality''. I think it is useful to have distinct terms for both notions.

[15] This definition is a close paraphrase of Kaplan 1989: 522 (cf. 547). As it stands, it is not sufficiently general, for in some frameworks the context of use will not always determine a unique circumstance of evaluation. For example, in indeterministic frameworks allowing overlapping worlds or ''histories'', the context of use will not pick out a single history (see Belnap and Green 1994; MacFarlane 2003). For a more generally applicable definition, we could replace (2) with: ''*p* is true when evaluated at all circumstances of evaluation compatible with *C*'' (e.g. at all moment/history pairs in which the moment is the moment of *C* and the history contains the moment of *C*). We can ignore this complication for present purposes.

The situation is much the same when we relativize sentence truth to contexts of assessment as well as contexts of use:

> *Sentence Truth and Proposition Truth II*: A sentence S is true at a context of use C_U and context of assessment C_A just in case for some proposition p,
>
> (1) S expresses p at C_U and C_A, and
> (2) p is true when evaluated at the circumstance determined by C_U and C_A.

As before, there are two distinct roles for contexts to play: (1) determining which proposition is expressed and (2) determining how that proposition is to be evaluated to yield a truth value for the sentence in context. Accordingly, there are two ways in which a sentence can be assessment-sensitive: it can be assessment-indexical, or the context of assessment can play a substantive role in determining the circumstance relative to which the proposition it expresses is to be evaluated. We can state the point more simply if we define proposition truth relative to contexts in the natural way:

> *Contexts-Relative Proposition Truth*: A proposition p is true at a context of use C_U and context of assessment C_A just in case p is true when evaluated at the circumstance determined by C_U and C_A.

Call a *proposition* assessment-sensitive just in case its truth varies with the context of assessment (keeping the context of use fixed). Then a sentence can be assessment-sensitive either by being assessment-indexical or by expressing an assessment-sensitive proposition. In the former case, we have expressive relativism; in the latter case, *propositional relativism*.

It might be thought that propositional relativism would require even more radical departures from orthodox semantics than expressive relativism. But that is not so. Granted, the form of propositional relativism I am advocating does require that the circumstances of evaluation to which propositional truth is relativized include an epistemic standards parameter in addition to a world parameter. But quite a few non-relativists have countenanced parameters of circumstances of evaluation besides the world parameter, so this hardly counts as a radical departure from standard semantic assumptions. For example, Kaplan's (1989) circumstances of evaluation include a time parameter. On Kaplan's

account, a tensed sentence like ''Socrates is sitting'' expresses the same proposition at every context of use; it nonetheless has different truth values at different contexts of use, because different contexts determine different circumstances (times and worlds) with respect to which this proposition is to be evaluated. King (2003) contemplates relativizing propositional truth to both worlds and standards of precision. On the view he considers (without endorsing or rejecting it), ''France is hexagonal'' expresses the same proposition at every context of use; it is true at a context of use just in case this proposition is true when evaluated with respect to the world of the context of use and the standards of precision in play at the context of use. Despite their appeal to parameters of circumstances of evaluation besides worlds, neither Kaplan nor King is a propositional relativist, because neither countenances assessment-sensitive propositions.

Nor is there anything particularly novel about having an *epistemic standards* parameter in the circumstances of evaluation. The non-relativist form of contextualism about ''knows'' defended in Kompa 2002 requires one too. On Kompa's view, ''know'' expresses the same relation at every context of use, but this relation is ''unspecific'', in the sense that ''what counts as having the property'' can vary with ''the context at hand'' (88)—that is, the context of use. Although Kompa does not develop her view in formal detail, it is hard to see how she could do so without adding an epistemic standards parameter to the circumstances of evaluation. The intension of the relation expressed by ''know'' would then be a function from worlds, times, and epistemic standards to extensions. This relation would be ''unspecific'' in the sense that there would be no answer to the question whether a particular person and fact fall into its extension at a time and world: only when an epistemic standard was specified would it have a definite extension. Kompa could then define context-relative sentence truth as in *Sentence Truth and Proposition Truth II*, above, taking the circumstances determined by a context of use C_U (and context of assessment C_A) to be $\langle w, t, e \rangle$, where $w =$ the world of C_U, $t =$ the time of C_U, and $e =$ the epistemic standards in play at C_U. ''Know'' would thus turn out to be use-sensitive but not use-indexical, just like tense on Kaplan's view, vague expressions on the view explored by King, and contingent eternal sentences on just about everyone's view.

The only difference between the relativist view I am advocating and Kompa's non-relativist, non-indexicalist form of contextualism is that I take the circumstances determined by a context of use C_U and context of

assessment C_A to be the ordered pair $\langle w, t, e \rangle$, where w = the world of C_U, t = the time of C_U, and e = the epistemic standards in play at C_A (not C_U).[16] In every other respect I can agree with Kompa. We can agree that propositional truth must be relativized to an epistemic standards parameter (in addition to a world and perhaps a time parameter). We can agree about which propositions are expressed by which sentences at which contexts of use. We can both accept the schematic principle 'Sentence Truth and Proposition Truth II' (provided Kompa does not mind the relativization to a context of assessment, which plays no substantive role in her account but also does no harm). We will of course disagree about the extension of the relation ''S is true at context of use C_U and context of assessment C_A'', because we disagree about how this schematic principle is to be filled in. (That is, we disagree about which circumstances of evaluation are ''determined'' by which contexts of use and assessment.) But this disagreement does not concern the theory of propositions. I conclude that if propositional relativism is objectionable, it is not because it requires radical revision to our existing theories of propositions.

5. MAKING SENSE OF RELATIVE TRUTH

I anticipate two objections to my proposal. First, that it is *ad hoc*. Good scientific practice dictates that we make central modifications to our theories only when they have great and wide-ranging explanatory value. Surely it is not a good idea to make structural changes to our semantic framework just to accommodate knowledge attributions. Second, that it is incoherent. It is one thing to talk of propositions or sentences being true with respect to one context of assessment, and not with respect to another. It is quite another thing to make sense of that talk, and there are reasons for doubting that any sense *can* be made of it.

5.1. Ad hoc?

To the first objection I have two replies. First, I believe that assessment sensitivity is not limited to knowledge-attributing sentences. I believe it

[16] I include the time parameter for illustrative purposes only; nothing hangs on its presence. King (2003) may be right that the best treatment of tense does not call for a time parameter in circumstances of evaluation, in which case I am happy to remove it.

is also the key to adequate semantic treatments of future contingents, epistemic modals, accommodation (in the sense of Lewis 1979 and 1980), terms like ''delicious'', and perhaps much else.[17] So the modification I propose is not tailor-made for a single use, but has much wider application.

Second, the structural changes that are required are less radical than one might think. As I have argued, propositional relativism requires minimal changes to existing semantic frameworks. These changes are conservative. They allow us to *describe* assessment sensitivity, but they leave open the possibility that there is no assessment sensitivity in any natural language. Existing accounts of the semantics of expressions that are not assessment-sensitive can be carried over essentially unchanged. (In these cases, the relativization to contexts of assessment will be an idle wheel, but a harmless one, because truth will not vary with the context of assessment.) Thus although one might object to the claim that ''know'' is assessment-sensitive, it is hard to see on what grounds one might object to the framework that makes it possible—unless one thinks that assessment-relative truth is simply incoherent.[18]

5.2. *Incoherent?*

What on earth can it *mean* to say that an assertion is true as assessed by me now, but false as assessed by me later; or true as assessed by me, but false as assessed by you? This is not the kind of question that can be answered by *defining* ''true at a context of use C_U and context of assessment C_A''. Indeed, I have already done that, in §4.2, for both sentences and propositions.[19] But our definitions leave us more or less in the position of Martian anthropologists, who know what counts as a

[17] I discuss future contingents in MacFarlane 2003 and the other issues in a book manuscript, in progress. For independent arguments for a relativist treatment of epistemic modals, see Egan *et al.* (forthcoming). For a relativist treatment of accommodation (somewhat different from mine), see Richard 2004. For a relativism motivated by ''delicious'' and the like, see Kölbel 2002.

[18] One might also worry that the extra degree of freedom I offer in constructing semantic theories does not come with sufficient counterbalancing constraints. The connections between contexts-relative truth and norms for assertion which I propose in the next section are meant to address this concern.

[19] These definitions are of course schematic, but when the semantic details are filled in, they will determine an extension for ''true at context of use C_U and context of assessment C_A.''

winning position in chess and other games, but fail to grasp the *significance* of winning (cf. Dummett 1959). We don't know what to *do* with a claim that a sentence (or proposition) is true relative to a context of use C_U and context of assessment C_A. And until we know that, we do not really understand relative-truth talk.

The charge of incoherence arises because a standard story about the significance of ''true at a context of use C_U'' cannot be extended to ''true at a context of use C_U and context of assessment C_A''. According to this story, truth is the internal aim of assertion. Of course, people may have all kinds of goals in making assertions—influencing others, showing off, giving directions, offering reassurance—and these goals may sometimes be better served by speaking falsely than by speaking the truth. But there is a sense in which a false assertion is always incorrect qua assertion, even if it succeeds in promoting these other goals. It may be useful to lie, but once your assertion has been shown to be false, you must withdraw it as mistaken. Dummett argues that the concept of truth gets its significance from this normative connection to the practice of assertion: just as it is part of the concept of winning a game that a player aims to win, so ''it is part of the concept of truth that we aim at making true statements'' (1978: 2). But if our primary grip on the notion of truth comes from our understanding of it as the internal aim of assertion, then the idea that truth might be relativized to a context of assessment just looks incoherent.[20] It does not make *sense* to aim to assert a proposition that is true at the context of use and the context of assessment, because there is no such thing as *the* context of assessment: each assertion can be assessed from indefinitely many distinct contexts.

At this point relativists typically say that the aim of assertion is to assert something that is true relative to the context of use and the asserter's own current context of assessment, which will of course be identical with the context of use (see Kölbel 2002: 125; Egan *et al.* forthcoming: 29). But this only gives a significance to ''true at C_U, C_A'' for the special case where $C_U = C_A$. The relativist has not told us what to do with ''true at C_U, C_A'' where C_U and C_A are distinct. As a result, the anti-relativist might justly charge that the relativist's

[20] For a development of this argument, drawing on Evans 1985, see Percival 1994: 196–8.

''true at *C*, *C*'' is just a notational variant of her own ''true at *C*'', and that ''true at C_U, C_A'' has not yet been given a sense when $C_U \neq C_A$.

In my view, the relativist should instead reject the whole idea of understanding truth as ''the aim of assertion''. This idea is pretty obscure anyway. Even if truth is *an* internal normative aim of assertion, it is certainly not the only such aim: it is also part of the practice of assertion that we strive to say what is relevant to the conversation at hand, and to say things that are appropriately justified (or on some accounts, known). Indeed, in his 1972 Postscript to ''Truth'', Dummett emphasizes that his talk of truth as the aim of assertion was intended as a placeholder for a more complex story about the role truth plays in our practice of assertion: ''What has to be added to a truth-definition for the sentences of a language, if the notion of truth is to be explained, is a description of the linguistic activity of making assertions; and this is a task of enormous complexity'' (Dummett 1978: 20).

Having given up the ''aim of assertion'' idea, what else might we say about the role truth plays in our practice of assertion? One plausible and widely accepted idea is that an assertion is a *commitment* to the truth of what is asserted (see e.g. Searle 1979: 12). To make an assertion—even an insincere or otherwise defective one—is, *inter alia*, to commit oneself to the truth of the proposition asserted (relative to its context of use).[21] But what is it to commit oneself to the truth of a proposition? How does one honor or violate such a commitment? Some philosophers seem to find ''commitment to truth'' intelligible without further analysis, but in my view, ''commitment to [noun phrase]'' is intelligible only when it can be glossed in terms of commitment to *do* something. For example, we can make sense of ''being committed to Al Gore'', but only as meaning something like ''being committed to *working for* (or perhaps *supporting*) Al Gore''. When no obvious agentive complement presents itself, we can't make any sense of the deontic construction at all. What would it mean, for example, to be committed to the color of the sky; or to the texture of a damp rose petal?

So, in committing myself to the truth of a proposition at a context of use, what exactly am I commiting myself to *doing* (or refraining from doing)? Well, suppose I assert ''Jake is in Boston''. If you ask ''How do

[21] *Inter alia*, because presumably asserting a proposition involves more than simply committing oneself to its truth. Plausibly, the commitment must be undertaken publicly, by means of an overt utterance; perhaps there are other conditions as well.

you know?'' or challenge my claim more directly, by giving reasons for thinking it false, then it seems to me that I have an *obligation* to respond, by giving adequate reasons for thinking that my claim was true, or perhaps by deferring to the person who told me. If I can't discharge this obligation in a way that meets the challenge, I must ''uncommit myself'' by retracting my assertion. If I neither withdraw the assertion nor reply to the challenge, I am shirking an obligation I incur not qua moral agent or friend or member of polite society, but simply qua asserter.

These observations suggest an answer to our question. What I have committed myself to doing, in asserting that Jake is in Boston, is vindicating my claim when it is challenged.[22] There may be no specific sanction for failing to follow through on this commitment. But if I fail too blatantly or too frequently, others may stop treating me as a being that is capable of undertaking this kind of commitment. They may still take my utterances as expressions of my beliefs, as we take a dog's excited tail wagging as an expression of its psychological state. They may even regard my utterances, if found to be reliable, as useful bits of information. But they will be treating me as a measuring instrument, not as an asserter. They will not take me to be *committing myself* to the truth of anything.

If this is right, then we should understand the ''commitment to truth'' incurred by an assertion as follows:

> *Assertoric Commitment*: In asserting that p at a context C_U, one commits oneself to providing adequate grounds for the truth of p (relative to C_U), in response to any appropriate challenge, or (when appropriate) to deferring this responsibility to another asserter on whose testimony one is relying. One can be released from this commitment only by withdrawing the assertion.[23]

[22] For this way of looking at assertoric commitment as a ''conditional task responsibility'' to vindicate a claim when it is challenged, see Brandom 1983 and ch. 3 of Brandom 1994. I do not develop the idea in quite the same way as Brandom, but I am much indebted to his work.

[23] Several philosophers have suggested to me that this account overgeneralizes, taking the norms of the seminar room to apply to assertions in general. It may be that ordinary asserters recognize no general obligation to justify their claims in the face of reasoned challenges (though it would not follow that they are not *bound by* such a norm). But even these sceptics ought to be able to accept a weaker norm requiring withdrawal of assertions that have been shown to be untrue (relative to the context of use). This would be enough for my purposes here.

The principle is schematic along many dimensions: to make it less schematic, one would have to say something about what kinds of challenges count as ''appropriate'', what grounds count as ''adequate'' responses to what kinds of challenges, and when it is appropriate to defer responsibility. I won't attempt to do any of this here. What is important for our purposes is that this account can be extended in a very natural way to allow for assessment-relative truth. For whenever an assertion is challenged, there are always two relevant contexts: the context in which the assertion was originally made and the context in which the challenge must be met. A natural way to give significance to doubly context-relative truth, then, would be to say that what must be established when an assertion is challenged is truth relative to the original context of use and the asserter's *current* context of assessment (at the time of the challenge):

> *Assertoric Commitment (Dual Contexts)*: In asserting that p at a context C_U, one commits oneself to providing adequate grounds for the truth of p (relative to C_U *and one's current context of assessment*), in response to any appropriate challenge, or (when appropriate) to deferring this responsibility to another asserter on whose testimony one is relying.[24] One can be released from this commitment only by withdrawing the assertion.[25]

Note that, although this account assumes that it makes sense to talk about contexts of assessment, it does not assume that propositional truth actually *varies* with the context of assessment. So non-relativists should be able to accept it, though for them the mention of ''one's current context of assessment'' will be an idle wheel. What we have, then, is a plausible story about the role of truth in our practice of assertion that gives a significance to talk of truth relative to a context of assessment, without prejudging the question whether we can actually

[24] In speaking of ''grounds for truth'', I do not mean to imply that the justification must be explicitly semantic. One can give grounds for the truth of a proposition p relative to context of use C_{U1} and context of assessment C_A simply by asserting another proposition at C_{U2} whose truth relative to C_{U2} and C_A entails, or is evidence for, the truth of p relative to C_{U1} and C_A. (The grounds one offers can themselves be challenged, of course, as can their status as grounds.)

[25] Those who accept only the weaker account of assertoric commitment in n. 23, above, may modify it as follows to accommodate assessment-relative truth: one who asserts that p at C_U is obliged to withdraw this assertion in context of assessment C_A if p is shown to be false relative to C_U and C_A.

assert anything whose truth is relative in this way. Indeed, this account gives us a way to test particular semantic hypotheses that make use of relative truth, by settling the *normative* consequences of these hypotheses.

We can now get a better feel for the assessment-sensitive semantics for ''know'' by examining its normative consequences. Suppose that Linda asserts, in a ''low standards'' context C_1, that Joe knew on 10 March that his car was in his driveway. If the assertion is challenged at this point, Linda must defend it by showing that

(*a*) Joe's car was in his driveway on 10 March, and
(*b*) Joe's epistemic position with respect to this fact was good enough on 10 March to meet the (low) standards in play at C_1.[26]

Suppose that a little while later, standards are raised. If Linda's assertion is challenged in this new context, C_2, she must defend it by showing that

(*a*) Joe's car was in his driveway on 10 March, and
(*b*) Joe's epistemic position with respect to this fact was good enough on 10 March to meet the (higher) standards in play at C_2.

In asserting that Joe knew on 10 March that his car was in his driveway, Linda takes on an open-ended commitment to show, whenever her assertion is (appropriately) challenged at a context *C*, that what she asserted is true by the standards in play at *C*—even if these standards are different from those that were in play when she made the assertion. If she lacks the resources to reply to a challenge, or if the challenge is unanswerable, then she is obliged to withdraw her assertion.

It seems to me that there is nothing incoherent about taking on such a commitment. Indeed, the facts about our use of ''know'' surveyed in §2, above suggest that we implicitly take ourselves to be bound by just such a commitment whenever we attribute knowledge.

6. CONCLUSION

According to the ''relativist'' semantics I have proposed, the epistemic standards relevant for determining whether someone can be truly said to ''know'' something are determined by the context of assessment, not

[26] To simplify the exposition, I am ignoring the possibility of deferring to the word of another.

the context of use. Consequently, in assessing knowledge claims made at different contexts for truth or falsity, one need not keep track of the standards that were in place at these contexts. The only relevant standards are the ones *now* in place.[27] This is why knowledge attributions can be reiterated and reported homophonically.

On this view, knowledge attributions are not as robustly objective as ordinary claims about the world. We must be prepared to withdraw a knowledge attribution if standards change, even if the subject's epistemic position is just as we thought it was. Relatedly, when we challenge others for having made false knowledge claims, we may be assessing them in light of standards higher than the ones they recognized when they made them. Isn't this unfair? Not unless retracting an assertion is always tantamount to admitting that the assertion was made irresponsibly: and of course it is not, even without assessment sensitivity in the picture. When standards rise, speakers withdraw their knowledge attributions and take them to have been false, but they needn't (and typically don't) take themselves to have acted irresponsibly in making them. One indication of this is that when standards fall again, they go right back to their old ways, rather than becoming more cautious in attributing knowledge. This is not so strange if we think of knowledge attributions as temporary record-keeping devices—tools for keeping track of a normative status keyed to ever-changing present circumstances—rather than straightforward statements of fact.

If I am right, then knowledge attributions made blamelessly and with full access to the relevant facts must sometimes be withdrawn as false. In my view, philosophers have been too quick to find this incoherent. Sceptics argue that we are right to withdraw our knowledge claims in the face of sceptical challenges; they conclude that these claims were not responsibly made in the first place. Dogmatists and contextualists argue that we are wrong to withdraw our knowledge claims, precisely because

[27] Keith DeRose might call this view ''single scoreboard semantics run amok'' (see DeRose 2004*a*). On this view, there is a ''single scoreboard'' not just for all parties to a single conversation, but for *all* uses of ''know'' as assessed from any one perspective. It seems to me that the very same arguments DeRose uses to support his view that the various parties to a conversation share a single scoreboard can be applied transtemporally to show that the semantics for ''know'' must be use-invariant. If we shouldn't ignore the fact that speakers in a single conversation take themselves to be contradicting and agreeing with each other in making knowledge claims, then we shouldn't ignore the fact that speakers take present knowledge claims to contradict or agree with past ones, even ones made when different standards were in play.

they were responsibly made. I say that both sides have part of the truth: we are right to withdraw our knowledge claims in the face of certain sceptical challenges, even though they were responsibly made and we haven't learned anything new. A relativist semantics for ''know'' allows us to understand how this can be.

REFERENCES

Belnap, N., and M. Green (1994) 'Indeterminism and the Thin Red Line', *Philosophical Perspectives*, 8: 365–88.

Brandom, R. (1983) 'Asserting', *Noûs*, 17: 637–50.

—— (1994) *Making it Explicit* (Cambridge, Mass. Harvard University Press).

Churchland, P. (1979) *Scientific Realism and the Plasticity of Mind* (Cambridge: Cambridge University Press).

Cohen, S. (2001) 'Contextualism Defended: Comments on Richard Feldman's ''Skeptical Problems, Contextualist Solutions'' ', *Philosophical Studies*, 103: 87–98.

DeRose, K. (2000) 'Now you Know it, Now you Don't', in *Proceedings of the Twentieth World Congress of Philosophy*, v. *Epistemology* (Bowling Green, Ohio: Philosophy Documentation Center), 91–106.

—— (2002) 'Assertion, Knowledge, and Context', *Philosophical Review*, 111: 167–203.

—— (2004*a*) 'Single Scoreboard Semantics', *Philosophical Studies*, 119: 1–21.

—— (2004*b*) 'The Problem with Subject-Sensitive Invariantism', *Philosophy and Phenomenological Research*, 68: 346–50.

Dummett, M. (1959) 'Truth', *Proceedings of the Aristotelian Society*, NS 59: 141–62.

—— (1978) *Truth and Other Enigmas* (Cambridge, Mass. Harvard University Press).

Egan, A., J. Hawthorne, and B. Weatherson (forthcoming) 'Epistemic Modals in Context', in G. Preyer and P. Peter (eds.), *Contextualism in Philosophy* (Oxford: Oxford University Press).

Evans, G. (1985) 'Does Tense Logic Rest upon a Mistake?', in *Collected Papers* (Oxford: Oxford University Press), 343–63.

Feldman, R. (2001) 'Skeptical Problems, Contextualist Solutions', *Philosophical Studies*, 103: 61–85.

Grice, P. (1989) *Studies in the Way of Words* (Cambridge, Mass.: Harvard University Press).

Hawthorne, J. (2004) *Knowledge and Lotteries* (Oxford: Oxford University Press).

Kaplan, D. (1989) 'Demonstratives: An Essay on the Semantics, Logic, Metaphysics, and Epistemology of Demonstratives and Other Indexicals', in J. Almog, J. Perry, and H. Wettstein (eds.), *Themes from Kaplan*. (Oxford: Oxford University Press), 481–566.

King, J. (2003) 'Tense, Modality, and Semantic Value', *Philosophical Perspectives*, 17, *Language and Philosophical Linguistics*, ed. J. Hawthorne and D. Zimmerman (Oxford: Blackwell), 195–245.

Kölbel, M. (2002) *Truth Without Objectivity* (London: Routledge).

Kompa, N. (2002) 'The Context Sensitivity of Knowledge Ascriptions', *Grazer Philosophische Studien*, 64: 79–96.

Lewis, D. (1979) 'Scorekeeping in a Language Game', *Journal of Philosophical Logic*, 8: 339–59.

—— (1980) 'Index, Context, and Content', in S. Kanger and S. Öhman (eds.), *Philosophy and Grammar* (Dordrecht: Reidel).

—— (1998) *Papers in Philosophical Logic* (Cambridge: Cambridge University Press).

MacFarlane, J. (2003) 'Future Contingents and Relative Truth', *Philosophical Quarterly*, 53(212): 321–36.

Percival, P. (1994) 'Absolute Truth', *Proceedings of the Aristotelian Society*, 94: 189–213.

Richard, M. (2004) 'Contextualism and Relativism', *Philosophical Studies*, 119: 215–42.

Rosenberg, J. F. (2002) *Thinking about Knowing* (Oxford: Oxford University Press).

Rysiew, P. (2001) 'The Context-Sensitivity of Knowledge Attributions', *Noûs*, 35(4): 477–514.

Schaffer, J. (2004) 'Skepticism, Contextualism, and Discrimination', *Philosophy and Phenomenological Research*, 69: 138–55.

Schiffer, S. (1996) 'Contextualist Solutions to Scepticism', *Proceedings of the Aristotelian Society*, 96: 317–33.

Searle, J. (1969) *Speech Acts* (Cambridge: Cambridge University Press).

—— (1979) *Expression and Meaning* (Cambridge: Cambridge University Press).

Stalnaker, R. (1978) 'Assertion', in P. Cole (ed.), *Syntax and Semantics*, ix. *Pragmatics* (New York: Academic Press).

Stanley, J. (2004) 'On the Linguistic Basis for Contextualism', *Philosophical Studies*, 119: 119–46.

This page intentionally left blank

9. Contrastive Knowledge

Jonathan Schaffer

I know a hawk from a handsaw
(Hamlet)

Does G. E. Moore know that he has hands? *Yes*, says the dogmatist: Moore's hands are right before his eyes. *No*, says the skeptic: for all Moore knows he could be a brain-in-a-vat. *Yes* and *no*, says the contrastivist: *yes*, Moore knows that he has hands rather than stumps; but *no*, Moore does not know that he has hands rather than vat-images of hands.

The dogmatist and the skeptic suppose that knowledge is a *binary*, *categorical* relation: *s* knows that *p*. The contrastivist says that knowledge is a *ternary*, *contrastive* relation: *s* knows that *p* rather than *q*.

I propose to develop the contrastive account of knowledge. Such an account requires five stages of development. One needs to report the *use* of knowledge ascriptions (§1), limn the *structure* of the knowledge relation (§2), show how the ascriptions *express* the relation (§3), analyze or otherwise *illuminate* the relation (§4), and resolve outstanding *paradoxes* (§5). On route, I will compare the contrastive account to *binary* accounts. Once home, I will compare contrastivism to *contextualism* (§6).

The view that emerges links knowledge to *inquiry* and to *discrimination*. There is no such thing as inquiring into *p*, unless one specifies: *as opposed to what*? There is no such thing as discriminating that *p*, unless one adds: *from what*? And likewise I will argue that there is no such thing as knowing that *p*, unless one clarifies: *rather than what*?

1. USE

The first stage of an account of knowledge is to report the *use* of knowledge ascriptions. What are knowledge ascriptions *for*? I propose:

(1) Knowledge ascriptions certify that the subject is able to answer the question.

I will now clarify, argue for, and address objections to (1).

Clarifications

''Knowledge ascriptions'' in (1) refers to tokens containing ''knows'' in the *informational sense*. In the terms of Gilbert Ryle (1949), (1) covers ''knows that'', not ''knows how''.[1] More accurately, (1) covers ''knows'' in the sense of *savoir* not *connaître* (French), and in the sense of *wissen* not *kennen* (German).

''Certify'' describes the act performed by the ascriber. In the terms of J. L. Austin (1962), such certification constitutes the *illocutionary force* of the utterance. In the terms of Robert Brandom (1994), such certification consists in the conferral of an *entitlement* (''You may answer the question''), with subsequent *commitment* to endorsing the answer (''Whatever you say'').

''Able to answer'' denotes an *epistemic capacity*. It is epistemic in that one may guess rightly without having the requisite ability (just as a blind throw may find the target). It is a capacity insofar as one need not actually speak or otherwise exercise the ability in order to possess it.

''The question'' denotes the *options* relevant in the context of ascription. The question need not be explicitly posed, but it is always recoverable from context, since a context may be modeled as ''the set of possible worlds recognized by the speakers to be the 'live options' relevant to the conversation'' (Robert Stalnaker 1999*a*: 84–5).

By way of illustration, imagine that Holmes and Watson are investigating who stole the sapphire. Here the live options might be: {Black stole the sapphire, Scarlet stole the sapphire, Mustard stole the sapphire}. Now imagine that Holmes finds Black's fingerprints on the lock. So Watson reports, ''Holmes knows who stole the sapphire.'' What Watson is *doing* with this speech act, according to (1), is giving his

[1] Though Ryle's ''knows that''/''knows how'' distinction does not mark the informational/acquaintance distinction accurately. First, Ryle's distinction misses other forms of knowledge ascription, such as ''knows who'', ''knows what'', and ''knows where'', which are informational. Second, Ryle's distinction obscures the fact that ''knows how'' is informational, as evident in ''I know how turtles reproduce'', and Monty Python's explanation of how to play the flute: ''Well, you blow in one end and move your fingers up and down the outside.'' See Jason Stanley and Timothy Williamson (2001) for further discussion.

stamp of approval to Holmes, for selecting who stole the sapphire. Watson is identifying someone able to answer the question. He is fingering an answerer.

Arguments

First, (1) *fits our practice*. In the case of Holmes and Watson, one expects Watson to report that Holmes knows who stole the sapphire, only when Holmes is able to answer the question. Or consider our practice of testing students. The professor attributes knowledge to the students on the basis of which questions they are able to answer (''Let's see what you know''). Or consider our practice of fielding questions. One may say ''I know'' or ''Ask Pam, she knows''. One fingers an answerer.

Second, (1) serves our goal of *scoring inquiry*. Our ultimate epistemic goal is truth, and our method for seeking truth is inquiry.[2] So it is apt for knowledge ascriptions to be directed to questions, to gauge the progress of inquiry. In this vein, Christopher Hookway remarks: ''The central focus of epistemic evaluation is . . . the activity of inquiry . . . When we conduct an inquiry, . . . we attempt to formulate questions and to answer them correctly'' (1996: 7).

Third, (1) *explains the other proposals in the literature*. For instance, according to Ludwig Wittgenstein, knowledge ascriptions serve to indicate when ''one is ready to give compelling grounds'' (1969: §243; also §§50, 483–5). While according to Edward Craig, the role of the knowledge ascription is ''to flag approved sources of information'' (1990: 11).

Wittgenstein's and Craig's proposals must be *relativized to questions*. If one is inquiring into *who* stole the sapphire, then the evidence of Black's fingerprints on the lock might constitute compelling grounds for ''Black stole the sapphire'', and the detective might count as an approved source of that information. But if one is inquiring into *what* Black stole, then the evidence of his fingerprints might not constitute compelling grounds for ''Black stole the sapphire'', and the detective might not count as an approved source. The fingerprints may help identify who did the stealing, but they may not help establish what was stolen. In an

[2] The *Peircean* (following C. S. Peirce 1877) may rephrase the argument of the main text as: ''Our ultimate epistemic interest is the fixation of belief. Our method for fixing belief is inquiry.'' The same directedness to answers would be called for.

inquiry into *what* Black stole, the owner's testimony that there was a sapphire in the safe might constitute compelling grounds for ''Black stole the sapphire'', and the owner might count as an approved source of that information. The owner's testimony may help identify what was stolen, but it may not help identify who stole it. While if one is inquiring into *how* Black obtained the sapphire (or *why* he stole it, etc.) then different evidential factors come to the fore. In short, what counts as compelling grounds, and who counts as an approved source, depends on which question is at issue.

Now (1) clarifies Wittgenstein's and Craig's proposals, by imposing the needed relativization to a question. And (1) *explains what is right* about these proposals, suitably relativized. What counts as compelling grounds relative to a question is just what counts as a basis for an answer. Who counts as an approved source relative to a question is just who is able to provide an answer.[3]

Objections

First, one might object that (1) is *overly intellectual* in its focus on answers. We routinely ascribe knowledge to *animals* (and infants, etc.), though they cannot answer questions or participate in inquiry. Thus, the objection concludes, (1) misconstrues our practice.

In reply, animals may be thought to have the *ability* to answer, which is all that (1) requires. That is, animals may have the *cognitive basis* by which the answer is reached, though they lack the means to express it. Thus Fido might know who feeds him, though he cannot express the answer save through his affections.[4]

Second, one might object that (1) is *socially disruptive* in its relativity to questions. We *traffic* in knowledge ascriptions, without tracking

[3] A further example: John Greco addresses the ''what are we doing?'' question by identifying: ''an important illocutionary force of knowledge attributions: namely, that when we credit knowledge to someone we mean to give the person credit for getting things right'' (2002: 111). What suffices for 'getting things right' is just what suffices for selecting the right answer.

[4] Our intuitions to ascribe knowledge to animals seem to sway with our inclinations to ascribe them the concepts involved. For instance, our inclination to say, ''Fido knows where he buried the bone'', seems to sway with our inclination to say that Fido possesses the concepts *bury* and *bone*. Thus, to the extent that we are willing to ascribe knowledge to animals, we are committed to their possessing the concepts that would form the cognitive basis for answering.

questions. For instance, if Watson tells Lestrade, ''Holmes knows that Black stole the sapphire'', then Lestrade may repeat Watson's words to Scotland Yard, in a different context with a different question on the table. Thus, the objection concludes, (1) undermines our practice.

In reply, trafficking in knowledge ascription must be regarded as a *risky act*, which is all that (1) entails. The careless trafficker may wind up doing something inappropriate. Imagine that, while Holmes and Watson were pursuing the question of who stole the sapphire, Lestrade and Scotland Yard were stuck on the question of whether what was stolen was a sapphire or a paste imitation. If Lestrade now repeats Watson's words to Scotland Yard, then Lestrade would have acted inappropriately, by representing Holmes *as if* he had tested the sapphire.

There is nothing special about knowledge ascriptions here. We traffic in *assertions* generally, while recognizing that repeating any assertion out of context is risky. Misunderstandings may arise when the originator and the repeater are in *conversational disequilibrium*. That is, if the originator and repeater have different presuppositions, then their assertions may be identical in word but not in deed. We redress misunderstandings if they count.

The ultimate test of (1), of course, is whether it coheres with a successful epistemology. I will argue (§2) that (1) calls for a contrastive view of knowledge. Whether this counts as a further argument for (1), or an objection to it, is left to the reader's judgment.

2. STRUCTURE

The second stage of an account of knowledge is to limn the *structure* of the knowledge relation. What is its *form*? I propose:

(2) The knowledge relation has the ternary, contrastive structure: K*spq*.

Here K is the knowledge relation, *s* is the subject, *p* is the proposition selected, and *q* is the proposition rejected.[5] K*spq* may thus be rendered as: *s* knows that *p* rather than *q*.

[5] The proposition *q* may be glossed as the disjunction of the 'relevant alternatives'. As such, two constraints on *q* are needed: (i) *q* must be non-empty, and (ii) *p* and all the disjuncts of *q* must be pairwise exclusive.

Objection

One might object that (2) is *implausibly radical* in contravening the widespread assumption that knowledge has the binary form: K*sp*. Have so many epistemologists been *wrong*?[6] Thus, the objection concludes, (2) deserves to be met with a blank stare, or at least with steeply arched brows.

In reply, it is unclear *why* the assumption of binarity is so widespread. For what it is worth, I have found no explicit arguments for binarity in the literature. Perhaps binarity is assumed because it reflects the *surface form* of knowledge ascriptions. After all, some knowledge ascriptions look binary: ''I know that I parked the car on Elm.'' But surface form is *equivocal*. There are interrogative ascriptions that do not look binary: ''I know where I parked the car.'' And there are declarative ascriptions that look explicitly contrastive: ''I know that I parked the car on Elm rather than Main''. In any case, surface form can mislead.

Perhaps binarity is assumed because it reflects the *intuitive adicity* of knowledge. But adicity is not so easily intuited. Our intuitive judgments merely provide evidence as to the acceptability of utterances (Noam Chomsky 1977). Anything more is *theory*.

Perhaps binarity is assumed because it is required to solve *theoretical problems*. But which? What have accounts of K*sp* produced but problems? *What if contrastivity works better*?

Arguments

First, (2) *fits* (1) by logging the question. That is, the contrastive structure K*spq* records the information about which question was asked, and so is the right form for the job of fingering who is able to answer.

To begin with, the ability to answer is *question-relative*. Some questions are harder to answer than others. The ability to answer p to the question on the table does *not* entail the ability to answer p to all other

[6] Some exceptions: Fred Dretske flirts with the contrastive view: ''To know that x is A is to know that x is A within a framework of relevant alternatives, B, C, and D. This set of contrasts . . . serve to define what it is that is known'' (1970: 1022). Bredo Johnsen describes the intuitive content of knowledge ascriptions as contrastive: ''what is known is always a contrastive proposition to the effect that P-rather-than-any-other-member-of-category-C is true'' (2001: 401), though he makes this point in service of *skepticism*. And Adam Morton and Anti Karjalainen (2003), as well as Walter Sinnot-Armstrong (2004), uphold contrastivism, though as a *revisionary* proposal.

questions in the field. Anyone who has devised an exam will recognize this—add a trick option, and the question will be harder. Compare:

(Q1) Is there a *goldfinch* in the garden, or a *raven*?
(Q2) Is there a *goldfinch* in the garden, or a *canary*?
(Q3) Is there a goldfinch in the *garden*, or at the *neighbor's*?

All can be answered by *p*: there is a goldfinch in the garden. But the ability to answer Q1 does not entail the ability to answer Q2 or Q3. Q1 is an easy question. While to answer Q2 one might need an ornithologist, and to answer Q3 one might need the homeowner. So fingering answerers requires logging the question, because the abilities to answer Q1–Q3 are different abilities.

Logging the question requires recording the alternatives. All well-formed questions are multiple-choice questions. As James Higginbotham writes, ''An *abstract question* [is] a nonempty *partition* Π of the possible states of nature into *cells*'' (1993: 196). These cells are the semantic image of a (possibly infinite) *multiple-choice slate*.[7]

The contrastive structure K*spq* logs the question, by recording the alternatives. Here {*p*, *q*} conforms to the multiple-choice slate—*p* corresponds to the selected answer and *q* to the disjunction of the rejected alternatives. Thus one who knows that *p*: there is a goldfinch in the garden, rather than *q1*: there is a raven in the garden, is able to answer Q1. While one who knows that *p* rather than *q2*: there is a canary in the garden, can answer Q2. And one who knows that *p* rather than *q3*: there is a goldfinch at the neighbor's, can answer Q3. Thus differences at *q* correspond to different abilities to answer different questions. Contrast-relative knowledge is question-relative knowledge, and so befits our question-relative usage.

The second argument for (2) is that contrastivity *models inquiry* by measuring progress. Inquiry is the engine of knowledge (§1), and it is driven by a question-and-answer process.[8] Drawing on Jaakko Hintikka (1975*a*, 1981), inquiry may be modeled as a cooperative game played

[7] The association of questions with multiple-choice slates is known as *Hamblin's dictum* (C. I. Hamblin 1958), and is implemented in Nuel Belnap and Thomas Steel's (1976) erotetic logic, and maintained in the leading linguistic treatments of interrogatives, such as that by Jeroen Groenendijk and Martijn Stokhof (1997).

[8] This is the *Deweyian view* of inquiry: ''Inquiry and questioning, up to a certain point, are synonymous terms.'' (1938: 105). See also Isaac Levi (1984), in which expansion of a belief corpus is directed by an ultimate partition over a set of possible answers to a

between Questioner and Answerer, represented by a sequence of question-and-answer pairs $<< Q_1, A_1 >, < Q_2, A_2 >, \ldots, < Q_n, A_n >>$. Progress in inquiry is movement through the sequence, so answers make for progress. Suppose the chemist is identifying a sample of potassium (K), via the following line of inquiry: << Q1: What element is the sample?, A1: Potassium>, < Q2: Is the sample ionized?, A2: No>>. To answer Q1, the chemist might run experiments (putting the question to nature) that test for atomic mass. To answer Q2, the chemist might run experiments that test for charge or reactivity (K and K^+ have nearly the same atomic mass, but while K is neutral and reactive, K^+ is positive and inert).[9]

The contrastive structure measures progress, because q measures which stage of inquiry has been concluded. The chemist progresses from ignorance through knowledge that the sample is K rather than some other element: $Kspq_1$; and then knowledge that the sample is K rather than K^+: $Kspq_2$. The epistemic state that corresponds to no progress is: $\sim Kspq_1$ & $\sim Kspq_2$; partial progress is: $Kspq_1$ & $\sim Kspq_2$; and complete progress is: $Kspq_1$ & $Kspq_2$. In general, progress can be pictured in terms of finding actuality in *widening regions* of logical space. To find w_α from amongst worlds w_1–w_m is to know that $\{w_\alpha\}$ rather than $\{w_1, w_2, \ldots, w_m\}$. To make further progress is to find w_α from amongst worlds w_1–$w_n (n > m)$, which is to know that $\{w_\alpha\}$ rather than $\{w_1, w_2, \ldots, w_m, \ldots, w_n\}$.[10] Thus differences at q correspond to different stages of inquiry. Contrast-relative knowledge is progress-relative knowledge, and so befits the structure of inquiry.

question. For an application to scientific progress, see Scott Kleiner (1988). As Matti Sintonen comments in this regard: ''If there is a philosophy of a working scientist it certainly is the idea that inquiry is a search for questions and answers.'' (1997: 234)

[9] Note that the entire inquiry is framed within certain presuppositions. At no point, for instance, does the chemist test the option: *the sample is but a dream*. If one looks at dichotomous keys, for instance, one never finds an entry for *pinch yourself*.

[10] On this view of progress, progress essentially consists in *replacing presupposition with evidence*. When the subject is able to answer Q1 and hence able to find w_α from amongst worlds w_1–w_m, the remainder of logical space is simply presupposed away. When the subject progresses through Q2 and is able to find w_α from amongst worlds w_1–$w_n (n > m)$, less is presupposed away and more is ruled out by evidence. The (ideal) limit of inquiry would consist in finding w_α from amongst all of logical space, which would be a full grasp of truth by evidence. Thus movement towards the limit consists in finding w_α from amongst widening spheres of logical space, which would be a greater grasp of truth by evidence, and a lesser need for presupposition. Of course, at each stage short of the limit, assumptions remain. But that does *not* mean that there had been no progress—not all assumptions are equal.

The third argument for (2) is that contrastivity *fits perception*, which is basically a discriminatory ability. Thus the psychophysicist S. S. Stevens remarks: ''When we attempt to reduce complex operations to simpler and simpler ones, we find in the end that discrimination or differential response is the fundamental operation. Discrimination is prerequisite even to the operation of denoting or 'pointing to,' '' (quoted by C. S. Watson 1973: 278). The discriminatory powers of perception are codified in Weber's Law, which states that just noticeable differences are well-described by: $\Delta S/S = K$. In words: the size of a just noticeable difference in stimulation S is a constant proportion K of the existing stimulus. For instance, in normal humans, just noticeable differences in tonal frequency are well-described by $K = .0025$ (at least for the central portion of the human range). Thus if the existing stimulus S is 1000 Hz, then differences of ± 2.5 Hz will be just noticeable.

The contrastive structure fits perceptual discrimination, by logging both the reported stimulus: *p*, and what the stimulus was discriminated from: *q*. Suppose that a normal human subject Norm hears a tone of S1 = 1000 Hz. Norm can discriminate S1 from a tone of S2 = 1005 Hz, but cannot discriminate S1 from S3 = 1001 Hz. Then he knows that *p*: the tone is 1000 Hz, rather than *q1*: the tone is 1005 Hz. But he does not know that *p*: the tone is 1000 Hz, rather than *q2*: the tone is 1001 Hz. In general, for a stimulus S and a perceiver whose just noticeable difference for such stimuli is $K = x$, this perceiver can know that he is perceiving S rather than any difference in S greater than or equal to KS, and cannot know that he is perceiving S rather than any lesser difference. Thus differences at *q* correspond to what the percept is being discriminated from. Contrast-relative knowledge is discrimination-relative knowledge, and so befits the nature of perception.

In the remaining sections I will add three more arguments for (2), namely that (2) is the best fit for decoding knowledge ascriptions (§3), illuminating the knowledge relation (§4), and resolving the closure paradox (§5).

Comparison

The ultimate test of contrastivity is how it compares to binarity.[11] How does K*spq* compare to K*sp*?

[11] Why not let knowledge come in *both* binary and contrastive forms? Because (i) this would require an ambiguity in ''knows'' that the evidence does not support, (ii) I will argue

I suspect that K*sp* induces systematic problems for lack of a contrast slot. Nothing in the K*sp* relation logs the queried alternatives, the stage of inquiry, or the discriminatory task. So there is no natural fit to fingering answerers, modeling inquiry, and measuring perception. Consider the subject who enjoys *merely partial success*. For instance, consider the subject who can answer, ''Goldfinch or raven?'' but not, ''Goldfinch or canary?''[12] Given binarity, he must either know that the bird is a goldfinch, or not (I leave it to the dogmatist and skeptic to dispute which). But if the subject knows, then his inability to answer, ''Goldfinch or canary?'' seems inexplicable. With a minimum of logical acumen, he ought to be able to apply his alleged knowledge to answer this further question. So partial success would explode into total victory. Whereas if the subject does not know that the bird is a goldfinch, then his ability to answer, ''Goldfinch or raven?'' seems inexplicable. He ought not to be able to answer where he is allegedly ignorant. So partial success would collapse into total defeat. K*sp* seems too impoverished to provide a stable account of partial success in answering, inquiry, and discrimination.[13]

My aim is to develop a contrastive view, not to refute the binary view in all its forms. *That* would be a Herculean task. Perhaps the binary theorist can find some devious strategy to model partial success. But I think it fair to conclude, at the least, that (2) provides *the more natural fit* to the contrast-relative tasks of answering, inquiry, and discrimination.

3. ENCODING

The third stage of an account of knowledge is to show how knowledge ascriptions *express* the knowledge relation. What is the *code*? I propose:

(3) Knowledge ascriptions encode K*spq*, by encoding relations to questions.

(§3) that the contrastive form fits all of our knowledge ascriptions, and (iii) I will suggest (§5) that the binary form is paradoxical.

[12] Or, to borrow a case from Dretske (1970), consider the zoo-goer who can answer, ''Zebra or mule?'', but not, ''Zebra or cleverly painted mule?''

[13] Perhaps the *contextualist* has a way to model partial success, in terms of the *plurality* of binary K_x relations they postulate as the range of semantic values for ''knows''. Here there is the added structure of a subscript to K. For further discussion of contextualism, see §6.

I will now defend (3) by exhibiting three main surface forms of knowledge ascription, and showing the mechanisms for question-relativity encoded in each.

Surfaces

There are three main types of knowledge ascription (in the informational sense of ''knows'': §1), which may be distinguished syntactically: (i) interrogative ascriptions, which employ a *wh*-headed complement phrase, such as: ''I know what time it is'', (ii) noun ascriptions, which employ a noun (determiner) phrase, such as: ''I know the time'',[14] and (iii) declarative ascriptions, which employ a *that*-headed complement phrase, such as: ''I know that it is midnight''. An account of encoding is responsible for every type of ascription.

Epistemologists, however, have focused nearly exclusively on declarative ascriptions.[15] Interrogative and noun ascriptions are typically ignored, or else hastily fitted to the Procrustean bed of K*sp*. Why the focus on declarative ascriptions? These seem relatively rare in natural language, especially when compared to interrogative ascriptions. So why the focus on such an unrepresentative sample? Perhaps the widespread focus on declarative ascriptions is due to the widespread assumption that knowledge is a binary relation (§2). Perhaps here is a case where theory dictates observation: ''Our theoretical presuppositions about the natural order tell us what to expect'' (Larry Laudan 1977: 15). In any case, it must not be *presumed* that declarative ascriptions are

[14] Noun ascriptions can express either the informational or practical sense of ''knows'' (§1). Here are three tests for whether a given noun ascription is informational or practical. First, only the practical sense supports comparatives: compare ?''I know the time very well'' and ?''I know the time better than I know Ben'', with ''I know Ann very well'' and ''I know Ann better than I know Ben''. Second, only the practical sense supports ''but not as such'' constructions: compare ?''I know the time but not as such'' to ''I know Ann but not as such''. Third, only the informational sense entails knowledge-*wh*: ''I know the time'' entails ''I know what the time is'', but ''I know Ann'' does not entail ''I know what Ann is'' (nor does it entail ''I know who she is'' or ''I know where she is'', etc.).

[15] Some exceptions: Hintikka (1975*b*) distinguishes the full spectrum of knowledge ascriptions, yet he classifies all the others as departures from the ''knows that'' form. And David Lewis (1982), Steven Boër and William Lycan (1986), and Stanley and Williamson (2001) discuss (respectively) ''knows whether'', ''knows who'', and ''knows how''. Yet even here Stanley and Williamson contrast ''question-embedding uses of 'know' and *normal* clausal-complement uses of 'know' '' (2001: 421; italics added), and all of these philosophers attempt to reduce knowledge-*wh* to knowledge that p. The exceptions prove the rule.

more fundamental. Perhaps it is the interrogative ascriptions that are more fundamental, in the sense that it is they that wear their logical forms closer to their surfaces.

Mechanisms

Starting with interrogative ascriptions (perhaps the most frequent in natural language), these embed *questions*. Questions present contrasts (§2). The mechanism of question-relativity is thus on the surface, in the *wh*-clause. So, for instance, if one says, ''I know who stole the bicycle'', then the embedded question ''who stole the bicycle'' presents a set of alternatives, such as: {Mary stole the bicycle, Peter stole the bicycle, Paul stole the bicycle}.[16] Here p is the selected answer, and q is the disjunction of the rejected alternatives. So if it was Mary who stole the bicycle, then to know who stole the bicycle is to know that p: Mary stole the bicycle, rather than q: Peter stole the bicycle or Paul stole the bicycle. In this vein, Higginbotham says: ''Mary knows who John saw'' should be interpreted as: ''Mary knows the (or an) answer to the question who John saw'' (1993: 205).

Here are three tests that confirm the question-relative treatment of interrogative ascriptions. First, differences at q can *affect truth-values*. For instance, suppose that Joe glances at George W. Bush speaking on television, and compare the following knowledge claims:

(I1) Joe knows whether Bush or Janet Jackson is the speaker.
(I2) Joe knows whether Bush or Will Ferrell is the speaker.[17]

Intuitively, I1 may be true but I2 false. Joe can discriminate Bush from Jackson, but perhaps only First Lady Laura Bush can discriminate Bush from Ferrell. In other words, Joe is able to answer whether Bush or Jackson is the speaker (this is an easy question—Joe knows the answer to *that*), but Joe cannot answer whether Bush or Ferrell is the speaker (this is a hard question—Joe can only *guess*). The difference in truth-value between I1 and I2 is not due to a difference in s or in p—the

[16] The set of alternatives is determined by (i) the contextually determined domain of quantification, and (ii) the matrix: x is a bicycle thief. So if the individuals in the domain are Mary, Peter, and Paul, then the set of queried alternatives is: {Mary stole the bicycle, Peter stole the bicycle, Paul stole the bicycle}.

[17] Background information: Janet Jackson is a pop diva who would be quite hard to confuse with Bush, while Will Ferrell is a skilled Bush impersonator.

subject is Joe and the true answer p is: Bush is the speaker. So the difference must lie elsewhere. The difference is at q, between $q1$: Jackson is the speaker, and $q2$: Ferrell is the speaker. The question is what is differentiating the truth-value.

To take another example (from §2), suppose that Ann sees a goldfinch in the garden, and compare the following claims:

(I3) Ann knows whether there is a goldfinch or a raven in the garden.
(I4) Ann knows whether there is a goldfinch or a canary in the garden.
(I5) Ann knows whether there is a goldfinch in the garden or at the neighbor's.

Intuitively, I3–I5 may differ in truth-value. I3 is a relatively easy item of knowledge. While I4 is harder, requiring some ornithology. And I5 is incommensurable, requiring some sense of the landscape. The difference in truth-value between I3–I5 is not due to a difference in s or in p—the subject is Ann and the true answer p is: there is a goldfinch in the garden. So the difference must lie elsewhere. The difference is at q, between $q3$: there is a raven in the garden, $q4$: there is a canary in the garden, and $q5$: there is a goldfinch at the neighbor's. The question is what is differentiating the truth-value.

A second confirmation for the question-relative treatment of interrogative ascriptions comes from *existential generalization*. If I know who stole the bicycle, then it follows that there is a question (namely, the question of who stole the bicycle) that I know the answer to. Likewise if I know what time it is, then it follows that there is a question (the question of what time it is) that I know the answer to. The question is what is being generalized on.

A third confirmation comes from *substitution*. If I know when Napoleon was born, and if the question of when Napoleon was born is a historical question, then it follows that I know the answer to a historical question. Likewise if I know why the sky looks blue, and if the question of why the sky looks blue is a scientific question, then it follows that I know the answer to a scientific question. The question is what is being substituted for.[18]

[18] For further discussion of interrogative ascriptions, see Schaffer n.d.

Turning to noun ascriptions, these are at least semantically equivalent to interrogative ascriptions. Thus, for instance, ''I know the time'' is semantically equivalent to ''I know what time it is'', and ''I know the murderer'' (in the informational sense) is semantically equivalent to ''I know who the murderer is''.[19] The mechanism of question-relativity is thus present in the interpretation of the noun phrase. So if it is noon, then to know the time is that it is to know that *p*: the time is noon, rather than *q*: the time is 1p.m. or 2p.m. or . . . or 11a.m. And if the murderer is Oswald, then to know the murderer is to know that *p*: Oswald is the murderer, rather than *q*: Castro is the murderer or the CIA is the murderer.

The question-relative treatment of noun ascriptions is confirmed by the same three tests as with interrogative ascriptions. First, differences at *q* can affect truth-value. For instance, suppose that, in context *c1*, the domain of quantification is {Bush, Jackson}, so that the question of who the speaker is denotes: {Bush is the speaker, Jackson is the speaker}. While in *c2*, the domain of quantification is {Bush, Ferrell}, so that the question of who the speaker is denotes: {Bush is the speaker, Ferrell is the speaker}. Then consider the utterance type:

(N1) Joe knows the speaker.

Intuitively, a token of N1 may be true if uttered in *c1*, but false if uttered in *c2*. After all, Joe knows whether Bush or Jackson is the speaker—which is what *N1* is semantically equivalent to in *c1*. But Joe does not know whether Bush or Ferrell is the speaker—which is what N1 is semantically equivalent to in *c2*. There is no difference in *s* or *p* here—the subject is Joe and the true answer *p* is: Bush is the speaker. So the difference in truth-value must lie elsewhere. The difference is at *q*, between *q1*: Jackson is the speaker, and *q2*: Ferrell is the speaker. The question is what is differentiating the truth-value.

A second confirmation for the question-relative treatment of noun ascriptions comes from *existential generalization*. If I know the time, then it follows that there is a question (namely, the question of what time it is) that I know the answer to. Likewise if I know the murderer, then it follows that there is a question (the question of who the

[19] Thus Irene Heim (1979) refers to this as the *concealed question* use of noun phrases, saying: ''As we naturally understand the sentence [''John knows Bill's telephone number''] we could paraphrase it as ''John knows what Bill's telephone number is''.

murderer is) that I know the answer to. The question is what is being generalized on.

A third confirmation comes from *substitution*. If I know the date Napoleon was born, and if the question of when Napoleon was born is a historical question, then it follows that I know the answer to a historical question. Likewise if I know the reason the sky looks blue, and if the question of why the sky looks blue is a scientific question, then it follows that I know the answer to a scientific question. The question is what is being substituted for.

Moving finally to declarative ascriptions (perhaps the rarest form in natural language), these inherit their contrasts from *context*. A context is an implicit question. According to Stalnaker, a context may be represented by a set of possible worlds, ''which includes all the situations among which speakers intend to distinguish with their speech acts'' (1999*b*: 99). This set is ''the set of possible worlds recognized by the speakers to be the 'live options' relevant to the conversation'' (1999*a*: 84–5). Thus a context is a set of options (§1). A set of options is the slate of a question (§2). So if one says, ''I know that Mary stole the bicycle'', in a context in which the identity of the bicycle thief is in question, then the value of p is: that Mary stole the bicycle, and q is: that some other suspect stole the bicycle. If one says this in a context in which Mary's behavior toward the bicycle is in question, then the value of p is: that Mary stole the bicycle, and q is: that Mary acted in some other way towards the bicycle. While if one says this in a context in which the nature of Mary's contraband is in question, then the value of p is: that Mary stole the bicycle, and q is: that Mary stole some other loot. In general, context provides the default source of contrasts.

The question-relative treatment of declarative ascriptions is confirmed by the same three tests as with interrogative and noun ascriptions. First, differences at q can affect truth-value. For instance, suppose that the context set for *c1* is: {Bush is the speaker, Jackson is the speaker}. While the context set for *c2* is: {Bush is the speaker, Ferrell is the speaker}. Then consider the utterance type:

(D1) Joe knows that Bush is the speaker.

Intuitively, a token of D1 may be true if uttered in *c1*, but false if uttered in *c2*. After all, if one is wondering whether the speaker is Bush or Jackson—which is the implicit question of *c1*—then one would do well to ask Joe. But if one is wondering whether the speaker is Bush or

Ferrell—which is the implicit question of *c2*—then Joe is not the one to ask. There is no difference in *s* or *p* here—the subject is Joe and the true answer *p* is: Bush is the speaker. So the difference in truth-value must lie elsewhere. The difference is at *q*, between *q1*: Jackson is the speaker, and *q2*: Ferrell is the speaker. The question is what is differentiating the truth-value.

To take the example of the goldfinch in the garden, suppose that the context set for *c1* is: {there is a goldfinch in the garden, there is a raven in the garden}, the context set for *c2* is: {there is a goldfinch in the garden, there is a canary in the garden}, and for *c3* is: {there is a goldfinch in the garden, there is a goldfinch at the neighbor's}. Then consider the utterance type:

(D2) Ann knows that there is a goldfinch in the garden.

Intuitively, what it takes for a token of D2 to be true differs among *c1*, *c2*, and *c3*. In other words, if one is wondering whether there is a goldfinch or a raven in the garden—which is the implicit question of *c1*—then one might ask virtually anyone. While if one is wondering whether there is a goldfinch or a canary in the garden—which is the implicit question of *c2*—then one might need to ask the ornithologist. And if one is wondering whether there is a goldfinch in the garden or at the neighbor's—which is the implicit question of *c3*—then one might need to ask the homeowner. There is no difference at *s* or *p*, only at *q*. The question is what is differentiating the truth-value.[20]

A second confirmation for the question-relative treatment of declarative ascriptions comes from *existential generalization*. If I know that the time is noon, then it follows that there is a question (namely, the question of what time it is) that I know the answer to. Likewise if I know that Oswald is the murderer, then it follows that there is a

[20] John Hawthorne suggests that the question-sensitivity of our intuitions here may be explained away, on grounds that ''the very asking of a question may provide one with new evidence regarding the subject matter'' (2004: 78). The idea is that Ann has different evidence in contexts *c1*, *c2*, and *c3*, concerning which question was asked of her. *But* this assumes that (i) Ann fields the question, and (ii) Ann trusts the questioner to select the likely options. Ann need not field the question. She might not be privy to the conversation at all. Others might be discussing what she knows. (This situation might arise when one is deciding who to ask—one tries to figure out *in advance* which third party knows the answer.) In any case, Ann need not trust the questioner to select the likely options. She might just play along. (Anyone who has questioned students will recognize this situation.)

question (here, the question of who is the murderer) that I know the answer to. The question is what is being generalized on.

A third confirmation comes from *substitution*. If I know that Napoleon was born in 1769, and if the question of when Napoleon was born is a historical question, then it follows that I know the answer to a historical question. Likewise if I know that the sky looks blue because of Rayleigh scattering (blue's short wavelength causes it to get scattered far more than the longer wavelength colors), then it follows that I know the answer to a scientific question. The question is what is being substituted for.

Here are four additional arguments for the question-relativity of declarative ascriptions. The first additional argument is that declarative ascriptions should *fit the pattern* of knowledge ascriptions generally. Since interrogative and noun ascriptions are question-relative (and since ''knows'' is not ambiguous here), declarative ascriptions should be expected to be question-relative too.

A second additional argument comes from *focus*. As Dretske recognized, focus is semantically efficacious in declarative ascriptions:

> Someone claiming to know that Clyde *sold* his typewriter to Alex is not (necessarily) claiming the same thing as one who claims to know that Clyde sold his typewriter *to Alex* . . . A person who knows that Clyde *sold* his typewriter to Alex must be able to rule out the possibility that he *gave* it to him, or that he *loaned* it to him . . . But he needs only a nominal justification, if he needs any justification at all, for thinking it was Alex to whom he sold it. (1981: 373)

Following David Sanford (1991), one can model the effect of focus by sets of relevant alternatives, as follows:

$$\text{I know that} \quad \left\{ \begin{matrix} \text{Mary} \\ \text{Peter} \\ \text{Paul} \end{matrix} \right\} \quad \left\{ \begin{matrix} \text{stole} \\ \text{begged} \\ \text{borrowed} \end{matrix} \right\} \quad \text{the} \quad \left\{ \begin{matrix} \text{bicycle} \\ \text{unicycle} \\ \text{tricycle} \end{matrix} \right\}$$

Thus if one says, ''I know that *Mary* stole the bicycle'', then the value of p is: that Mary stole the bicycle, and q is: that Peter or Paul stole the bicycle. If one says, ''I know that Mary *stole* the bicycle'', then the value of p is: that Mary stole the bicycle, and q is: that Mary begged or borrowed the bicycle. While if one says, ''I know that Mary stole the *bicycle*'', then the value of p is: that Mary stole the bicycle, and q is: that Mary stole the unicycle or the tricycle. The semantic efficacy of focus is thus explained: differences in focus determine differences in the

proposition expressed. Focus is a mechanism of contrastivity.[21] Where focus is semantically effective, it is because contrasts are semantically operative.

A third additional argument comes from the *binding* test. Suppose that Sally has aced her exam. Here one might boast on her behalf: ''On every question, Sally knew the answer.'' This has a natural reading on which it is semantically equivalent to: ''On the first question, Sally knew the answer *to that question*; on the second question, Sally knew the answer *to that question*; etc.'' Here the quantifier is binding q.[22]

A fourth and final additional argument comes from *explicit contrasts*. One can directly articulate the contrasts with ''rather than''-clauses. For instance, if one says, ''I know that there is a goldfinch in the garden rather than a raven'', then the value of p is: there is a goldfinch in the garden, and q is: there is a raven in the garden. While if one says, ''I know that there is a goldfinch in the garden rather than a canary'', then the value of p is: there is a goldfinch in the garden, and q is: there is a canary in the garden. Whereas if one says, ''I know that there is a goldfinch in the garden rather than at the neighbor's'', then the value of p is: there is a goldfinch in the garden, and q is: there is a goldfinch at the neighbor's. The ''rather than''-clause is a mechanism of contrastivity. It explicitly articulates q.

The binary surface form of declarative ascriptions may thus be misleading. There are many *precedents* for misleading surfaces. For instance, ''Ann prefers chocolate'' looks to have the binary form: s prefers x. But it should be obvious on reflection that there must be an implicit contrast (to vanilla? to double chocolate chip? to peace on earth?), which is what Ann prefers chocolate *to*. To take another example, ''Rayleigh scattering explains why the sky looks blue'' looks to have the binary form: C explains E. But it has been argued that there must be an implicit contrast (rather than red? rather than violet?), which

[21] Thus Mats Rooth (1992) proposes the *alternative semantics* approach to focus, on which focus adds a semantic marker whose value is a contextually determined set of alternatives. So ''I know that *Mary* stole the bicycle'' gets semantically interpreted as [. . . that $[\text{Mary}]_F$ stole . . .], where $[\text{Mary}]_F$ induces a dual interpretation, one of which is Mary, and the other of which are the other suspects.

[22] The binding test is due to Barbara Partee (1989), and is used extensively by Stanley, who maintains: ''[B]ound readings within a clause are due to the existence of a variable binding operator standing in a certain structural relationship to a co-indexed variable in that clause'' (2000: 412).

is what Rayleigh scattering makes a difference to.[23] Or consider, ''I asked Ann where she was going. Ann answered that she was going to the bar.'' The second sentence looks to have the binary form: *s* answered that *p*. But it should be obvious on reflection that answering is question-relative.

The binary surface form of declarative ascriptions may have misled Moore. When Moore declared, ''I know that I have hands'', perhaps he was misusing the language. Thus Wittgenstein writes: ''[C]an one enumerate what one knows (like Moore)? Straight off like that, I believe not.—For otherwise the expression 'I know' gets misused'' (1969: §6). Wittgenstein suggests that Moore must have ''been thinking of something else in the interim and is now saying out loud some sentence in his train of thought'' (1969: §465; also §§350, 423, 553). Perhaps the preceding train of thought functions to generate a contrast-setting question.[24]

The audience can accommodate Moore by charitably imputing an easy question. For instance, on hearing, ''I know that I have hands'', one might glance to see whether Moore has hands or stumps. Or one might look a bit closer, to see whether he has hands or prostheses. (What *does* one look for?) Perhaps this is why Moorean declarations seem undeniable, yet empty.

Comparison

How does (3) compare to a binary view of encoding? That is, what are the prospects for interpreting various types of knowledge ascription as expressing K*sp*?

[23] Background information: Rayleigh scattering explains why the sky looks blue rather than red, because blue's short wavelength causes it to get scattered around ten times more than longer wavelength colors like red. But Rayleigh scattering does not explain why the sky looks blue rather than violet. In fact, since violet is an even shorter wavelength than blue, Rayleigh scattering predicts that the sky should look violet. What explains why the sky looks blue rather than violet is that our visual system is relatively insensitive to violet. Contrastive views of explanation are defended by Bas van Fraassen (1980), Alan Garfinkel (1981), and Peter Lipton (1991), *inter alia*.

[24] Revealingly, Moore himself uses focused and overtly contrastive ascriptions in key passages. He begins his ''A Defence of Common Sense'' with the focused ascription that he knows ''that there exists at present a living human body which is *my* body'' (1959*a*: 33). And he begins ''Certainty'' by listing his convictions in contrastive format: ''I am at present, as you all can see, in a room and not in the open air; I am standing up, and not either sitting or lying down; I have clothes on, and am not absolutely naked; I am speaking in a fairly loud voice, and am not either singing or whispering or keeping quite silent;'' (1959*b*: 227). Perhaps it is here that Moore captures the content of common sense knowledge.

I suspect that binary views face systematic problems with respect to all types of knowledge ascription. (Here I continue to focus on *invariantist* binary views, postponing discussion of contextualism until §6.) Consider the interrogative ascription: ''Ann knows whether there is a goldfinch or a raven in the garden.'' The natural way to chop this ascription to fit the Procrustean bed of K*sp*, is to treat *p* as: there is a goldfinch in the garden. In general, the natural way to fit interrogative ascriptions into the binary mold is to treat them as expressing K*sp*, where *p* is the true answer to the question posed by the *wh*-clause.[25]

The binary treatment of interrogative ascriptions, though, is counter-intuitive. It implies that ''Ann knows whether there is a goldfinch or a raven in the garden'', ''Ann knows whether there is a goldfinch or a canary in the garden'', and ''Ann knows whether there is a goldfinch in the garden or at the neighbor's'' all express the same proposition. (Or at least, that all have the same truth conditions). When intuitively these can differ in truth-value.[26]

My aim is to develop a contrastive view, not to refute binary views. Perhaps the binary theorist can find some devious strategy to encode interrogative ascriptions (similar issues arise with respect to the other types of ascription). But I would suggest, at this point, that (3) supplies *the more natural code* for the full range of question-relative knowledge ascriptions.

[25] Thus Higginbotham proposes the rule: ''know $(x,\hat{}\boldsymbol{\pi}) \leftrightarrow (\exists p)$ (know(x,p) & p answers $\boldsymbol{\pi}$)'' (1996: 381). Instances of this rule are implicit in Hintikka's treatment of ''knows who'', Lewis's treatment of ''knows whether'', and Stanley and Williamson's treatment of ''knows how''. Thus, for Hintikka, ''*a* knows who *b* is'' is analyzed as: $(\exists x)$ *a* knows that $(b = x)$ (1975*b*: 4). For Lewis, ''Holmes knows whether . . . if and only if he knows the true one of the alternatives presented by the 'whether'-clause, whichever one that is'' (1982: 194). And for Stanley and Williamson, ''Hannah knows how to ride a bicycle'' is ''true if and only if, for some contextually relevant way *w* which is a way for Hannah to ride a bicycle, Hannah knows that *w* is a way for her to ride a bicycle''. From which they conclude: ''Thus, to say that someone knows how to *F* is always to ascribe them knowledge-that'' (2001: 426).

[26] A less natural possibility is to transform *p* into a big conditional. Here ''Ann knows whether there is a goldfinch or a raven in the garden'' is to be transformed (somehow) into: ''Ann knows that if (there is a goldfinch or a raven in the garden), then there is a goldfinch in the garden.'' But this gives the wrong truth-value *when all the options are false*. For instance, ''Moore knows whether he has tentacles or flippers'' seems false, since Moore has neither tentacles nor flippers. But the 'corresponding' conditional is: K*m*$((p \vee q) \supset p)$, where *p* is: that Moore has tentacles, and *q* is: that Moore has flippers. And this knowledge claim is *true* (or at least the binary theorist should think it true), since Moore should know that the antecedent of the conditional is false, and Moore knows that conditionals with false antecedents are true.

4. KNOWLEDGE

The fourth stage of an account of knowledge is to analyze or otherwise *illuminate* the relation. What is *knowledge*? I propose:

(4) K*spq* iff: (i) *p*, (ii) *s* has proof that *p* rather than *q*, and (iii) *s* is certain that *p* rather than *q*, on the basis of (ii).

I should emphasize from the outset that (4) is the least important and least promising part of the contrastive view. It is the least important insofar as K*spq* is compatible with virtually any analysis of knowledge (even none at all). And it is the least promising insofar as the history of philosophical analyses suggests that counterexamples are inevitable. Thus (4) is merely intended as a *useful gloss*.

Clarifications

Overall, (4) is a contrastive implementation of the contextualist idea that knowledge is the elimination of relevant alternatives (Austin 1946; Dretske 1981; Lewis 1996; Ram Neta 2002).

Piecewise, the first condition is the *truth* condition. (Note that since *p* and *q* are mutually exclusive, *p*'s truth implies *q*'s falsity.)

The second condition is a contrastive interpretation of *justification*. It is a form of *restricted infallibilism* about evidence. It is infallibilist insofar as it requires proof, which is conclusive evidence, evidence that could not possibly obtain without *p* being true. But it is restricted insofar as the space of possibilities open to disproof is restricted to: $\{p\} \cup \{q\}$.[27]

The third condition is a contrastive interpretation of *belief* (plus a provision that belief and justification must be appropriately related via *basing*[28]). It is a form of *restricted indubitabilism* about belief. It is indubitabilist insofar as it requires certainty, which is an absence of any

[27] I have not said what *evidence* consists in, nor whether the notion can be reduced. Though what I say is compatible with Lewis's (1996) conception of one's evidence as one's total experience. Lewis defines *elimination* as follows: possibility *p* is eliminated for *s* (at *t*) iff *p* is inconsistent with *s*'s total experience *e* (at *t*). *S* has conclusive evidence that *p* rather than *q*, on this interpretation, iff *q* is eliminated for *s*. (Notice that the actuality possibility cannot be eliminated; thus *p*, if true, is ineliminable.)

[28] Basing is a hybrid of *causation* and *rationality*: one's proof must be a rationalizing, non-deviant cause of one's certainty. For further discussion of basing see Keith Allen Korcz (2000).

doubt that p is true. But it is restricted insofar as the space of possibilities open to doubt is restricted to: $\{p\} \cup \{q\}$.

Arguments

First, (4) fits (1) by *comprising the ability to answer*. That is, the analysis in (4) is the right form for the task of fingering answerers as per (1), because to meet (4) *is* to be an answerer. In this way, (4) implements Hector-Neri Castañeda's idea that, ''knowledge involves essentially the non-doxastic component of a power to answer a question'' (1980: 194).

The first condition, the truth condition, is required to fit (1). That is, being able to select the truth is a necessary condition on being able to answer the question. Questions with no true alternatives involve false *presuppositions*,[29] and ought to be rejected rather than answered.

The second condition, the contrastive justification condition, is also required to fit (1)—having proof for p rather than q is a necessary condition on being able to answer: $p \vee q$? As long as one's evidence is compatible with multiple queried alternatives, the inquiry cannot be concluded. This comports with the methodological insight of Sherlock Holmes: ''It is an old maxim of mine that when you have excluded the impossible, whatever remains, however improbable, must be the truth'' (*The Adventure of the Beryl Coronet*).

The third condition, the contrastive belief condition, is also required to fit (1)—being certain that p rather than q is a necessary condition on being able to answer: $p \vee q$? As long as one is in doubt, the inquiry is still open. This comports with the Peircean view of doubt as the irritant that spurs inquiry. (The basing relation is required as well: if one's certainty is not based on the proof, then the inquiry has not been closed on proper grounds.)

Perhaps meeting all three conditions is still insufficient for being able to answer. But what could be lacking? Imagine taking a multiple choice exam, having proof that all but one answer is wrong, and being certain of the true answer on this basis. What could be lacking, as far as knowing the answer?

The second argument for (4) is that it resolves numerous problem cases in the literature, including *lottery cases* and *Gettier cases*, via

[29] Question Q presupposes proposition p iff p is entailed by all answers to Q (Belnap and Steel 1976).

restricted infallibilism. Lottery cases beg for infallibilism: the ticket holder does not know in advance that her ticket will lose rather than win, no matter how long the odds, because her evidence remains fallible—she might be wrong, she might win, she does not know that she will lose. Gettier cases also beg for infallibilism: the passerby who sees a clock stopped twenty-four hours ago on 3p.m. does not know that it is now 3p.m. rather than 4p.m., despite some evidence for a true belief, because his evidence remains fallible—he might be wrong, the clock might be off, he does not know what time it is. Here the fallibility of the connection between evidence and truth is what opens up the possibility of a merely accidental correlation.[30] (Such an infallibilism does not induce skepticism, since the infallibilism is restricted. Knowledge is still possible, when the alternatives in *q* are eliminable.)

Objections

First, (4) faces *the problem of the giveaway question*. The giveaway question arises when *p* and *q* are both dubious hypotheses for *s*, *p* is luckily true, and *q* is easily eliminable. For instance, suppose that Poirot can prove that it was Mayerling who was murdered, but has no evidence that it was Darrow who did the deed. Then, on (4), Poirot can count as knowing that Darrow killed Mayerling rather than that Darrow killed Japp. Yet intuitively, it might seem that Poirot knows nothing of the sort—he need not even know who Darrow is.

In reply, perhaps Poirot does know that Darrow killed Mayerling rather than Japp. After all, if Poirot were to engage the question, ''Did Darrow kill *Mayerling*, or *Japp*?'', he would be able to answer properly—he can eliminate all but one option. Poirot would pass the test. This is an epistemic achievement. The knowledge claim marks this achievement. It distinguishes Poirot's epistemic standing from that of Poirot's sidekick Hastings, who does not even know who was murdered. Poirot at least knows that it was *Mayerling* rather than *Japp* who Darrow murdered.[31] Or try: Poirot knows whether Darrow killed *Mayerling* or *Japp*.

[30] For further discussion of the restricted infallibilist solution to lottery and Gettier cases, see Lewis (1996), Stewart Cohen (1998*a*), and Mark Heller (1999).

[31] In this vein, Johnsen imagines that Milan Kundera might just happen to be in Ventimiglia, and claims that he (Johnsen) would at least know that Kundera is in Ventimiglia rather than *Johnsen's office* (2001: 405).

A second reply (which I reserve as backup) would be to add a further condition to (4). The most natural addition would require some sort of *positive evidence for* p. This would entail that Poirot does not know that Darrow murdered Mayerling rather than Japp, on grounds that Poirot lacks evidence for the proposition that Darrow killed Mayerling. Here there is room to explore a mixture of fallibilism and infallibilism, on which *s* must have infallible evidence that *p* rather than *q*, plus fallible evidence that *p*. I leave this for further exploration.[32] As indicated above, I am merely aiming for a useful gloss here.

Second, one might object that (4) *induces skepticism*. The contrastivist promises to resist skepticism, by allowing Moore to know that he has hands rather than stumps. But, the objection runs, (4) does not allow for this, since there are stump-possibilities that Moore cannot eliminate, such as possibilities in which Moore has stumps but is dreaming of hands, or has stumpy arms stapled onto his envatted brain. Thus, the objection concludes, (4) disallows knowledge.

In reply, there are possibilities that Moore can eliminate, which is what (4) requires for knowledge. Here it will help to leave the shifty 'that'-clauses of English behind, and speak directly of the worlds they denote. There are plenty of worlds that Moore can eliminate, including worlds in which he veridically perceives his stumps. And there are plenty of worlds that Moore cannot eliminate, including actuality and its skeptical variants. In general, for any subject *s* and true proposition *p*, *s* will have a *discriminatory range R* over *p*, where *R* is the union of those $\sim p$-worlds which *s* is able to discriminate from actuality. For all nonempty subsets *R*- of *R*, *s* is in a position to know that *p* rather than that *R*- obtains. Whereas for all nonempty subsets *S*— of the complement of *R*, $\sim$K*spS*— holds.

So does Moore know that he has hands rather than stumps? *Yes*, in a sense. What Moore knows can be more fully described as follows: he knows that he has hands rather than *stumps that are apparent*. Or more fully: Moore knows that he has hands rather than *stumps that he would veridically perceive*. Fuller descriptions are always available. Which worlds these descriptions denote is contextually variable. Thus, strictly speaking, what follows from (4) is that ''Moore knows that he has hands rather than that he has stumps'' is *true* in contexts in which ''that he has

[32] Dretske expresses some ambivalence on this point, saying that the subject, ''needs only a nominal justification, if he needs any justification at all'' for the non-contrasted aspect of the knowledge claim (1981: 373).

stumps'' denotes worlds within Moore's discriminatory range *R*. The context-invariant truth is of the form: Moore knows $\{w_\alpha\}$ rather than $\{w_1, w_2, \ldots, w_m\}$.

Comparison

How does (4) compare to various binary views of knowledge? If the task is to provide a finite, non-circular, and intuitively fitting set of necessary and sufficient conditions, all views may prove equally hopeless. If the task is merely to provide a useful gloss of a relation (a decent approximation), perhaps (4) proves best.

The advantage of (4), shared only by some versions of contextualism, is the ability to steer between, ''the rock of fallibilism and the whirlpool of skepticism'' (Lewis 1996: 221), by implementing a restricted infallibilism. This is an advantage insofar as fallibilism is *implausible*, *arbitrary*, and *lottery-wracked*. Fallibilism is implausible insofar as it licenses the breathtaking conjunction: ''I might be wrong, though I still know.'' Fallibilism is arbitrary insofar as any line of evidence (or shading of a penumbra) below 1 is arbitrary. Fallibilism is lottery-wracked insofar as any line below 1 will be exceeded by evidence that, in a suitably large lottery, a given ticket is a loser. Implementing a restricted infallibilism is also an advantage insofar as unrestricted infallibilism is *skeptical*. These points are all controversial, and I cannot defend them here. This is left to the reader's judgment. But I would suggest, for these reasons, that (4) offers the *more illuminating gloss* of knowledge, rivaled only by contextualism.

5. SKEPTICISM

The fifth and final stage of an account of knowledge is to resolve outstanding *paradoxes*. How does contrastive knowledge *help*? I propose:

(5) Contrastive knowledge resolves the closure paradox.

Paradox

The closure paradox is typically formulated in binary terms, as follows:

(C1) Moore knows that he has hands.
(C2) Moore doesn't know that he is not a brain-in-a-vat.

(C3) If Moore doesn't know that he is not a brain-in-a-vat, then he doesn't know that he has hands.[33]

These premises are individually plausible, but conjointly contradictory.

There are four main replies to the closure paradox from within a binary framework: the *skeptic* denies C1, the *dogmatist* denies C2, the *denier of closure* denies C3, and the *contextualist* denies that C2 and C3 entail the falsity of C1 (by maintaining that the denotation of ''knows'' shifts, rendering the argument equivocal). These positions have been extensively debated.[34] So I will simply state what I find objectionable about each position, to set the stage for the contrastive solution.

I object to skepticism and dogmatism on two parallel counts. First, the denials of C1 and C2 strike me as *absurd*. At least, some explanation is needed of their plausibility. Second, skepticism and dogmatism *collapse distinctions*.[35] Suppose that Student, Assistant, and Professor are visiting the zebras at the zoo. Student is remarkably ignorant, and can't even discern a zebra from a mule; Assistant can discern a zebra from a mule by its stripes, but cannot discern a zebra from a cleverly painted mule; Professor can discern a zebra even from a cleverly painted mule by anatomical features that no mere paint job can disguise. The skeptic confuses Student with Assistant, denying that either knows that the beast is a zebra, since neither can eliminate the painted mule hypothesis. The dogmatist confuses Assistant with Professor, maintaining that both know that the beast is a zebra, since both can eliminate the unpainted mule hypothesis. Both skepticism and dogmatism thereby distort partial knowledge.[36]

[33] This formulation is found in Keith DeRose (1995) and Stephen Schiffer (1996), *inter alia*. See Peter Unger (1975) for arguments that this is the *root* skeptical argument. See Anthony Brueckner (1994), Cohen (1998*b*), and Jonathan Vogel (n.d.) for further discussion of how closure relates other skeptical concerns such as underdetermination.

[34] For a defense of skepticism, see Unger (1975); for a defense of dogmatism, see Peter Klein (1981), Ernest Sosa (1999), and James Pryor (2000); for a defense of the denial of closure, see Dretske (1971) and Robert Nozick (1981); for a defense of contextualism, see G. C. Stine (1976), Cohen (1988 1999), DeRose (1995), Lewis (1996), and Neta (2002).

[35] Heller levels this criticism at the skeptic: ''[Skeptical] standards fail to draw the distinctions that are important to us. Even though neither my wife nor I can rule out the possibility of an evil genius deceiving us about where the leftovers are, she is in a better epistemic position than I am'' (1999: 119).

[36] Though see Schaffer (2004*b*) for a defense of the skeptic from these objections. Overall, I would rate skepticism the second-best option.

I object to the denial of closure on two counts. First, the denial of C3 seems *absurd*, at least without some explanation of its plausibility. Second, denying closure *collapses inferences*. Surely deduction transmits knowledge. How could it not, given that our ultimate epistemic interest is truth, and deduction preserves truth? How could it not, given that mathematical proof is deductive and mathematical proof yields knowledge? Pending a replacement for C3, the anti-closure view cripples knowledge.[37]

I object to contextualist solutions on four counts. First, the compatibility of C1 and C2 seems *absurd*, at least without some explanation of the appearance of incompatibility.[38]

Second, the way that C1 and C2 are rendered compatible is *overly concessionary* to both skepticism and dogmatism. For the contextualist concedes that dogmatism holds in the courtroom, so that there one can count as knowing that one is not a brain-in-a-vat. But surely one can never know so much. And the contextualist concedes that skepticism holds in the classroom, so that there one cannot count as knowing that one has hands. But surely one can never know so little. Thus the contextualist is stuck with the implausibilities of both views, and their subsequent conflations. In any given context, the contextualist must either confuse Student with Assistant, or Assistant with Professor. In no context can the contextualist successfully distinguish all three.

Third, the contextualist machinery turns our knowledge attributions *manic*. The contextualist swings from highs of dogmatism to lows of skepticism, at the mere drop of a skeptical scenario. Surely our dispositions to ascribe knowledge are more stable (Johnsen 2001: 395; see also Dretske 1991: 192; Richard Feldman 1999: 106).

Fourth, contextualism renders ''knows'' too shifty to score inquiry consistently (§2). Scoring inquiry requires being able to evaluate how a subject performs through a sequence of questions. This requires having epistemic vocabulary that can keep a consistent score through a range of contexts. But ''knows'' as the contextualist conceives it cannot keep a consistent score, because ''knows'' as the contextualist conceives it is continually warped by the present context.

[37] See Williamson for a defense of closure based on the idea that ''deduction is a way of extending one's knowledge'' (2000: 117). For extended discussion see Hawthorne (2004: 31–50).

[38] As Schiffer notes in a criticism of the contextualist solution, ''If that's the solution, what the hell was the *problem*?'' (1996: 329).

Resolution

The contrastivist rejects the closure paradox as formulated, since C1–C3 all concern binary knowledge. I will now argue, on behalf of (5), that contrastivism (i) dissolves the paradox, (ii) explains the plausibility of its premises, and (iii) answers all the objections leveled above at the other approaches.

Contrastivism dissolves the paradox by revealing how ordinary knowledge and skeptical doubt are compatible: they concern different contrasts. Moore knows that he has hands rather than stumps. Moore does not know that he has hands rather than vat-images of hands. In interrogative terms, Moore knows whether he has hands or stumps, but does not know whether he has hands or vat-images of hands. In general, for any subject s and proposition p, s is in position to know that p rather than q for any proposition q within s's discriminatory range (§4). Whereas for any q that extends beyond s's discriminatory range, $\sim Kspq$.

Some of the inferential relations that hold between contrastive knowledge states can be adduced from the notion of discriminatory range. A valid schema will preserve discrimination of truth. It will preserve the elimination of all-but-p. Here are two valid schemas:

Expand-p: if $p_1 \rightarrow p_2$ then $Ksp_1q \rightarrow Ksp_2q$[39]
Contract-q: if $q_2 \rightarrow q_1$ then $Kspq_1 \rightarrow Kspq_2$

And here are four *in*valid schemas, which do not preserve discrimination of truth:

*Contract-p: if $p_2 \rightarrow p_1$ then $Ksp_1q \rightarrow Ksp_2q$
*Expand-q: if $q_1 \rightarrow q_2$ then $Kspq_1 \rightarrow Kspq_2$
*Replace-p: $Ksp_1q \rightarrow Ksp_2q$
*Replace-q: $Kspq_1 \rightarrow Kspq_2$.

[39] These schemas are only valid as *idealizations*. Expand-p, for instance, needs limitation to prevent the p-worlds from *swallowing* q. Contrastivity needs to be preserved under p-expansion. So a more accurate statement of expand-p would be: if $p_1 \rightarrow p_2$ and $p2 \cap q = \varnothing$, then $Ksp_1q \rightarrow Ksp_2q$. Expand-$p$ should also be restricted to cases of competent deduction (here I am following Williamson 2000). So an even more accurate statement of expand-p would be: if (i) $p_1 \rightarrow p_2$, (ii) $p_2 \cap q = \varnothing$, (iii) s competently deduces p_2 from p_1, and (iv) s comes to be certain that p_2 rather than q on the basis of (iii), then $Ksp_1q \rightarrow Ksp_2q$. These details won't matter for what follows.

Since Replace-q is invalid, one cannot use the fact that Moore knows that he has hands rather than stumps to infer that Moore knows that he has hands rather than vat-images of hands. The fact that the vat possibility lies outside Moore's discriminatory range does not entail that the stumps possibility does too.

Ordinary knowledge concerns discriminations in a limited range. Skeptical doubts reveal the limits of that range. Since the existence of possibilities outside one's discriminatory range does not imply the absence of any possibilities inside that range, skeptical doubts do not imply any absence of ordinary knowledge. Thus ordinary knowledge and skeptical doubt are compatible. Paradox dissolved.

Why then are the premises of the paradox so plausible? The contrastivist explanation is that (i) we charitably accommodate binary knowledge ascriptions by imputing a question (§3), and (ii) the natural questions for C1–C3 in fact generate contrastive truths. Starting with C1, the natural question would concern whether Moore has hands or is some sort of amputee. Indeed, the only implicit questions for C1 that would generate falsity would be those concerning skeptical scenarios, supplying of which would be both unnatural and unaccommodating. In the case of C2, the implicit question that leaps out concerns whether Moore is handed or envatted. Since Moore cannot discriminate between these alternatives, we naturally assent to C2. And finally in the case of C3, we naturally interpret it as embedded in an inquiry that concerns whether Moore is handed or envatted. So we naturally think of C3 as saying: if Moore does not know that he's not a brain-in-a-vat rather than a brain-in-a-vat, then he doesn't know that he's a hand-owner rather than a brain-in-a-vat. This has the form: $\sim Ksp_1 \sim p_1 \rightarrow \sim Ksp_2 \sim p_1$, where $p_2 \rightarrow p_1$ (hands entails not-vatted). This is a valid inference, as it is an instance of the contrapositive of Expand-p.

Putting this together, the contrastive reformulation of closure is:

(C1′) Moore knows that he has hands rather than stumps.
(C2′) Moore does not know that he is handed rather than envatted.
(C3′) If Moore doesn't know that he's not envatted rather than envatted, then he doesn't know that he's handed rather than envatted.[40]

[40] This is the reformulation that preserves the truth of each premise. Alternatively the paradox could be reformulated so as to preserve the incompatibility of the premises via: C3′*: If Moore doesn't know that he's handed rather than envatted, then he doesn't know

To put the reformulation in interrogative terms:

(C1″) Moore knows whether he has hands or stumps.
(C2″) Moore does not know whether he has hands or is envatted.
(C3″) If Moore does not know whether he is non-envatted or envatted, then he doesn't know whether he is handed or envatted.

Each premise is true. There is no paradox. The plausibility of each of C1–C3 is due to our naturally processing them as something like C1′–C3′ (equivalently: C1″–C3″) respectively.

Contrastivism, finally, answers all the objections leveled above against skepticism, dogmatism, the denial of closure, and contextualism. With respect to skepticism and dogmatism, contrastivism explains the plausibility of C1 and C2, as per the previous paragraph. And contrastivism captures the distinctions that skepticism and dogmatism collapse. Student does not know that the beast is a zebra rather than a mule. Assistant knows that the beast is a zebra rather than a mule, but does not know that the beast is a zebra rather than a painted mule. Professor knows that the beast is a zebra rather than a mule, and that the beast is a zebra rather than a painted mule. What distinguishes these characters is their discriminatory ranges.

With respect to the denial of closure, contrastivism explains the plausibility of C3, as above. And contrastivism captures the inferences that the denier of closure disallows, via Expand-*p* and Contract-*q*. In particular, Expand-*p* preserves the sense in which deductive proof is knowledge-transmitting.

With respect to contextualism, the contrastivist can explain the apparent incompatibility of C1 and C2 as due to neglect of the covert contrast variable. And covert variables can induce confusion among competent speakers. The compatibility of C1′ and C2′ allows the contrastivist to avoid conceding dogmatism in one context and skepticism in another, as the contextualist must. Ordinary knowledge and skeptical doubt do not need to be cordoned off into separate contexts. They coexist in both the courtroom and the classroom. Moore always knows that he has hands rather than stumps, and never knows that he has

that he has hands rather than stumps. But C3′* is false—just because ''Hands or vat-images of hands?'' falls beyond Moore's discriminatory range does not imply that ''Hands or stumps?'' does too.

hands rather than vat-images of hands. The context-invariance of C1′ and C2′ provides the stability that contextualism precludes. The invocation of skeptical possibilities does not change which discriminations *s* can make one whit. Thus one can track *s*'s discriminatory range through a sequence of questions, and thereby properly keep score of inquiry.

Comparison

Contrastivism reveals that the closure paradox is *an artifact of binarity*. Contrastivism provides the following recipe for binary paradoxes. First, find an easy question that *s* can successfully answer by *p*. This will generate a context in which ''*s* knows that *p*'' encodes a true proposition: Kspq_1. Treat this as binary knowledge: K*sp*. Second, find a hard question that *s* cannot answer involving *p*. This will generate a context in which ''*s* knows that *p*'' encodes a false proposition: Kspq_2. Treat this as binary ignorance: ∼K*sp*. Third, conjoin and tremble. Skeptical scenarios merely help provide hard questions for the second step (''Or has she just dreamt the whole episode?'')

For all we philosophers might fret over skepticism, ordinary inquiries never shipwreck on skeptical possibilities. No court case has ever been dismissed due to the closure paradox (''Your Honor, that witness knows nothing!''). Ordinary inquiries succeed because ordinary questions are restricted. *The wile of the skeptic is to shift the question*. Thus resolving the closure paradox requires rendering knowledge in a structure that logs the question: the contrastive structure.

6. CONTEXTUALISM

Epistemic contrastivism is cousin to the family of *epistemic contextualisms*. It might prove useful, by way of epilogue, to clarify the relations.[41]

Contextualisms feature three main family traits, which I label *indexicalism*, *relevantism*, and *equivocationism*. Indexicalism is the thesis that ''knows'' functions like an indexical in having a stable character but a context-dependent content. Relevantism is the thesis that what one knows is determined by a set of relevant alternatives. Equivocationism

[41] See Schaffer (2004*a*) for a more extended discussion of these issues.

is the thesis that the closure paradox involves an equivocation between the contents of ''knows'' generated by the first two premises (§5).[42] To clarify the relations between contrastivism and the family of contextualisms, it will prove most helpful to compare contrastivism to indexicalism, relevantism, and equivocationism directly, as separate positions.

Indexicalism

Contrastivism and indexicalism are similar in the following way. On both theories, a binary knowledge ascription may be true in one context, and false in another.

But contrastivism and indexicalism differ in two main ways. First, the mechanism of context-dependence is different. With indexicalism, it is the content of the relation denoted by ''knows'' that is contextually shifty. With contrastivism, it is the value of the contrast relatum q that is shifty. Thus indexicalism, but not contrastivism, is committed to the postulation of context-dependence without representation in logical form.[43]

Second, the extent of context-dependence is different. With indexicalism, since it is the occurrence of the term ''knows'' that induces shiftiness, every knowledge ascription must be shifty. With contrasti-

[42] While more recent contextualisms (such as DeRose 1995; Lewis 1996) exhibit indexicalism, relevantism, and equivocationism together, these traits are independent. Indexicalism does not entail relevantism, since the context-dependence of ''knows'' might turn on something other than relevance, such as the degree of justification required by the stakes. Cohen (1988) is perhaps best read this way. And indexicalism does not entail equivocationism, since, for instance, ''knows'' might not be variable enough for skeptical doubts. DeRose (1995) allows though does not endorse this position. Relevantism does not entail indexicalism, since relevance might be determined purely in terms of s's objective situation, with no reference to the context of utterance. Dretske (1991) and Hawthorne (2004) endorse this view. And relevantism does not entail equivocationism, since, for instance, skeptical possibilities might never be relevant. Austin (1946) takes this line. Equivocationism, finally, does not entail either indexicalism or relevantism, since the equivocation might be due to polysemy (with neither sense indexicalized or involving a relevance function). Norman Malcolm's (1952) distinction between the ''strong'' and ''weak'' senses of ''knows'' might serve as a prototype for such a view.

[43] Stanley (2000) argues that it is implausible to postulate context-dependence that is unrepresented in logical form, except for the cases of the obvious indexicals, demonstratives, and pronouns. Stanley's argument applies against indexicalism but not contrastivism. There are plenty of precedents (including ''prefers'' and ''explains'': §3) for verbs with additional contrast slots, while there seem to be no precedents for verbs that are indexicalized.

vism, since it is the value of *q* that is shifty in binary ascriptions, interrogative, noun, and overtly contrastive ascriptions must be relatively stable, since these at least partially fix the value of *q*. This seems intuitively correct: ''Moore knows that he has hands'' seems shiftier than ''Moore knows that he has hands rather than stumps''. Further, this stable form of knowledge ascription is required by the scorekeeping function of knowledge (§5).

Relevantism

Contrastivism and relevantism are similar in the following way. On both theories, whether one knows is calculated with reference to a set of alternatives.

But contrastivism and relevantism differ in two main ways. First, what one knows is different. With relevantism, by eliminating the relevant alternatives, one knows that *p*. With contrastivism, one knows that *p* rather than *q*. The relevantist is still in the grip of binarity.

Second, the alternatives are generated in different ways. With relevantism, the alternatives are generated by a relevance function. With contrastivism, the alternatives are generated by an explicit or implicit question (§3). But what is 'relevance'? By far the best account of relevance is to be found in Lewis (1996).[44] But Lewis's account is subject to counterexamples (see Vogel 1999). Worse, it is (i) *imprecise*, (ii) *epistemically tailored*,[45] and (iii) *ad hoc* in certain respects (such as why resemblance with respect to evidence is non-salient). The contrastivist mechanisms (§3), on the other hand, are (i) *precise*, (ii) *linguistically general* mechanisms, and their application is (iii) *motivated* by the role of knowledge in inquiry.

Equivocationism

Contrastivism and equivocationism are similar in the following way. On both theories, ordinary knowledge and skeptical doubts are compatible.

[44] Lewis's account may be the only serious account of relevance. Dretske (1981: 373–7) makes a number of programmatic remarks, but otherwise one finds little of substance on this topic in the entire literature. Not for nothing does Sosa warn that relevantism ''will remain unacceptably occult'' (1986: 585). See also Vogel (1999).

[45] Lewis begins by invoking the linguistic mechanism of *quantifier domain restriction*. This much is linguistically general. But then most of Lewis's subsequent rules of ignoring are epistemically tailored.

But contrastivism offers a better solution to the closure paradox in four main ways (§5): (i) contrastivism provides a better explanation of the apparent incompatibility of ordinary knowledge and skeptical doubt; (ii) contrastivism avoids conceding dogmatism in some contexts and skepticism in the others, by allowing ordinary knowledge and skeptical doubts to be compatible *in the same context*: ''Moore knows whether he has hands or stumps; but he does not know whether he has hands or vat-images of hands''; (iii) contrastivism avoids manic swings from dogmatism to skepticism thereby; and (iv) contrastivism allows ''knows'' to serve its inquiry-scoring function, since one can keep a consistent score through a range of contexts. Assistant can successfully answer the question: ''Zebra or [normal] mule?'' After it emerges that Assistant cannot answer the question: ''Zebra or painted mule?'', one can still report Assistant's previous success: ''At least he knows whether the beast is a zebra or a normal mule.''[46]

REFERENCES

Austin, J. L. (1946) 'Other Minds', *Proceedings of the Aristotelian Society*, 20: 149–87.

—— (1962) *How to Do Things with Words* (Cambridge).

Belnap, Nuel, and Thomas Steel (1976) *The Logic of Questions and Answers* (New Haven).

Boër, Steven, and William Lycan (1986) *Knowing Who* (Cambridge).

Brandom, Robert (1994) *Making it Explicit* (Cambridge).

Brueckner, Anthony (1994) 'The Structure of the Skeptical Argument', *Philosophy and Phenomenological Research*, 54: 827–35.

Castañeda, Hector-Neri (1980) 'The Theory of Questions, Epistemic Powers, and the Indexical Theory of Knowledge', in Peter French, Theodore Uehling, Jr., and Howard Wettstein (eds.), *Midwest Studies in Philosophy*, v. *Studies in Epistemology* (Minneapolis), 193–238.

Chomsky, Noam (1977) *Essays on Forms and Interpretation* (Amsterdam).

Cohen, Stewart (1988) 'How to be a Fallibilist', *Philosophical Perspectives*, 2: 91–123.

[46] Thanks to Kent Bach, Martijn Blaauw, Thomas Blackson, John Collins, Fred Dretske, Tamar Gendler, Ed Gettier, John Hawthorne, Mark Heller, Michael Huemer, Bredo Johnsen, Antti Karjalainen, David Lewis, Adam Morton, Roald Nashi, Ram Neta, Barbara Partee, Walter Sinnott-Armstrong, Jonathan Vogel, Brian Weatherson, and audiences at the Bellingham Summer Philosophy Conference, the Pacific APA, the University of Colorado, and the University of Florida.

——(1998*a*) 'Contextualist Solutions to Epistemological Problems: Scepticism, Gettier, and the Lottery', *Australasian Journal of Philosophy*, 76: 289–306.

——(1998*b*) 'Two Kinds of Skeptical Argument', *Philosophy and Phenomenological Research*, 52: 143–59.

——(1999) 'Contextualism, Skepticism, and the Structure of Reasons', *Philosophical Perspectives*, 13: 57–89.

Craig, Edward (1990) *Knowledge and the State of Nature: An Essay in Conceptual Synthesis* (Oxford).

DeRose, Keith (1995) 'Solving the Skeptical Problem', *Philosophical Review*, 104: 1–52.

Dewey, John (1938) *Logic: The Theory of Inquiry* (New York).

Dretske, Fred (1970) 'Epistemic Operators', *Journal of Philosophy*, 67: 1007–23.

——(1981) 'The Pragmatic Dimension of Knowledge', *Philosophical Studies*, 40: 363–78.

——(1991) 'Knowledge: Sanford and Cohen', in Brian McLaughlin (ed.), *Dretske and his Critics* (Oxford), 185–96.

Feldman, Richard (1999) 'Contextualism and Skepticism', *Philosophical Perspectives*, 13: 91–114.

Garfinkel, Alan (1981) *Forms of Explanation: Rethinking the Questions in Social Theory* (New Haven).

Greco, John (2002) 'Knowledge as Credit for True Belief', in Linda Zagzebski and Michael DePaul (eds.), *Intellectual Virtue: Perspectives from Ethics and Epistemology* (Oxford), 111–34.

Groenendijk, Jeroen, and Martin Stokhof (1997) 'Questions', in Johan F. A. K. van Benthem and Alice G. B. ter Meulen (eds.), *Handbook of Logic and Language* (Amsterdam), 1055–124.

Hamblin, C. L. (1958) 'Questions', *Australasian Journal of Philosophy*, 36: 159–68.

Hawthorne, John (2004) *Knowledge and Lotteries* (Oxford).

Heim, Irene (1979) 'Concealed Questions', in R. Bäuerle, U. Egli, and A. von Stechow (eds.), *Semantics from Different Points of View* (Berlin), 51–60.

Heller, Mark (1999) 'Contextualism and Anti-Luck Epistemology', *Philosophical Perspectives*, 13: 115–29.

Higginbotham, James (1993) 'Interrogatives', in Kenneth Hale and Samuel Jay Keyser (eds.), *The View from Building 20: Essays in Honor of Sylvain Bromberger* (Cambridge), 195–228.

——(1996) 'The Semantics of Questions', in Shalom Lappin (ed.), *The Handbook of Contemporary Semantic Theory* (Oxford), 361–83.

Hintikka, Jaakko (1975*a*) 'Answers to Questions', *The Intensions of Intentionality and Other New Models for Modalities* (Dordrecht), 137–58.

Hintikka, Jaakko (1975*b*) 'Different Constructions in Terms of the Basic Epistemological Verbs: A Survey of Some Problems and Proposals', *The Intensions of Intentionality and Other New Models for Modalities* (Dordrecht), 1–25.

—— (1981) 'On the Logic of an Interrogative Model of Scientific Inquiry', *Synthese*, 47: 69–84.

Hookway, Christopher (1996) 'Questions of Context', *Proceedings of the Aristotelian Society*, 96: 1–16.

Johnsen, Bredo (2001) 'Contextualist Swords, Skeptical Plowshares', *Philosophy and Phenomenological Research*, 62: 385–406.

Klein, Peter (1981) *Certainty: A Refutation of Skepticism* (Minneapolis).

Kleiner, Scott (1988) 'Erotetic Logic and Scientific Inquiry', *Synthese*, 74: 19–46.

Korcz, Keith Allen (2000) 'The Causal-Doxastic Theory of the Basing Relation', *Canadian Journal of Philosophy*, 30: 525–50.

Laudan, Larry (1977) *Progress and its Problems: Towards a Theory of Scientific Growth* (Berkeley, Calif.).

Levi, Isaac (1984) 'Abduction and Demands for Information', *Decisions and Revisions: Philosophical Essays on Knowledge and Value* (Cambridge).

Lewis, David (1979) 'Scorekeeping in a Language Game', *Journal of Philosophical Logic*, 8: 339–59.

—— (1982) 'Whether Report', in Tom Pauli (ed.), *Philosophical Essays Dedicated to Lennart Åqvist on his Fiftieth Birthday* (Uppsala), 194–206.

—— (1996) 'Elusive Knowledge', *Australasian Journal of Philosophy*, 74: 549–67.

Lipton, Peter (1991) *Inference to the Best Explanation* (New York).

Malcolm, Norman (1952) 'Knowledge and Belief', *Mind*, 51: 178–89.

Moore, G. E. (1959*a*) 'A Defence of Common Sense', *Philosophical Papers* (London), 32–59.

—— (1959*b*) 'Certainty', *Philosophical Papers* (London), 227–51.

Morton, Adam, and Antti Karjalainen (2003) 'Contrastive Knowledge', *Philosophical Explorations*, 6: 74–89.

Neta, Ram (2002) 'S Knows that p', *Noûs*, 36: 663–81.

Nozick, Robert (1981) *Philosophical Explanations* (Cambridge).

Partee, Barbara (1989) 'Binding Implicit Variables in Quantified Contexts', in *Proceedings of the Chicago Linguistics Society*, 25 (Chicago), 342–65.

Peirce, C. S. (1877) 'The Fixation of Belief', *Popular Science Monthly*, 12: 1–15.

Pryor, James (2000) 'The Skeptic and the Dogmatist', *Noûs*, 34: 517–49.

Rooth, Mats (1992) 'A Theory of Focus Interpretation', *Natural Language Semantics*, 1: 75–116.

Ryle, Gilbert (1949) *The Concept of Mind* (Chicago).

Sanford, David (1991) 'Proper Knowledge', in Brian McLaughlin (ed.), *Dretske and his Critics* (Oxford), 38–51.
Schaffer, Jonathan (2004*a*) 'From Contextualism to Contrastivism', *Philosophical Studies*, 119: 73–103.
—— (2004*b*) 'Skepticism, Contextualism, and Discrimination', *Philosophy and Phenomenological Research*, 69: 138–55.
—— (n.d.) 'Knowing the Answer', unpublished typescript.
Schiffer, Stephen (1996) 'Contextualist Solutions to Scepticism', *Proceedings of the Aristotelian Society*, 96: 317–33.
Sinnott-Armstrong, Walter (2004) *Pyrrhonian Skepticism* (Oxford).
Sintonen, Matti (1997) 'Explanation: The Fifth Decade', in Matti Sintonen (ed.), *Knowledge and Inquiry: Essays on Jaakko Hintikka's Epistemology and Philosophy of Science* (Amsterdam), 225–38.
Sosa, Ernest (1986) 'On Knowledge and Context', *Journal of Philosophy*, 83: 584–5.
—— (1999) 'How to Defeat Opposition to Moore', *Philosophical Perspectives*, 13: 141–53.
Stalnaker, Robert (1999*a*) 'Assertion', *Context and Content* (Oxford), 78–95.
—— (1999*b*) 'On the Representation of Context', *Context and Content* (Oxford), 96–113.
Stanley, Jason (2000) 'Context and Logical Form', *Linguistics and Philosophy*, 23: 391–434.
—— and Timothy Williamson (2001) 'Knowing How', *Journal of Philosophy*, 98: 411–44.
Stine, G. C. (1976) 'Skepticism, Relevant Alternatives, and Deductive Closure', *Philosophical Studies*, 29: 249–61.
Unger, Peter (1975) *Ignorance: A Case for Scepticism* (Oxford).
Van Fraassen, Bas (1980) *The Scientific Image* (Oxford).
Vogel, Jonathan (1999) 'The New Relevant Alternatives Theory', *Philosophical Perspectives*, 13: 155–80.
—— (n.d.) 'Varieties of Skepticism', unpublished typescript.
Watson, C. S. (1973) 'Psychophysics', in Benjamin B. Wolman (ed.), *Handbook of General Psychology* (Englewood Cliffs, NJ), 275–306.
Williamson, Timothy (2000) *Knowledge and its Limits* (Oxford).
Wittgenstein, Ludwig (1969) *On Certainty*, ed. G. E. M. Anscombe and G. H. von Wright (Oxford).

This page intentionally left blank

10. Paradox and the A Priori

Stephen Schiffer

1. A MCKINSEY PARADOX

A paradox is a set of apparently mutually incompatible propositions each one of which is apt to seem plausible, at least when considered on its own. A well-known paradox in the literature on privileged access and externalism due to Michael McKinsey (1991) may for present purposes be stated, at least initially, as the set consisting of the following four propositions:

(1) John is a priori justified in believing that he believes that if dogs bark, then dogs bark.
(2) John is a priori justified in believing that dogs have existed if he—or anyone else—believes that if dogs bark, then dogs bark.
(3) For any propositions p, q, if one is a priori justified in believing both p and (q if p), then, ceteris paribus, one is a priori justified in believing q.
(4) Even when cetera are paria, one cannot be a priori justified in believing that dogs have existed.

Evidently, (1)–(4) are mutually incompatible, and the prima-facie plausibility of each of these propositions may be glossed in the following way.

Re (1)

The traditional Kantian conception of a priori knowledge is knowledge that is ''independent of experience''. This alludes to the way in which one is *justified* in *believing* the proposition one knows. Whether or not knowledge is a complex state that includes belief,[1] knowing p entails

[1] See Williamson (2000) for an argument that it is not.

being justified in believing *p*, and while one's belief may enjoy being justified in more than one way, there is at least one way of being justified that sustains one's knowledge. What makes one's knowing *p* a priori is that one's knowing *p* is sustained by one's being a priori justified in believing *p*. The 'a priori' in 'a priori knowledge' pertains to the way in which one is justified in believing that which one knows. Now, I am aware that some philosophers are reluctant to allow that one can know a priori that one believes any proposition, and these philosophers will deny that anyone is ever a priori justified in believing that he or she believes any given proposition. But if one sticks to the definition of a priori as ''independent of experience'', then it does seem that John, a normal New Yorker (if there is such a thing), is a priori justified in believing that he believes that if dogs bark, then dogs bark.[2] For the way in which he is justified does indeed *seem*—at least when considered on its own, prior to reflection on the paradox—to be ''independent of experience''. While John has, and knows that he has, the same evidence we have for believing that he believes that if dogs bark, then dogs bark, he does not believe that he believes that if dogs bark, then dogs bark *on the basis of* that evidence. His being justified in believing that he believes that if dogs bark, then dogs bark does not seem to consist in his having any reasons for holding that belief, empirical or otherwise; nor does his being justified seem to consist in his having an experience of any particular kind. The idea of one's being justified in believing a proposition without one's having any particular kind of perception or sensation or any a posteriori or a priori reason may at first strike one as implausible, but it is actually a familiar phenomenon that is arguably manifested in, e.g., the way you are justified in believing that it is wrong to torture children just for the fun of it, or the way you are justified in believing that if there are numbers, then there are numbers. We may even in these cases speak of a person's justification for her belief, provided we take this to mean

[2] As will presently be apparent, one of my purposes in this paper is to engage the a posteriori/a priori distinction as it is currently understood in epistemology. But apart from that concern, much of this paper would be unchanged if the McKinsey paradox were stated not in terms of a priori justification, but in terms of a stipulated sense of non-evidential justification, where evidential justification would include the way in which one's perceptual beliefs are justified by one's sense experiences but would not include the way in which one is justified in believing that, say, one believes that there are prime numbers.

whatever it is that constitutes her being justified, where this need not consist in any reasons she has for believing what she does.

[The notion of justification as whatever constitutes one's being justified in believing a proposition is vague. I would *like* to think that the vague notion suffices for my purposes in this paper, but that may be naïve, and even if it is not, it would be nice to have some idea of what an acceptable precisification of the vague notion might look like. Perhaps a few precisifications would suit my purposes in this paper; I offer the following as a stab at one of them.

The form '*J* is a justification for *x*'s believing *p*' has more than one use, and some of them imply neither that *x* believes *p*, nor, if *x* does believe *p*, that *x* is justified in believing *p*. As I shall use the form, however, *J*'s being a justification for *x*'s believing *p* entails that *x* believes *p*, that *x* is justified in believing *p*, and that *J* is what constitutively makes *x* justified in believing *p*. I say '*a* justification', as opposed to '*the* justification', to allow for overdetermination with respect to what makes *x* justified in believing *p*.[3] Yet, as I already confessed, even with these stipulations, the locution is still pretty vague as regards how justifications are to be individuated or what it is for something to belong to, or be part of, one, and I doubt that these questions enjoy determinately correct answers with any helpful degree of precision. So, I shall further stipulate that—at least to a first approximation[4]—*J* is a justification for *x*'s believing *p* iff *J* is a maximal sequence of sets of true propositions such that:

(i) each set constitutes a non-trivial metaphysically sufficient condition for every succeeding member of the sequence;
(ii) each set constitutes a non-trivial metaphysically sufficient condition for *x*'s being justified in believing *p*; and
(iii) each proposition in each member set is essential to the set's constituting a metaphysically sufficient condition for *x*'s being justified in believing *p* (that is, each such proposition is a necessary part of a sufficient but not necessarily necessary condition for *x*'s being justified in believing *p*).

[3] I do not feel a need to add that *J* is an ''epistemic'' justification for *x*'s believing *p*, because I do not think there can be moral or prudential justifications *for believing p*; such justifications can only be justifications for *making it the case that one believes p*, say, by taking a drug that one believes will cause one to believe that one is a very lovable person.

[4] I would not be shocked to learn that what follows falls short of the mark—that is, of what I need for this paper—in one way or another.

To this I add the following gloss:

- To say that a set of propositions s constitutes a metaphysically sufficient condition for a set of propositions s' is to say that the conjunction of the members of s metaphysically entails the conjunction of the members of s'.
- A set of propositions constitutes a *trivial* metaphysically sufficient condition for itself, but I shall not attempt any further precisification of what it is to constitute a ''non-trivial'' metaphysically sufficient condition for a set of propositions. (A consequence of the non-triviality requirements is that the fact that x is justified in believing p is not part of what justifies x in believing p.)
- Nor will I attempt to accommodate either degrees of belief or degrees of justification, although the notion of partial belief will make a cameo appearance in what follows. The mere notion of being justified in believing a proposition should suffice for my purposes.
- I shall say that a proposition is included in—or is a part or component of—a justification J which x has for believing p if that proposition either (a) is a member of one of J's member sets or (b) is directly entailed by some conjunction of propositions that are members of one or more of those sets. Entailment here includes metaphysical necessitation, and I will only gesture by example at the intended sense of 'directly' in 'directly entails'. Suppose that the fact that I came to believe q in such-and-such a way entails that I am a posteriori justified in believing q. Then that fact also entails that [(I am a posteriori justified in believing q) and $(68 + 57 = 125)$] and that [(I am a posteriori justified in believing q) or $(68 + 57 = 5)$]. Here, only the proposition that I am a posteriori justified in believing q is ''directly'' entailed. The subtle (a)-or-(b) disjunctive condition is intended to enable me to say that x's being a priori/a posteriori justified in believing a certain proposition q is ''part of'' one's justification for believing p, when the proposition that one is a priori/a posteriori justified in believing q is not itself a member of any of J's member sets.
- The reason I do not count each minimally sufficient condition as a distinct justification for x's believing p—and the reason my definition takes its complex form—is that one sufficient condition for x's being justified may supervene on another, and in that case we

should not want to say that *x*'s being justified in believing *p* is overdetermined. Such supervenience can happen in one of at least two ways. First, a sufficient condition whose specification requires epistemic notions such as *justification* may supervene on a condition that is specifiable in non-epistemic terms. Second, a sufficient condition that is specifiable in non-epistemic terms may supervene on a condition that is specifiable in more basic non-epistemic terms, in the way that, say, biological facts supervene on physical-chemical facts.

- Roughly speaking, a justification *J* for *x*'s believing *p* is *maximal* just in case no other justification can be derived from *J* by ''inserting'' a new set of true propositions anywhere in *J*.]

Re (2)

We may suppose John is an intelligent undergraduate who has had a few philosophy courses. John is familiar with Putnam's twin earth thought experiments; he has read about unicorns in *Naming and Necessity*; and he is familiar with the notion of object-dependent concepts and propositions from his reading of Gareth Evans and more recent things in the theory of propositional content. As a result, John has come to accept the philosophical theory according to which it is necessarily the case that: (i) the property of being a dog is the property of belonging to a natural kind to which something belongs if, but only if, it is a dog; (ii) this natural kind is individuated partly in terms of a certain evolutionary lineage that would not exist if dogs had never existed; and (iii) propositions involving our concept *dog*—that is, propositions to which we may refer using that-clauses containing the word 'dog'—are individuated partly in terms of the property of being a dog, so that those propositions would not exist if that property did not exist. And from all this, John recognized, it follows that dog-propositions are dog-dependent, which is to say that propositions involving our concept *dog* would not exist if dogs had never existed. The reasoning that led to John's philosophical belief—a belief in a proposition that is either necessarily true or necessarily false—was of the kind with which we philosophers are very familiar, and it is apt to seem (at first glance, anyway[5]) that we would say that it is a priori.

[5] In case you are already protesting that John's reasoning will arguably presuppose that dogs have existed, and that that is not a proposition John can be a priori justified in

Re (3)

This closure principle presupposes the more familiar closure principle that if one is justified in believing both p and (q if p), then, ceteris paribus, one is justified in believing q. The more familiar closure principle is really better stated as the principle that if one is justified in believing the conjunction [p & (q if p)], then, ceteris paribus, one is justified in believing q.[6] The reason this is a better statement is not that a person who believes two propositions might not put them together so that she believes the conjunction; that contingency is already reasonably taken to be subsumed by the ceteris paribus clause. The reason is that—by my lights, at any rate—believing a proposition is just a matter of believing it to a contextually relevant high degree,[7] and it may be that, while both the degree to which a rational person believes p and the degree to which she believes (q if p) pass the relevant threshold for deeming the person to believe both propositions *tout court*, the degree to which she believes the conjunction [p & (q if p)] is below that threshold. I propose simply to bypass this complication by assuming throughout that, if in the cases at issue one is justified in believing both p and (q if p), then one is also justified in believing the conjunction [p & (q if p)], and, when it simplifies the exposition, I shall write as though the closure principle is formulated in terms of such a conjunction.[8]

believing, let me counsel patience and remind you that I am at this point not speaking *in propria persona*, but merely glossing the prima-facie plausibility of (2) and the other members of the set of mutually incompatible propositions.

[6] What *explains* the truth of this closure principle, assuming it is true? That is a very important question, but one I shall not here try to answer. I will, however, venture the thought that it is constitutive of being a rational believer who possesses the concepts of conjunction and the conditional that it is impossible for such a person, when functioning normally, not to believe q when she believes [p & (q if p)]. This sort of claim is a key ingredient both in commonsense functionalist accounts of propositional attitudes, such as those in (Lewis 1983*b*) and (Loar 1981), and in more recent work on concepts and justification, such as (Peacocke 1992) and (Boghossian 2003).

[7] I am aware that there are those who deny this—e.g., Harman (1986: 22–4), Williamson (2000: 99), and Peacocke (2003: 113–15)—but I am unpersuaded by their reasons.

[8] But why not replace the (1)–(4) formulation of the McKinsey paradox with the one consisting of the following three mutually incompatible propositions?

- John is a priori justified in believing that [(he believes that if dogs bark, then dogs bark) & (dogs have existed if he believes that if dogs bark, then dogs bark)].
- For any propositions p, q, if one is a priori justified in believing [p & (q if p)], then, ceteris paribus, one is a priori justified in believing q.
- Even when cetera are paria, one cannot be a priori justified in believing that dogs have existed.

Anyway, when the closure principle (3) is understood in the now stipulated way, it is apt to seem plausible, and I shall often suppress the ceteris paribus qualification, since it pertains to general considerations of rationality which we may take to be in place. Very few philosophers would deny that if one is justified in believing [p & (q if p)], then, ceteris paribus, one is justified in believing q, and it is apt to seem hard to see how we can get a false principle by putting 'a priori' before 'justified'. After all, the fact that you are justified in believing q cannot require anything that is not entailed by the fact that you are justified in believing [p & (q if p)]. Given that, how can your justification for believing q depend on experience when your justification for believing [p & (q if p)] is independent of experience?

Re (4)

It may be hard to find anyone willing to deny this.[9] Paul Boghossian (1998: 275) expressed a very widely held view when he wrote that the proposition that water exists is ''clearly not knowable a priori'', and this conviction, I suspect, extends to the thought that one might be a priori justified in believing that dogs have existed. If we could be a priori justified in believing that there are, or have been, dogs, then it may seem that we could in principle be a priori justified in believing any empirical fact, and that is apt to seem preposterous.

So much for the prima-facie plausibility of the mutually incompatible propositions comprising the paradox set. What has to give?

2. A SHARPENING OF THE ISSUES

The traditional Kantian conception of a priori knowledge is knowledge that is ''independent of experience'', and thus knowledge sustained by one's being justified ''independently of experience'' in believing that which one knows. I think it is fair to construe the intended sense of 'independent of experience' as meaning that, when one is a priori justified in believing p, then one's being justified in the way in which one is

I could; but, as will presently be clear, it is important that the paradox be stated in a way that makes it easy separately to address John's justification for believing that he believes that if dogs bark, then dogs bark and his justification for believing dogs have existed if he believes that if dogs bark, then dogs bark.

[9] An exception is Sawyer (1998).

justified—one's justification or entitlement or warrant (in this paper I use these notions interchangeably) for believing *p*—does not itself entail one's being a posteriori justified in believing any other proposition. That may sound circular, but it is not; it merely means that however we precisify the vague notion of being independent of experience, an experience-independent way of being justified in believing a proposition cannot itself entail one's being justified in believing some other proposition in an experience-dependent way. Thus, according to the traditional Kantian conception of the a priori, you are not a priori justified in believing a proposition if you are justified in believing that proposition in a way that entails being a posteriori justified in believing some other proposition.

There is something puzzling about this conception of the a priori. The a priori/a posteriori distinction is supposed to be mutually exclusive and jointly exhaustive of ways of knowing and ways of being justified, and no one balks at the idea that one may be a posteriori justified in believing a proposition even though the way in which one is justified requires being a priori justified in believing a certain other proposition. For example, my a posteriori justification for believing that Jane and Anthony are second cousins may essentially involve my being a priori justified in believing that two people are second cousins if a parent of one is a first cousin of a parent of the other. Considerations of symmetry therefore suggest (but by no means entail) that our conception of the a priori might allow one to be a priori justified in believing a proposition even though the way in which one is justified entails being a posteriori justified in believing a certain other proposition. There are in fact examples which give some support to this symmetric conception of the a priori. Here are three such examples.

(i) Most philosophers who work on the problem of vagueness believe that a proposition is knowable only if it is determinately true. If, for example, it is indeterminate whether Harold is bald, then these philosophers would agree that one cannot know either that Harold is bald or that he is not bald. I take it that, whether or not this widely held belief counts as knowledge (or is even true), it is a belief that is justified, and that the justification philosophers have for believing it is one they regard as a priori. It is the result of the sort of armchair conceptual analysis and theorizing that characterize analytical philosophy. At the same time, *part* of the justification these philosophers have for believing that proposition about vagueness is the empirical proposition that they have neither heard nor themselves been able to think of any counter-

examples to it. Thus, if I am right that this justification counts as a priori as philosophers use that expression of art, then one's justification for believing a proposition may count as a priori even if that justification includes one's being a posteriori justified in believing a certain empirical proposition. If in such a case what one is a priori justified in believing is also something one knows, then there are instances of one's knowing a proposition a priori where one's justification for accepting the proposition is not wholly independent of the character of one's experience.

(ii) A student taking her first logic course is given a homework assignment in which she is asked to determine whether a certain formula is a theorem of propositional logic. She proves that it is a theorem, but not having complete confidence in her newly acquired skills, does not fully believe that it is. When she then speaks to her friend Bob, whom she knows to be good at logic, and he tells her that he came up with the same proof, she then fully believes that the formula is a theorem. I believe many philosophers would be content to class the student's justification as a priori, even though part of that justification is her being a posteriori justified in believing a certain empirical proposition, that is, that her friend came up with the same proof.

There is a better way to gloss this example. We need again to advert to the fact that beliefs come in degrees:[10] one can believe a proposition more or less firmly, to a greater or lesser degree, and to believe a proposition *tout court* is just to believe it to a contextually relevant high degree. Let us pretend that degrees of belief can be measured in the interval [0, 1], 0 representing complete disbelief, 1 complete belief.[11] Then we may suppose that before she spoke with her friend, the student believed to, say, degree .8 that the formula was a theorem. At that time, her justification for believing that proposition to degree .8 was entirely a priori; it was just her a priori confidence in the proof she constructed. But when she learns that her friend Bob came up with the same proof, she comes to believe to, let us suppose, degree .95 that the formula is a theorem. At that point, her justification for believing *to degree .95* that the formula is a theorem includes her a posteriori justified belief that

[10] See p. 278 above.

[11] The issue of partial belief is very complex, and there is more than one way in which I am presently indulging in simplification. See e.g. Schiffer (2003: ch. 5). But I think the simplifications are benign in the context of this paper, since the points I am using them to make would be unaffected by a replacement of the simplifications with the considerably more complex real McCoy.

Bob came up with the same proof of the formula, yet, notwithstanding this, there is, I believe, some inclination to say that her justification for believing that the formula is a theorem is a priori.[12] Whether or not this is the best way to speak—a question I shall presently address—we should notice that this is not a case of overdetermination. The student would not believe to degree .95 that the formula was a theorem just on the basis of believing that her friend thought he proved that it was. Her single justification for believing to degree .95—and thus, in context, for believing *tout court*—that the formula is a theorem is constituted by the purely a priori reasons she had for believing to degree .8 that her proof was sound together with the extent to which she was able to take the fact that Bob came up with the same proof to be empirical evidence of the proof's soundness.

An actual example of this sort is provided by Andrew Wiles's justification for believing that his proof of Fermat's Last Theorem is sound. Wiles worked out his proof in private and did not show it to other mathematicians until he felt he had it right. No doubt his degree of belief in the soundness of his proof went up when his work was confirmed by other mathematicians, so that when he finally fully believed he had proved the theorem, his justification for believing it had this ineliminable a posteriori element.

(iii) Accepting as you do the law of excluded middle, you believe that Giuseppe Verdi did or did not write the song ''I'm Too Sexy for My Shirt''. But you know that that instance of excluded middle entails that Verdi existed, and you are not justified in believing that Verdi did or did not write the song unless you are justified in believing that he existed. But even though your justification for believing that Verdi existed is a posteriori, I think that many philosophers would say you are a priori justified in believing the logical truth that Verdi did or did not write ''I'm Too Sexy for My Shirt''.

[12] Christopher Peacocke (2004: 28) correctly observes that, ''there is a distinction between what gives us access to the entitling conditions for a priori knowledge, and the entitling conditions themselves''. But I believe he makes an overstatement when he adds that ''possession of what the thinker knows to be a proof (a tree-structure of contents) provides an a priori entitlement to accept a logical or arithmetical proposition'' (2004: 28). Whether someone in possession of what he knows to be a proof has a priori knowledge that the proof is a proof depends on whether he is a priori justified in believing that the proof is sound. A person may have only an a posteriori justification for believing that a proof is sound, and the student example suggests that a person may have an a priori justification with an a posteriori component for believing that a proof is sound.

Well, do these examples show that one may be a priori justified in believing a proposition even though one's justification includes the fact that one is a posteriori justified in believing a certain other proposition? I doubt that the current use of 'a priori justified belief' among philosophers is such that it could deliver a determinate verdict. Fortunately, a determinate verdict is not required; it is possible to resolve the McKinsey paradox without getting embroiled in verbal disputes generated by a tendentious use of the labels 'a priori' and 'a posteriori'. Notwithstanding this, however, there is a distinction implicit in the examples already before us to which we should attend before attempting a resolution.

Certain indisputably a priori justifications are wholly a priori, in that one's a priori justification for believing a certain proposition does not contain an a posteriori justification for believing any proposition, and certain indisputably a posteriori justifications are wholly a posteriori, in that one's a posteriori justification for believing a certain proposition does not contain an a priori justification for believing any proposition. Then there are mixed justifications, justifications that have both a priori and a posteriori components, and here we find an interesting division. In certain of these cases we would say that the justification as a whole is indisputably and determinately a posteriori, notwithstanding its a priori component. My justified belief that Jane and Anthony are second cousins is such an example: we would say that I am a posteriori justified in believing that they are second cousins, but an essential part of my justification is my being a priori justified in believing that two people are second cousins if each has a parent that is a first cousin of a parent of the other. In certain other cases, such as the examples (i)–(iii), philosophers' intuitions appear to be mixed. None of these examples, I believe, is such that nearly all philosophers would judge them to be examples of indisputably, or determinately, a posteriori justifications. But philosophers evidently fall into four groups as regards any one of the examples (i)–(iii). Some will say the justification is a posteriori; some will say that it is a priori; some will say that it is neither a priori nor a posteriori, but rather a justification that is partly a priori and partly a posteriori; and some will not know what to say. As I said in the preceding paragraph, it may be that none of these responses is determinately wrong; and, as I also said, such indeterminacy does not preclude a determinate resolution of the McKinsey paradox. Still, we have an interesting distinction that demands an explanation.

The interesting distinction is between the mixed justifications that are indisputably a posteriori, such as my a posteriori justification for believing that Jane and Anthony are second cousins, and those, such as (i)–(iii), which are arguably neither determinately a posteriori nor determinately a priori. I believe there is a principled basis for this distinction.[13] To a first approximation, a mixed justification for *x*'s believing *p* is determinately a posteriori when, and only when, two conditions are satisfied: first, the a priori component by itself does not justify *x* in believing *p* to any degree; and second, what the a priori component does do is enable *x* to take the a posteriori component to be *evidence* that *p* is true (that is, all other things being equal, the a posteriori component does not justify *x* in believing *p* to any degree in the absence of the a priori component, but given the a priori component, *x* is justified in taking the a posteriori component to be evidence that *p* is true). In the second cousin example, my evidence for believing that Jane and Anthony are second cousins is whatever evidence I have for thinking that each has a parent who is a first cousin of a parent of the other, but I would not be able to take that evidence as evidence that Jane and Anthony are second cousins if I did not know that two people are second cousins if a parent of one is a first cousin of a parent of the other. At the same time, my being a priori justified in believing *that two people are second cousins if each has a parent who is a first cousin of a parent of the other* by itself gives me no reason to believe to any degree that Jane and Anthony are second cousins. In the mixed cases that are arguably neither determinately a priori nor determinately a posteriori, the a priori component does not serve *merely* to enable the a posteriori component to be taken to be evidence that *p* is true. In some of these cases, the a priori component in itself justifies *x* in believing *p* to *some* degree, the a posteriori component being that in the justification which justifies *x* in believing *p* to the higher degree to which *x* in fact believes *p*. This is what is going on in examples (i) and (ii). In other cases, such as the Verdi example (iii), the a posteriori component (e.g. your evidence that Verdi existed) is needed just to secure that there is a proposition which is an instance of a propositional schema (in this case the schema *p or not-p*, which represents the law of excluded middle) that one is a priori entitled

[13] This explanation was inspired by remarks made by Celia Teixeira during the discussion of an earlier draft of this paper at the 2003 European Summer School in Analytical Philosophy.

to believe has only true instances. Presently we will see that there is still another way in which a mixed justification may fall on the not-clearly-a-posteriori side.

I formulated the McKinsey paradox as the set consisting of the following four mutually incompatible propositions:

(1) John is a priori justified in believing that he believes that if dogs bark, then dogs bark.
(2) John is a priori justified in believing that dogs have existed if he believes that if dogs bark, then dogs bark.
(3) For any propositions p, q, if one is a priori justified in believing both p and (q if p), then, ceteris paribus, one is a priori justified in believing q.
(4) Even when cetera are paria, one cannot be a priori justified in believing that dogs have existed.

But in view of the somewhat equivocal nature of a priori justification that has come to light, let me now offer these two *stipulative* definitions:

> A person is *purely a priori justified* in believing p, and thus has a *pure a priori justification* for believing p, just in case her justification for believing p—or one of her justifications for believing p, should it be overdetermined that she is justified in believing p—does not include her being a posteriori justified in believing some other proposition.
>
> A person is *impurely a priori justified* in believing p, and thus has an *impure a priori justification* for believing p, just in case (a) her justification for believing p—or one of her justifications for believing p—is not determinately a posteriori but (b) does include her being a posteriori justified in believing some other proposition.

Then our initial paradox set, (1)–(4), may give way to these two precisifications:[14]

[14] Other precisifications are possible involving mixed cases—e.g. ones which would require the (obviously false) closure principle that if one is purely a priori justified in believing one of the propositions p and (q if p) and impurely a priori justified in believing the other, then, ceteris paribus, one is impurely a priori justified in believing q—but these add nothing of relevance to what is already covered by the two displayed.

(1p) John is purely a priori justified in believing that he believes that if dogs bark, then dogs bark.

(2p) John is purely a priori justified in believing that dogs have existed if he believes that if dogs bark, then dogs bark.

(3p) For any propositions p, q, if one is purely a priori justified in believing both p and (q if p), then, ceteris paribus, one is purely a priori justified in believing q.

(4p) Even when cetera are paria, one cannot be purely a priori justified in believing that dogs have existed.

(1i) John is impurely a priori justified in believing that he believes that if dogs bark, then dogs bark.

(2i) John is impurely a priori justified in believing that dogs have existed if he believes that if dogs bark, then dogs bark.

(3i) For any propositions p, q, if one is impurely a priori justified in believing both p and (q if p), then, ceteris paribus, one is impurely a priori justified in believing q.

(4i) Even when cetera are paria, one cannot be impurely a priori justified in believing that dogs have existed.

At this point, the following possible resolution, which I shall call *Resolution A*, is apt to suggest itself. (As will presently be clear, I am *not* putting Resolution A forward as the resolution I accept.)

As regards (1p)–(4p), Resolution A holds, (3p) and (4p) are true, but (1p) and (2p) are false.

(4p) is true, because the only justification a belief that dogs have existed can enjoy is an a posteriori justification.

(3p) is true for the following reason. We are supposing that if one believes both p and (q if p) and cetera are paria, then one believes the conjunction [p & (q if p)], and it is conceptually impossible for a normal person to be justified in believing [p & (q if p)] but not to be justified in believing q (this is in part because it is impossible for a normal person to believe [p & (q if p)] and not to believe q). Now suppose that one is justified in believing q only if one believes q on the basis of justification J. Since being justified in believing [p & (q if p)] entails being justified in believing q, it follows that one's justification for believing [p & (q if p)] must entail J. If J is either an a posteriori or impure a priori justification, then it follows that one's justification for believing [p & (q if p)] cannot be a pure a priori justification, since in either event it means that part of what makes one justified in believing

[*p* & (*q* if *p*)] is that one is a posteriori justified in believing a certain proposition.

Neither (1p) nor (2p) is true for the following reason. Since John accepts the philosophical theory according to which dog-propositions are dog-dependent, and since he knows that both the proposition *that he believes that if dogs bark, then dogs bark* and the proposition *that dogs have existed if he believes that if dogs bark, then dogs bark* are dog-propositions, he will not be justified in believing either proposition unless he is justified in believing that dogs have existed. But the only justification anyone can have for believing that dogs have existed is an a posteriori justification.

As regards (1i)–(4i), Resolution A holds, (1i), (2i), and (4i) are true, but (3i) is false.

(1i) and (2i) are true because, while John's being a posteriori justified in believing that dogs have existed is part of what makes him justified in believing both *that he believes that if dogs bark, then dogs bark* and *that dogs have existed if he believes that if dogs bark, then dogs bark*, the a priori components of those latter two justifications do not function to enable John to take the fact that dogs have existed as evidence for the truth of either belief, and this, as noted above, precludes those justifications from being determinately a posteriori.

(4i) is true for the same reason that (4p) is true: the only justification a belief that dogs have existed can enjoy is an a posteriori justification.

And finally, the closure principle (3i) is obviously false, since it is obvious that it is compatible with one's being impurely a priori justified in believing [*p* & (*q* if *p*)] that one is either a posteriori justified or purely a priori justified in believing *q*. For example, by the above displayed stipulative definition of impure a priori justification, one is impurely a priori justified in believing that [Verdi did or did not write ''I'm Too Sexy for My Shirt'' & (Verdi existed if he did or did not write that song)], but one is only a posteriori justified in believing that Verdi existed.

So much for Resolution A. I agree with some of its claims, but disagree with others.

I agree with Resolution A that (3p) is true (relative to the stipulated understanding of it) and that (3i) is false.

I disagree with Resolution A that (4p) and (4i) are true; I believe that both are false. Resolution A claims that it is *impossible* for anyone to be either purely or impurely a priori justified in believing that dogs have

existed, and this because it supposes that the only justification a belief that dogs have existed *can* enjoy is an a posteriori justification. But, as I will presently try to show, it is *possible*—albeit unlikely—for someone to be purely a priori justified in believing that dogs have existed, and since that is possible, it is also possible for someone's justification for that belief to be an impure a priori justification by virtue of containing an a posteriori justification for some proposition other than that dogs have existed (perhaps the a posteriori element pertains to the proposition that no one has been able to think of a counter-example to a certain theoretical component of the total justification).

Subject to an important qualification, I agree with Resolution A that (2p) is false and that (2i) is true. The important qualification is this. Given Resolution A's claim that it is *impossible* for anyone to be purely a priori justified in believing that dogs have existed, it is committed to saying that it is impossible for (2p) to be true and that, therefore, (2i) cannot be false by virtue of John's being purely a priori justified in believing *that dogs have existed if he believes that if dogs bark, then dogs bark*. I, however, want to say that, *as John has been described* (i.e. as a normal, well-educated person living in New York as we know it), (2p) is false and (2i) is true, but that he *might have been* such that (2p) was true and (2i) was false, and this, as I said in the preceding paragraph, because John might have been such that he was purely a priori justified in believing that dogs had existed.

I do not agree with Resolution A that (1p) is false and (1i) is true. My description of John leaves open whether he is purely or impurely a priori justified in believing that he believes that if dogs bark, then dogs bark. Given that, as things are, John's justification for believing that dogs have existed can only be a posteriori, that a posteriori justification must be part of either his justification for believing [α] *that he believes that if dogs bark, then dogs bark* or his justification for believing [β] *that dogs have existed if he believes that if dogs bark, then dogs bark*. It *must* be part of his justification for believing [β], but it *need not* be part of his justification for believing [α].

It must be part of John's justification for believing [β], because it is impossible for him to believe [β] and not to believe that dogs have existed, and he would not believe that dogs had existed unless he was a posteriori justified in believing by empirical evidence that dogs had existed. For John has to believe [β] given his ability to recognize

dog-propositions and his accptance of the philosophical theory that are dog-dependent. But in exactly the same way, he has to believe, at least on reflection, any conditional like [β] but with a different dog-proposition as the antecedent. Thus, John has to believe

> that dogs have existed if (he believes that dogs have existed if he believes that if dogs bark, then dogs bark),

and thus, by the closure of belief under known entailment, has to believe that dogs have existed.

But no a posteriori justification for believing that dogs have existed need be part of John's justification for believing [α]. He need not believe [α] even partly on the basis of believing that dog-propositions are dog-dependent, and it is consistent with my description of John that his belief in [α] is much more secure than his belief in the philosophical theory that entails [β], so that he believes that should it transpire that there never were any dogs, then that would merely show that his philosophical view [β] was false. He may justifiably have such a conviction even if dog-propositions are dog-dependent. Here is an analogy. I believe that it is a necessary condition for experiencing pain that I be in a certain functional state, but should it transpire that I cannot be in that functional state, I would conclude that experiencing pain does not entail being in that functional state, not that I had never experienced pain. To be sure, John's conviction that dog-propositions are dog-dependent may be so great that should he come to think that dogs have never existed, then he would also conclude that no proposition was ever referred to by his that-clause 'that if dogs bark, then dogs bark', and in that event his justification for believing [α], as well as his justification for believing [β], would contain his a posteriori justification for believing that dogs have existed. That is why my description of John leaves the truth-values of (1p) and (1i) open.

Resolution A does not resolve the McKinsey paradox, but if what I have claimed in listing my agreements and disagreements with Resolution A is correct, then we will have a resolution when we see how one might be purely a priori justified in believing that dogs have existed, and, correlatively, why, as things are, John can have only an a posteriori justification for believing it. In the end, we shall see that *the resolution of the McKinsey paradox turns essentially on a point that has nothing to do either with externalism or with a priori justification.*

3. RESOLVING THE MCKINSEY PARADOX

Ceteris paribus, one is justified in believing q if one is justified in believing [p & (q if p)]. Suppose that one is justified in believing [p & (q if p)] and that cetera are paria. Then one is justified in believing q, and there are at least two possible scenarios as regards one's being justified in believing q.

Inherited Justification

In this scenario,[15] one has a justification for believing q that is inherited from a justification one has for believing [p & (q if p)], and therefore that justification for believing [p & (q if p)] does not include a justification one has for believing q which is independent of one's being justified in believing [p & (q if p)]—that is, a justification one has for believing q which one would have even if one were not justified in believing [p & (q if p)]. If one's being justified in believing q is not overdetermined, then one is justified in believing q wholly on the basis of being justified in believing [p & (q if p)]; one believes q on the basis of believing [p & (q if p)], and one's justification for believing q just is (so to speak) one's justification for believing [p & (q if p)].

Here is a mundane example of Inherited Justification. I believe that [(Smith's child is ill) & (Smith will not attend today's colloquium if his child is ill)], and my justification for believing that conjunction contains no independent justification that I have for believing that Smith will not attend today's colloquium. Here I believe that Smith will not attend the colloquium on the basis of believing that [(Smith's child is ill) & (Smith will not attend today's colloquium if his child is ill)], and my justification for believing that Smith will not attend the colloquium derives entirely from my justification for believing that conjunction. In this example of Inherited Justification one comes to believe for the first time that Smith will not attend the colloquium on the basis of one's coming to believe that [(Smith's child is ill) & (Smith will not attend today's colloquium if his child is ill)]. There are also examples of Inherited Justification in which one is already justified in believing q prior to

[15] I trust it is clear that what follows is not a *definition* of some intuitive notion of inherited justification but is merely a label for the stipulated scenario. Likewise for the scenario labeled '*Uninherited Justification*'.

becoming justified in believing [p & (q if p)]. This would be true of Lester in the following example. On Monday he cannot see how to prove a certain mathematical proposition q but is a posteriori justified in believing q on the basis of being told that it is true by a mathematician who would know. There is another mathematical proposition p such that on Tuesday Lester sees for the first time how to prove both p and (q if p), where neither proof relies on q, thus becoming purely a priori justified in believing [p & (q if p)], and thus via Inherited Justification also becoming purely a priori justified in believing q, in addition to his already being a posteriori justified in believing it.

Uninherited Justification

In this scenario, a justification one has for believing q is not inherited from a particular justification one has for believing [p & (q if p)]. Suppose that neither one's being justified in believing [p & (q if p)] nor one's being justified in believing q is overdetermined. Then one's justification for believing q is one that one has independently of being justified in believing [p & (q if p)]. In this case, there is no sense in which one believes q on the basis of believing [p & (q if p)]; rather, one believes either p or (q if p)—and thus believes their conjunction—partly on the basis of believing q. To take a trivial example, since I am justified in believing that [(I am wearing a grey shirt) & (Jones was at the meeting if I am wearing a grey shirt)], I am justified (ceteris paribus) in believing that Jones was at the meeting. But I believe that Jones was at the meeting because I saw her there, and that is also my justification for believing *that Jones was at the meeting if I am wearing a grey shirt*: since I know that Jones was at the meeting, I know that she was there whether or not I am wearing a grey shirt.

There are interesting differences among cases of Uninherited Justification, and one kind of case is of special interest with respect to present concerns. In cases of this kind, one would not believe [p & (q if p)] on the basis of a justification that did not include an independent justification of a certain kind K for believing q—that is, a justification of kind K for believing q that was not itself a justification for believing [p & (q if p)]—because one knows that if q were true, one would have, and know that one has, an independent justification of kind K for believing q. For example, I know that if my pants were on fire, I would have, and know that I have, sensory evidence of an expected sort that that was

so. Consequently, for any proposition *p*, I would not believe [*p* & (my pants are on fire if *p*)] unless my justification for believing that conjunction included the expected sort of sensory justification for my believing that my pants are on fire. Thus, as things actually stand with me and my environment, I can have no Inherited Justification for believing that my pants are on fire. This is because there can be no proposition *p* such that I can believe [*p* & (my pants are on fire if *p*)] on the basis of a justification that does not include the obvious sort of sensory justification for believing that my pants are on fire: since I know that I would be justified in that way in believing that my pants were on fire if they were on fire, I would not believe anything that entailed that my pants were on fire unless I was already justified in the relevant way in believing that my pants were on fire. (It should be clear that the point of the pants example in no way depends on the fact that I do not now believe that my pants are on fire. For the same sort of reasons that are operative in that example, I cannot now have an Inherited Justification for believing that I am alive.)

As opposed to examples of the preceding sort, nothing now prevents me from acquiring an Inherited Justification for believing that I have 15.6 GB of used space on my computer's C drive. As it happens, I neither believe nor disbelieve that proposition, and I expect not to be justified in believing that I do, or do not, have 15.6 GB of used space on my computer's C drive unless I do certain things to find out. So, when I click on the right place and read that I have 15.6 GB of used space on my computer's C drive, nothing interferes with my believing that I have 15.6 GB of used space on my computer's C drive. As things actually stand, I cannot have an Inherited Justification for believing that my pants are on fire or that I am alive, but I can easily have an Inherited Justification for believing that I have 15.6 GB of used space on my computer's C drive.

We are now in a position to see why John is in no position to become purely a priori justified in believing that dogs have existed.[16] Even if he had never entertained the thought that dog-propositions are dog-dependent, John, a normal, well-educated New Yorker, would implicitly know that it would be extremely unlikely for dogs to have existed without his having

[16] As already noted (p. 291 above), the fact that John is now a posteriori justified in believing that dogs have existed does not per se preclude him from also becoming purely a priori justified in believing that proposition.

empirical evidence which justified him in believing that dogs had existed (that is to say—more or less—that John's subjective conditional probability that he has good empirical evidence that dogs have existed, given that dogs have existed, is high). Consequently, given that cetera are paria, there can be no proposition *p* such that John would believe [*p* & (dogs have existed if *p*)] unless his justification for believing that conjunction included an empirical-evidence-based a posteriori justification for his believing that dogs had existed, and from this it follows that John cannot have a purely a priori inherited justification for believing that dogs have existed which derives from his believing both *that he believes that if dogs bark, then dogs bark* and *that dogs have existed if he believes that if dogs bark, then dogs bark*. Since it is pretty clear that no one can be in a position to have a purely a priori *un*inherited justification for believing that dogs have existed, we have explained why John is in no position to become purely a priori justified in believing that dogs have existed.[17] Since we have seen that, and how, John might well be purely a priori justified in believing that he believes that if dogs bark, then dogs bark, the foregoing also explains why, as things actually are with John and the world, he can at best be impurely a priori justified in believing the philosophical theory that dog-propositions are dog-dependent, and thus in believing *that dogs have existed if he believes that if dogs bark, then dogs bark*.

We are also in a position to see how it is metaphysically possible for someone to be purely a priori justified in believing that dogs have existed. To see this, it will be helpful if we first switch to a different example, to a sketch of a near-fetched scenario in which a purely a priori justified belief in a contingent proposition combines with a purely a priori justified belief in a necessary proposition in a way that leads to one's being purely a priori justified in believing a contingent proposition of a sort that one might not at first have supposed could be justifiably believed on a priori grounds.

[17] If my claim about what John implicitly knows is correct, then he is also not in any position to have an impurely a priori justified belief that dogs have existed. But what if he merely implicitly believes that [it is fairly unlikely that dogs should have existed without his having *some* empirical evidence that that was so]? Might he then have an impurely a priori justified belief that dogs had existed? That question will implicitly get an affirmative answer when I show below how it is *metaphysically possible* for John to be purely a priori justified in believing that dogs have existed.

On Sunday, John, our neophyte philosopher from the previous example, is purely a priori justified in believing that he has certain beliefs. For example, he believes that he believes that there are prime numbers greater than 7, and that belief is not based on any sort of empirical evidence or made justified by any experience or sensation.[18] We may even assume that his belief is empirically indefeasible: even if John were to learn that, like a chocolate Easter bunny, his head was hollow, that would just prove to him that you do not need a stuffed head to have beliefs. Since John is purely a priori justified in believing that he has a particular belief, there is no problem in allowing that on Sunday he is also purely a priori justified in believing that he has beliefs. Also on Sunday, John is aware of the philosophical issue of whether having beliefs necessitates having information-processing states with sentential structure, but he has read nothing about it and given it no thought.[19] He has no opinion on that question one way or the other. He neither believes nor disbelieves that having beliefs necessitates having information-processing states with sentential structure, and this is not because he is acquainted with considerations both for and against that thesis which cancel each other out. It is because he is not aware of any relevant considerations. Now, it would be possible for John not to have an opinion one way or the other about the claim that having beliefs *necessitates* having information-processing states with sentential structure while also having some reason to think that people do in fact have some information-processing states with sentential structure. As it happens, however, John has no reason of any kind either to believe or to disbelieve that he or anyone else has information-processing states with sentential structure; he is not aware of any relevant considerations one way or the other; he has no opinion on the matter, and he can see no reason why the fact of the matter, whatever it is, would manifest itself in empirical evidence to which he would have access.

On Monday things begin to change when John delves into the literature on cognitive architecture and is impressed with various a priori

[18] A person will believe that she believes that such-and-such only if she believes that she exists, but I take it that one is purely a priori justified in believing that one exists by virtue of one's being purely a priori justified in believing that one believes that such-and-such. This is the moral of Descartes's *Cogito*.

[19] This example was suggested to me by Martin Davies's discussion of cognitive architecture (Davies 1998, 2003). As will soon be apparent to those familiar with Davies's papers, Davies and I disagree about what the example is an example of.

arguments by, among others, Brian Loar (1981), Martin Davies (1992), and myself (Schiffer 1993)[20] —arguments whose conclusions entail that having beliefs necessitates having information-processing states with sentential structure. John spends the rest of that day and all day Tuesday and Wednesday rereading these works and thinking hard about their arguments, which he finds more and more persuasive. By Thursday, John has reconstructed an a priori argument he believes is sound and whose conclusion is that having beliefs necessitates having information-processing states with sentential structure. It is doubtful that John can be said to *know* that his argument is sound, even if it is, but our issue is about justified belief, and it does seem to me plausible that John's philosophizing should have been well enough conducted so that he is purely a priori justified in believing that his argument is sound and, therefore, that having beliefs necessitates having information-processing states with sentential structure (the idea is that belief states are either identical to such states or else realized by them). John is also, of course, purely a priori justified in believing *that if (he has beliefs and having beliefs necessitates having information-processing states with sentential structure*), *then he has information-processing states with sentential structure*. In this way, John comes to believe that he has information-processing states with sentential structure, even though he has no empirical evidence or other a posteriori justification for this belief. But John's newly acquired belief was derived from two purely a priori justified beliefs, and so is itself purely a priori justified. To be sure, John is well aware that future scientific research might prove him wrong, might discover that our internal information-processing states do not have sentential structure, and if that should transpire, then, while he would not stop believing that he has beliefs, he would cease to believe the philosophical theory he is currently purely a priori justified in believing. Nevertheless, John's justification for believing that he has information-processing states with sentential structure is an instance of Inherited Justification. Moreover, since his justification is inherited from pure a priori justifications, it is itself a pure a priori justification.

Let me briefly recap John's progress, and register a slight qualification. John began in an initial state in which he was purely a priori justified in believing that he had beliefs and in which he had no reason to believe or to disbelieve that he or anyone else had

[20] These three works are cited in Davies (2003).

information-processing states with sentential structure. Thus, when John encountered the philosophical argument and was persuaded by it on wholly a priori grounds, nothing interfered with his being purely a priori justified in believing both that he has beliefs and that he has information-processing states with sentential structure if he has beliefs. And since pure a priori closure (suitably understood) is correct, it follows that John was then also purely a priori justified in believing that he had information-processing states with sentential structure. The slight qualification to which I just alluded is that a pure a priori justification was stipulated to be a justification with *no* a posteriori component, but all that really matters for the issues at hand is that John's justification for accepting the philosophical argument about cognitive architecture does not in any way rely on his being a posteriori justified in believing that anyone actually has information-processing states with sentential structure. It is irrelevant for present purposes if, for example, John's justification, like the philosopher's justification for believing that a proposition is knowable only if it is determinately true, includes the fact that he has not been able to think of any counter-examples to a certain claim. But rather than complicate the discussion with such qualifications, it is harmless and expositionally convenient to suppose that John's justification for accepting the philosophical argument about cognitive architecture is purely a priori, as well it might be, so long as that justification does not include an a posteriori justification for believing that he, or anyone else, has information-processing states with sentential structure.

That completes my characterization of John as regards his views about cognitive architecture and my case for concluding that on Thursday he is purely a priori justified in believing that he has information-processing states with sentential structure. In view of this, it is clear what we should say about the mutually incompatible propositions (1*)–(4*):

(1*) John is purely a priori justified in believing that he has beliefs.

(2*) John is purely a priori justified in believing that he has information-processing states with sentential structure if he has beliefs.

(3*) For any propositions p, q, if one is purely a priori justified in believing both p and (q if p), then, ceteris paribus, one is purely a priori justified in believing q.

(4*) Even when cetera are paria, one cannot be purely a priori justified in believing that anyone has information-processing states with sentential structure.

What we should say about (1*)–(4*) is that (1*)–(3*) are true, but that (4*) is false, since cetera are paria and John *is* purely a priori justified in believing that he has information-processing states with sentential structure.

What I propose we should say about (1*)–(4*) stands in marked contrast with what I proposed we should say about the mirroring set of propositions (1p)–(4p):

(1p) John is purely a priori justified in believing that he believes that if dogs bark, then dogs bark.

(2p) John is purely a priori justified in believing that dogs have existed if he believes that if dogs bark, then dogs bark.

(3p) For any propositions p, q, if one is purely a priori justified in believing both p and (q if p), then, ceteris paribus, one is purely a priori justified in believing q.

(4p) Even when cetera are paria, one cannot be purely a priori justified in believing that dogs have existed.

Now (3p) = (3*), and we may deem John to be such that (1p) is true. But (2p) is false: given that John is a normal New Yorker, he will be justified in believing *that dogs have existed if he believes that if dogs bark, then dogs bark* only if he has empirical evidence which a posteriori justifies his believing that dogs have existed. And while I suggested—a suggestion I have yet to explain or justify—that (4p) is false owing to its being *metaphysically possible* for a person to be purely a priori justified in believing that dogs have existed, I also implied that if we replace (4p) with

(4′p) Even though cetera are paria, John is not purely a priori justified in believing that dogs have existed

we shall have replaced it with a true proposition, while maintaining a set of mutually incompatible propositions. At the same time, it follows from what I said that its counterpart,

(4**) Even though cetera are paria, John is not purely a priori justified in believing that he has information-processing states with sentential structure

is false, since cetera are paria and John actually is purely a priori justified in believing that he has information-processing states with sentential structure.

What explains these differences between Structure and Dog (as I shall call these two scenarios)? We know the answer. John is in a position to have a justification for believing *that he has information-processing states with sentential structure* which is inherited from a justification—even a pure a priori justification—he has for believing that [(he has beliefs) & (he has information-processing states with sentential structure if he has beliefs)], but he is not in a position to have *any* justification for believing that dogs have existed which is inherited from a justification he has for believing that [(he believes that if dogs bark, then dogs bark) & (dogs have existed if he believes that if dogs bark, then dogs bark)]. And the reason this is so turns much less on the a priori/a posteriori distinction than it does on the distinction between Inherited and Uninherited Justification. More exactly, what it crucially turns on is the fact that, prior to coming to believe the philosophical theory that having beliefs necessitates having information-processing states with sentential structure, John did not believe, implicitly or otherwise, that he would have any kind of reason, let alone one based on empirical evidence, to believe that he had information-processing states with sentential structure, if that were so, whereas prior to coming to believe the philosophical theory that dog-propositions are dog-dependent, John did implicitly believe that he would have good empirical reasons to believe that dogs had existed, if that were so. The Dog side of that difference puts severe constraints on what inherited justifications John can have for believing that dogs have existed in a way that explains why he is precluded not only from being purely a priori justified in believing the philosophical theory that dog-propositions are dog-dependent, a theory that if true is necessarily true, but from having any justification for believing that dogs have existed which is *inherited* from any justification he has for believing the conjunction. But the Structure side of that difference itself leaves entirely unconstrained what kinds of inherited justifications John can have for believing that he has information-processing states with sentential structure in a way that explains why he is not precluded from being purely a priori justified in believing the philosophical theory that having beliefs entails having information-processing states with sentential structure, also a theory that if true is

necessarily true. As I said, the difference between Dog and Structure turns for the most part on the distinction between Inherited and Uninherited Justification.

It remains to explain why (4p) is false, why, that is, it is *metaphysically possible* for someone to be purely a priori justified in believing that dogs have existed. The reason it is metaphysically possible is simply that it is metaphysically possible for someone to be in the same position *vis-à-vis* the proposition that dogs have existed that John was in *vis-à-vis* the proposition that he has states with sentential structure prior to his accepting the philosophical argument about cognitive architecture. That is to say, it is possible for someone to have the concept *dog* but have no reason either to believe or to disbelieve that dogs had ever existed. A possible world in which a normally situated normal person who had the concept *dog* but had no reason to believe or to disbelieve that dogs had existed is probably very distant indeed from the actual world, but it need not be. Suppose that in the near future there is a nuclear holocaust in which only a very few people and other mammals survive, none of which are dogs, and all physical evidence of dogs which could be recognized as such is destroyed. The world that remains is in a virtual state of nature. A version of English survives, however, and one of the surviving adults speaks on occasion to his surviving daughter about dogs, describing them and drawing pictures of them (he is a good artist) in considerable detail. Unfortunately, the adult's mind was disturbed by the holocaust, and as his daughter matures, it becomes clear to her that her father's renderings of what the world was like before the holocaust are not to be trusted. At a certain point in her development, she has the concept *dog* but neither believes nor disbelieves that there ever were any dogs, and she can see no reason why the fact of the matter, whatever it is, should reveal itself to her in empirical evidence to which she has access. Now her father was a philosophy professor in a major New Jersey department, and he is never more lucid as when he is explaining to his daughter that theory of natural kind concepts according to which (*a*) *dog* is a natural kind concept if it is metaphysically possible that there were dogs and (*b*) dog-propositions are dog-dependent if *dog* is a natural kind concept. At the same time, the daughter has become convinced on the basis of a priori conceivability considerations that the existence of dogs is metaphysically possible, and we may suppose that her belief in that proposition is purely a priori justified, even if the

argument sustaining it is not altogether sound.[21] In this way, the daughter acquires via Inherited Justification the defeasible but purely a priori justified belief that dogs have existed, and we see that it is metaphysically possible for someone to be purely a priori justified in believing that dogs have existed.

That concludes my resolution of the McKinsey paradox with which I began this paper, but it will be instructive to look at the problem again from a slightly different angle, one that relates the results reached about Structure and Dog to what Crispin Wright (1985) has called ''transmission of warrant'', and to the classical Cartesian argument for skepticism about perceptual knowledge.

4. MCKINSEY PARADOX, SKEPTICISM, INHERITED JUSTIFICATION, AND TRANSMISSION OF WARRANT

Consider the following version of the classical Cartesian skeptical argument, where it is understood that I know that if there is a blue cube before me, then I am not a BIV (a bodiless brain in a cubeless vat whose every sensory experience is caused by electrochemical stimulations administered by a computer):

(1) I am not justified in believing that there is a blue cube before me unless I have a justification for believing that I am not a BIV which is independent of my current sensory experience.
(2) I have no such justification.
(3) ∴ I am not justified in believing that there is a blue cube before me.

Philosophers are divided in their response to this sort of argument. One big division would occur over premise (1). Some philosophers, such as James Pryor (2000) and Christopher Peacocke (2003), would reject this

[21] Is there a *sound* purely a priori argument whose conclusion is that dog-propositions are dog-dependent? I doubt it, and this may be a further difference between Dog and Structure, since it does seem to me more plausible that there is a sound purely a priori argument to show that having beliefs necessitates having information-processing states with sentential structure. But this paper is about the McKinsey paradox when stated in terms of belief *justification*, not in terms of *knowledge* (the terms in which that paradox is usually stated), and it is of course possible for someone to become purely a priori justified in believing a proposition, even though the a priori argument that sustains her belief has a false premise.

premise. These philosophers would hold that one needs an independent justification for disbelieving that one is a BIV only if one has reason to suspect that one might be a BIV; in ordinary circumstances, wherein one has no such reason, one's experience as of a blue cube's being before one directly justifies one in believing that there is a blue cube before one, and thereby also justifies one in disbelieving that one is a BIV, provided one is also justified in believing that one is not a BIV if there is a blue cube before one. Other philosophers, such as Crispin Wright (2003) and Martin Davies (2003),[22] would accept (1). Their view, roughly speaking, is that in order for one's experience as of *p*'s being the case to justify one in believing *p* one must be entitled independently of that experience to disbelieve any hypothesis which one knows to be both incompatible with *p* and such that one would have precisely the same sort of experience if that hypothesis were true. (Disagreement about premise (1) goes along with disagreement about premise (2), and here things can get sticky.[23] My focus now is just on premise (1).)

The dividing issue just rehearsed may be rejoined as an issue about Inherited Justification. Can one have a justification for believing that one is not a BIV that is inherited from one's justification for believing both that there is a blue cube before one and that one is not a BIV if there is a blue cube before one, or is it that such justification cannot be inherited, because, given one's awareness of the entailment, in order to be justified in believing that there is a blue cube before one, one must be independently justified in believing that one is not a BIV? Both sides of the dispute should accept the (appropriately qualified) closure principle that, ceteris paribus (and we may assume throughout that cetera are paria), one is justified in believing *q* if one is justified in believing both *p* and (*q* if *p*). The issue is about whether one can justifiably believe both *that there is a blue cube before one* and *that one is not a BIV if there is a blue cube before one* without being independently justified in believing that one is not a BIV. Those in the Peacocke–Pryor camp say that is not required, that one can inherit one's justification for believing that one is not a BIV from one's justification for believing the two other propositions. Those in the Davies–Wright camp deny this; they hold that a justification for believing that one is not a BIV cannot be inherited from a justification one has for believing the other two propositions, and

[22] Davies (forthcoming), however, indicates a change of mind.
[23] See Schiffer (2004).

this because, given awareness of the entailment, one cannot be justified in believing that there is a blue cube before one unless one is independently justified in believing that one is not a BIV.

The dividing issue may also be rejoined as an issue about transmission of warrant. John, the subject of our previous examples, was stipulated to be a normal, rational, intellectually mature member of our society, and we may ask about his epistemic position with respect to this Moorean inference:

> *Cube*
> There is a blue cube before me.
> If there is a blue cube before me, then I am not a BIV.
> ∴ I am not a BIV.

It seems to me that both those who accept premise (1) of the Cartesian skeptical argument and those who reject it ought to agree that two things are true of John *vis-à-vis* Cube. First, John cannot actively believe Cube's premises without at the same time actively believing its conclusion, and second, John is justified in believing Cube's conclusion if he is justified in believing its premises. Given this, the dividing issue re-emerges in the following way.

Recall that on my stipulated use of 'warrant' and 'justification', those two terms are used interchangeably, and one's justification for believing a proposition, when one is justified in believing it, is whatever makes one justified in believing it. Further, in order to avoid other presently irrelevant complexities, I shall assume that if John is justified in believing that he is not a BIV, then his being so justified is not overdetermined, that is, that he has just one justification for that belief; and I shall also assume that John has no reason to suspect that he may be a BIV. Relative to these stipulations, the division between those who reject premise (1) of the Cartesian argument and those who accept it comes to this.

Those who reject premise (1), such as Peacocke and Pryor, will claim that John is justified in believing Cube's premises, and thus in believing its conclusion; that his justification for believing its conclusion just is his justification for believing the conjunction of Cube's two premises, and therefore his justification for believing the first premise does not include a justification for believing the conclusion. For these theorists, John's justification for believing the first premise is, roughly, the fact that he seems to see a blue cube before him while having no reason to

suspect that any defeating hypothesis may be true, and his justification for believing the second premise is the obvious a priori justification, which neither side of the dividing issue challenges. As I understand talk of ''warrant transmission'',[24] these theorists would hold that in this case John's warrant, or justification, for the premises of Cube transmits to its conclusion.

Those who accept premise (1) of the Cartesian argument, such as Davies and Wright, will claim that John, who has the whole inference in mind, is not justified in believing Cube's first premise unless he has an independent justification for believing its conclusion. Consequently, if he is justified in believing the conjunction of its premises, then his justification for believing that conjunction cannot be his justification for believing Cube's conclusion. His justification for believing the conjunction of the two premises will include his justification for believing the conclusion, assuming he has such a justification, but it will include other things, such as his seeming to see a blue cube, which are extraneous to his justification for believing that he is not a BIV. If he is justified in believing that he is not a BIV, that justification is one that he would have even if he was not justified in believing either of Cube's premises. As I understand talk of ''transmission of warrant'', these theorists would hold that, if John has a warrant for Cube's premises, it does not transmit to its conclusion.

It should be clear that, at least for the cases at hand, the issue about transmission of warrant is identical to the issue about Inherited Justification: in these cases, warrant is transmitted just in case justification is inherited. In fact, I find talk of warrant transmission to be at best misleading in a couple of ways, and I can make no good sense of the notion other than in terms of the distinction between Inherited and Uninherited Justification. Putting that point aside, however, I shall now continue the discussion mostly in terms of transmission of warrant, since I want presently to connect my views on these matters with those of Crispin Wright.

My own position (Schiffer 2004) on the Cartesian argument, and thus on Cube, differs from that of both camps but is much closer in spirit to

[24] There is reason to doubt whether the conditions for transmission of warrant proposed by theorists such as Wright and Davies actually capture the phenomenon they meant to capture; see Silins (forthcoming). I intend my claims about warrant transmission (relative to my stipulations about 'warrant') to cohere with the intended application of that expression of art.

the Wright–Davies camp, which holds that premise (1) of the Cartesian argument is true, and that therefore warrant fails to be transmitted in Cube, since one's sense experience as of a blue cube's being before one will not justify one in believing that there is a blue cube before one unless one is justified in disbelieving that one is a BIV in a way that is independent of that experience (and, of course, all other similar sensory experience). But my purpose now is not to address that debate. Rather, let us assume that the correct response to Cartesian skepticism entails that, for the reasons I gave, warrant fails to be transmitted in Moorean inferences such as Cube and ask how this might bear on the McKinsey paradox which is this paper's primary concern.

So let us return to the Dog setup and imagine John contemplating this inference:

> *Dog**
> I believe that if dogs bark, then dogs bark.
> Dogs have existed if I believe that if dogs bark, then dogs bark.
> ∴ Dogs have existed.

It follows from my proposed resolution of the McKinsey paradox, relative to my stipulations about John, that Dog*, like Cube, suffers from transmission failure, since, as I argued, John's justification for believing the conclusion of Dog* is not inherited from his justification for believing its premises. John's mixed warrant for the second premise —that is, his justification for believing the premise—will not transmit to the conclusion, because it itself includes John's independent a posteriori warrant for the conclusion. The question I want to consider now is an instance of one raised by Crispin Wright in a couple of recent articles: does transmission of warrant fail in Dog* for the same sort of reason it fails in Cube (assuming it does fail in Cube)?

Wright (2000) defends a view which suggests that he would answer yes (although, as I shall presently note, he qualifies that view in an even more recent paper). Reconstructed in terms best suited both to my formulations of Cube and Dog* and to my own construal of transmission failure, his proposal suggests the following unitary account of transmission failure in those two inferences.

(*a*) Assume that John is justified in believing the premises of both Cube and Dog*. Then what suffices for, and explains, the transmission failure in both Cube and Dog* is that in both cases there is a proposition *C* such that (i) part of John's warrant for the first premise consists in

his being in a state that is subjectively indistinguishable from the state he would be in if *C* were true and (ii) *C* would be true if the conclusion were false.

(*b*) In Cube, *C* = the proposition that the thinker is a BIV who is now having a visual experience as of a blue cube's being directly in front of him, and it is clear how the explanation goes.

(*c*) In Dog*, we have the following. (i) The first premise—the proposition that John believes that if dogs bark, then dogs bark—entails (on the assumption that the second premise is a necessary truth) the conclusion, the proposition that dogs have existed. (ii) *C* = the proposition ''that the seeming-thought which [John] attempt[s] to express by ['I believe that if dogs bark, then dogs bark'] is content-defective owing to the reference failure of the purported natural kind term''[25] 'dog' in John's language, and thus (iii) part of John's warrant for the first premise consists in his being in a state which is subjectively indistinguishable from the state he would be in if *C* were true.

I do not think this attempt to give a unitary explanation of the two kinds of transmission failure succeeds. I have some reservations about whether Wright's proposal adequately explains the transmission failure in Cube, but I grant that it offers a reasonable first shot.[26] Roughly speaking, it tells us that the warrant provided by my visual experience as of seeing a blue cube cannot transmit to a warrant for believing that I am not a BIV because I would have just the same visual experience if I were a BIV. There are several reasons why this account doesn't seem to explain the transmission failure in Dog*.

First, and most important, while John's state of believing Dog*'s first premise is subjectively indistinguishable from the state he would be in if *C* were true, it is not the case that John's warrant for that premise consists even in part in his being in a state that is subjectively indistinguishable from the state he would be in if *C* were true. That is not an unreasonable thing to say about John's warrant for the first premise in

[25] Wright (2000: 156).

[26] Let *C** be the proposition that I seem to see a blue cube before me but there is not one there. It seems to me false that I am not justified in believing that there is a blue cube before me unless I have an *independent* justification for disbelieving *C**: that would introduce a circularity from which it would be impossible to escape. Hypotheses suitable for making a Cartesian skeptical argument must be hypotheses whose truth would *explain* one's relevant sense experience. Yet it follows from Wright's proposal that *C** is an acceptable value of his '*C*' and that, consequently, it should be suitable for incorporation into a Cartesian argument.

Cube, because part of his warrant for believing that there is a blue cube before him is that he is having a visual experience of a certain sort. But there is no sensation or experience of any kind that is part of John's warrant for believing that he believes that if dogs bark, then dogs bark. There is no Nagelian ''what it is like'' to be in that belief state; in the sense of ''subjective indistinguishability'' in play, John's state of believing *that he believes that if dogs bark, then dogs bark* is just as subjectively indistinguishable from his state of believing that there are prime numbers greater than 7 as it is from the state he would be in if *C* were true. The account of what justifies John in believing *that he believes that if dogs bark, then dogs bark* makes no reference to anything that would make ''subjective indistinguishability'' relevant. (It might be protested that Wright's point is merely that for all John noninferentially knows for certain, he is suffering from an ''illusion of content'': he thinks there is a proposition to which 'the proposition that he believes that if dogs bark, then dogs bark' refers and which he believes, but if there never were any dogs or dog-like creatures, there would be no proposition to which the singular term refers, and thus he would be mistaken about what he believes. This response, however, is implicitly addressed in the next objection.)

Second, if the Wright-suggested account of transmission failure in Cube and Dog* were correct, we should have transmission failure when John reasons

> *Structure**
> I have beliefs.
> If I have beliefs, then I have information-processing states with sentential structure.
> ∴ I have information-processing states with sentential structure.

For here *C* might be the proposition that none of John's information-processing states have sentential structure, which would induce a kind of illusion of content, given the truth of the second premise: in uttering the first premise John would not be expressing the belief he in fact is and takes himself to be expressing, since John would have no beliefs to express. But if what I said before about (1*)–(4*) is correct, there is not transmission failure in this case, for in this case John's purely a priori justification for believing the conclusion is inherited from his purely a priori justification for believing both premises. As a result of being purely a priori justified in believing the premises of Structure*, John

becomes purely a priori justified (albeit defeasibly) in believing its conclusion. John remains purely a priori justified in believing that he has beliefs without an independent warrant for believing that he has information-processing states with sentential structure, notwithstanding that John would be suffering a kind of illusion of content in producing the first premise should his second premise be true and he lacks states with sentential structure.

Third, as Wright makes clear in his most recent publication on this subject (2003), he would concede that it cannot be supposed that 'dog' in John's idiolect would fail to refer if dogs had never existed. Perhaps in the relevant subjectively indistinguishable state of affairs John's 'dog' refers to things that look and behave exactly like dogs but belong to a species other than *Canis familiaris*.[27]

Fourth, implicit in Wright's account of transmission failure in Dog* is the claim that the assumption that dog-propositions are dog-dependent is needed to explain why warrant fails to be transmitted in that inference. But even if dog-propositions are not dog-dependent, John would still have his warrant for believing the premises of Dog*, and, as that warrant necessitates John's being a posteriori justified in believing that dogs have existed, it would still fail to transmit to the proposition that dogs have existed.

Fifth, since John might cease to believe only his second premise if he should come to doubt the conclusion of Dog*, it cannot be that the failure of John's warrant for the premises of Dog* to transmit to its conclusion is to be explained in terms of the failure of his warrant for the first premise to transmit to the conclusion. If what I said in the preceding section is correct, transmission failure in this case is explained by John's having a mixed justification for believing *that dogs have existed if he believes that if dogs bark, then dogs bark* by virtue of the dependence of that justification on his having an independent a posteriori justification for believing that dogs have existed.

I do not think a unitary account can be given of the transmission failures in Cube and Dog*. In Cube, transmission of warrant fails because if one were a BIV, *one would have just the same experiential warrant one actually has* for believing that there is a blue cube before one—namely, one's seeming to see a blue cube before one—and *that is*

[27] Wright's revised view, I believe, does not escape the other objections I am raising to the view it revises.

why one will not be justified in believing that there is a blue cube before one unless one is independently warranted in disbelieving that one is a BIV. In Dog*, however, when transmission fails, as it does in John's case (I argued that in certain possible worlds it does not fail), it is because of the connection John takes to obtain between dogs having existed and his having empirical evidence of a certain kind that dogs have existed. Owing to what John takes that connection to be, he will not, as explained in the preceding section, believe any proposition that entails that dogs have existed unless his having empirical evidence of a certain kind renders him a posteriori justified in believing that dogs have existed. I said John might be such that only his justification for his philosophical theory of the dog-dependence of the concept *dog* was empirically defeasible by evidence that dogs never existed, but, as already indicated, nothing much changes if we suppose he would conclude that there never were any dog-propositions if he came to believe there never were any dogs. That would just show that he had mixed justifications for both premises, each justification dependent on his having an a posteriori justification for the proposition that dogs have existed. Besides, if John were to be suffering an illusion of content, he would not have the justification he has for believing that he believes that if dogs bark, then dogs bark. For that justification, I should think, essentially includes the fact that he does believe that if dogs bark, then dogs bark.[28]

REFERENCES

Boghossian, Paul (1998) 'What the Externalist can Know a Priori', in C. Wright, B. C. Smith, and C. Macdonald (eds.), *Knowing our own Minds* (Oxford), 271–84.

——(2003) 'Blind Reasoning', *Proceedings of the Aristotelian Society, supplementary volume*, 77: 225–48.

Davies, Martin (1992) 'Aunty's own Argument for the Language of Thought', in J. Ezquerro and J. M. Larrazabal (eds.), *Cognition Semantics and Philosophy: Proceedings of the First International Colloquium on Cognitive Science* (Dordrecht), 235–71.

[28] Thanks to Yuval Avnur, Paul Boghossian, Emma Borg, Jonathan Dancy, Alice Drewery, Tamar Szabó Gendler, Allan Gibbard, Hanjo Glock, John Hawthorne, Paul Horwich, Nikola Kompa, Anna-Sara Malmgren, Susana Nuccetelli, Christopher Peacocke, Jim Pryor, Sven Rosenkranz, Josh Schechter, Celia Teixeira, and Crispin Wright.

—— (1998) 'Externalism, Architecturalism, and Epistemic Warrant', in Crispin Wright, Barry C. Smith, and Cynthia Macdonald (eds.), *Knowing our own Minds* (Oxford), 321–61.
—— (2003) 'The Problem of Armchair Knowledge', in S. Nuccetelli (ed.), *New Essays on Semantic Externalism and Self-Knowledge* (Cambridge), 23–57.
—— (forthcoming) 'Epistemic Entitlement, Warrant Transmission, and Easy Knowledge', *Proceedings of the Aristotelian Society, supplementary volume*, 78.
Ezquerro, Jesús, and Jesús M. Larrazabal (eds.) (1992) *Cognition, Semantics, and Philosophy: Proceedings of the First International Colloquium on Cognitive Science* (Dordrecht).
Harman, Gilbert (1986) *Change in View: Principles of Reasoning* (Cambridge).
Jackson, Frank, and Michael Smith (eds.) (2004) *Oxford Handbook to Contemporary Philosophy* (Oxford).
Lewis, David (1983*a*) *Philosophical Papers*, i (Oxford).
—— (1983*b*) 'Radical Interpretation', *Philosophical Papers*, i (Oxford), 108–32.
Loar, Brian (1981) *Mind and Meaning* (Cambridge).
McKinsey, Michael (1991) 'Anti-Individualism and Privileged Access', *Analysis*, 51: 9–16.
Nuccetelli, Susanna (ed.) (2003) *New Essays on Semantic Externalism and Self-Knowledge* (Cambridge).
Peacocke, Christopher (1992) *A Study of Concepts* (Cambridge).
—— (2003) *The Realm of Reason* (Oxford).
—— (2004) 'The A Priori', in F. Jackson and M. Smith (eds.), *Oxford Handbook to Contemporary Philosophy* (Oxford).
Pryor, James (2000) 'The Skeptic and the Dogmatist', *Noûs*, 34: 517–49.
Sawyer, Sarah (1998) 'Privileged Access to the World', *Australasian Journal of Philosophy*, 76: 523–33.
Schiffer, Stephen (1993) 'Actual-Language Relations', *Philosophical Perspectives*, 7: 231–58.
—— (2003) *The Things we Mean* (Oxford).
—— (2004) 'Skepticism and the Vagaries of Justified Belief', *Philosophical Studies*, 119: 161–84.
Silins, Nicholas (forthcoming) 'Transmission Failure Failure', *Philosophical Studies*.
Williamson, Timothy (2000) *Knowledge and its Limits* (Oxford).
Wright, Crispin (1985) 'Facts and Certainty', *Proceedings of the British Academy*, 71: 429–72.

Wright, Crispin (2000) 'Cogency and Question-Begging: Some Reflections on McKinsey's Paradox and Putnam's Proof', *Philosophical Issues*, 10: 140–63.
—— (2003) 'Some Reflections on the Acquisition of Warrant by Inference', in Susanna Nuccetelli (ed.), *New Essays on Semantic Externalism and Self-Knowledge* (Cambridge), 57–79.
—— Barry C. Smith, and Cynthia Macdonald (eds.) (1998) *Knowing our own Minds* (Oxford).

11. Scepticism, Rationalism, and Externalism

Brian Weatherson

This paper is about three of the most prominent debates in modern epistemology. The conclusion is that three prima-facie appealing positions in these debates cannot be held simultaneously.

The first debate is **scepticism vs. anti-scepticism**. My conclusions apply to *most* kinds of debates between sceptics and their opponents, but I will focus on the inductive sceptic, who claims we cannot come to know what will happen in the future by induction. This is a fairly weak kind of scepticism, and I suspect many philosophers who are generally anti-sceptical are attracted by this kind of scepticism. Still, even this kind of scepticism is quite unintuitive. I am pretty sure I know (1) on the basis of induction.

(1) It will snow in Ithaca next winter.

Although I am taking a very strong version of anti-scepticism to be intuitively true here, the points I make will generalize to most other versions of scepticism. (Focusing on the inductive sceptic avoids some potential complications that I will note as they arise.)

The second debate is a version of **rationalism vs. empiricism**. The kind of rationalist I have in mind accepts that some deeply contingent propositions can be known a priori, and the empiricist I have in mind denies this. Kripke showed that there are *contingent* propositions that can be known a priori. One example is *Water is the watery stuff of our acquaintance*. ('Watery' is David Chalmers's nice term for the properties of water by which folk identify it.) All the examples Kripke gave are

Previous versions of this paper were presented at Cornell University, the Inland Northwest Philosophy Conference, and the Syracuse Workshop on the A Priori. Each time I received valuable feedback. Thanks also to David Chalmers, Earl Conee, Harold Hodes, Nicholas Sturgeon, a reader for *Oxford Studies in Epistemology*, and, especially, Tamar Szabó Gendler for very helpful comments on various drafts.

of propositions that are, to use Gareth Evans's term, deeply necessary (Evans 1979). It is a matter of controversy presently just how to analyse Evans's concepts of deep necessity and contingency, but most of the controversies are over details that are not important right here. I will simply adopt Stephen Yablo's recent suggestion: a proposition is deeply contingent if it could have *turned out* to be true, and could have *turned out* to be false (Yablo 2002).[1] Kripke did not provide examples of any *deeply* contingent propositions knowable a priori, though nothing he showed rules out their existence.

The final debate is a version of **internalism vs. externalism** about epistemic justification. The internalist I have in mind endorses a very weak kind of access internalism. Say that a class of properties (intuitively, a determinable) is *introspective* iff any beliefs an agent forms by introspection about which property in the class (which determinate) she instantiates are guaranteed to not be too badly mistaken.[2] (Since 'too badly' is vague, 'introspective' will be vague too, but as we'll see this will not matter to the main argument.) My internalist believes the following two claims:

- Which propositions an agent can justifiably believe supervenes in which introspective properties she instantiates, and this is knowable a priori.[3]

[1] If you prefer the 'two-dimensional' way of talking, a deeply contingent proposition is one that is true in some possible world 'considered as actual'. See Chalmers (2004) for a thorough discussion of ways to interpret this phrase, and the broader notion of so-called 'deep' contingency. Nothing that goes on here will turn on any of the fine distinctions made in that debate—the relevant propositions will be deeply contingent in every plausible sense.

[2] That a property is introspective does not mean that whenever a subject instantiates it she is in a position to form a not too badly mistaken belief about it. Even if the subject instantiates the property she may not possess sufficient concepts in order to have beliefs about it. And even if she has the concept she may simply have more pressing cognitive needs than forming certain kinds of belief. Many agents have no beliefs about the smell in their ordinary environment much of the time, for example, and this does not show that phenomenal smell properties are not introspective. All that is required is that if she forms any beliefs by introspection about which determinate she instantiates, the beliefs are immune from massive error. The need for a restriction to beliefs formed by introspection here was pointed out to me by Earl Conee, who noted that agents who are disposed to believe whatever they are told by a particular testifier can be massively mistaken about anything whatsoever. Although I won't always be explicit about the restriction, when I talk in the text about beliefs an agent has about her introspective properties, I'll be talking solely about such beliefs that are formed by introspection.

[3] There is a delicate ambiguity in this expression to which a referee drew my attention. The intended meaning is that for any two agents who instantiate the same introspective

- There exist some introspective properties and some deeply contingent propositions about the future such that it's a priori that whoever instantiates those properties can justifiably believe those propositions.

My externalist denies one or the other of these claims. Typically, she holds that no matter what introspective properties you have, unless some external condition is satisfied (such as the reliability of the connection between instantiating those properties and the world being the way you believe it is) you lack justification. Alternatively, she holds that the connection between introspective properties and justification is always a posteriori. (Or, of course, she might deny both.)

My argument will be that the combination of anti-scepticism, empiricism and internalism is untenable. Since there is quite a bit to be said for each of these claims individually, that their combination is untenable means we are stuck with a fairly hard choice: accept scepticism, or rationalism, or externalism. Of the three, it *may* seem that externalism is the best, but given how weak the version of internalism that I'm using is, I think we should take the rationalist option seriously.[4] In this paper I'll just argue against the combination of anti-scepticism, empiricism, and internalism, and leave it to the reader to judge which of the three to reject.

Very roughly, the argument for the trilemma will be as follows. There are some propositions q such that these three claims are true.

(2) If anti-scepticism is true, then I either know q a priori or a posteriori.

(3) If internalism and empiricism are true, I do not know q a priori.[5]

(4) If internalism is true, I do not know q a posteriori.

Much of the paper will be spent giving us the resources to find and state such a q, but to a first approximation, think of q as being a proposition

properties, belief in the same propositions is justified. What is not intended is that, if there is an agent who justifiably believes p, and the introspective properties they instantiate are $F_1, \ldots, F_n$, then any agent who instantiates $F_1, \ldots, F_n$ is justified in believing p. For there might be some other introspective property F_{n+1} they instantiate that justifies belief in q, and q might be a defeater for p. The 'unintended' claim would be a very strong, and very implausible claim about the subvenient basis for justification.

[4] Rationalism is supported by BonJour (1997) and Hawthorne (2002), and my argument owes a lot to each of their discussions.

[5] Aesthetically it would be preferable to have the antecedent of this claim be just that empiricism is true, but unfortunately this does not seem to be possible.

like *I am not a brain-in-a-vat whose experiences are as if I was a normal person*.[6] The important features of *q* are that (*a*) it is entailed by propositions we take ourselves to know, (*b*) it is possibly false, and (*c*) if something is evidence for it, then any evidence is evidence for it. I will claim that by looking at propositions like this, propositions that say in effect that I am not being misled in a certain way, it is possible to find a value for *q* such that (2), (3), and (4) are all true. From that it follows that either scepticism or externalism or rationalism is true.

For most of the paper I will assume that internalism and anti-scepticism are true, and use those hypotheses to derive rationalism. The paper will conclude with a detailed look at the role internalism plays in the argument, and this will give us some sense of what an anti-sceptical, empiricist externalism may look like.

1. A SCEPTICAL ARGUMENT

As mentioned, among the many things I know about the future, one of the firmest is (1).

(1) It will snow in Ithaca next winter.

I know this on the basis of inductive evidence about the length of meteorological cycles and the recent history of Ithaca in winter. The inductive sceptic now raises the spectre of Winter Wonderland, a kind of world that usually has the same meteorological cycles as ours, and has the same history, but in which it is sunny every day in Ithaca next winter.[7] She says that to know (1) we must know that (5) is false, and we do not.

(5) I am living in Winter Wonderland.

Just how does reflection on (5) affect my confidence that I know (1)? The sceptic might just appeal to the intuition that I do not know that (5) is false. But I don't think I have that intuition, and if I do, it is much

[6] i.e. I am not a brain-in-a-vat* in the sense of Cohen (1999).

[7] If she is convinced that there is no possible world with the *same* history as ours and no snow in Ithaca next winter, the sceptic will change her story so Winter Wonderland's past differs imperceptibly from the past in our world. She doesn't think this issue is particularly relevant to the *epistemological* debate, no matter how interesting the scientific and metaphysical issues may be, and I agree with her.

weaker than my intuition that I know (1), and that I can infer (5) from (1). James Pryor (2000: 527–8) has suggested that the sceptic is better off using (5) in the following interesting argument.[8]

Sceptical Argument 1

(6) Either you don't know you are not living in Winter Wonderland; or, if you do know that, it is because that knowledge rests, in part, on your inductive knowledge that it will snow in Ithaca next winter.

(7) If you are to know (1) on the basis of certain experiences or grounds *e*, then for every *q* which is ''bad'' relative to *e* and (1), you have to be in a position to know *q* to be false in a non-question-begging way—i.e. you have to be in a position to know *q* to be false antecedently to knowing that it will snow next winter on the basis of *e*.

(8) (5) is ''bad'' relative to any course of experience *e* and (1).

C. You can't know (1), that it will snow next winter, on the basis of your current experiences.

An alternative hypothesis *q* is ''bad'' in the sense used here iff (to quote Pryor) ''it has the special features that characterise the sceptic's scenarios—whatever those features turn out to be'' (2000: 527). To a first approximation, *q* is bad relative to *p* and *e* iff you're meant to be able to know *p* on the basis of *e*, but *q* is apparently compatible with *e*, even though it is not compatible with *p*.

Pryor argues that the best response to the external world sceptic is **dogmatism**. On this theory you can know *p* on the basis of *e* even though you have no prior reason to rule out alternatives to *p* compatible with *e*. Pryor only defends the dogmatic response to the external world sceptic, but it's worth considering the dogmatist response to inductive scepticism. According to this response, I *can* come to know I'm not in Winter Wonderland on the basis of my experiences to date, even though I didn't know this a priori. So, dogmatism is a version of empiricism, and it endorses (6).[9] The false premise in this argument, according to the

[8] Pryor is discussing the external world sceptic, not the inductive sceptic, so the premises here are a little different to those he provides.

[9] To be sure, it is consistent with the letter of dogmatism that we could have known some *other* kinds of deeply contingent propositions a priori, so it is not constitutive of

dogmatist, is (7). We can know it will snow even though the Winter Wonderland hypothesis is bad relative to this conclusion and our actual evidence, and we have no prior way to exclude it.

Pryor notes that the sceptic could offer a similar argument concerning justification, and the dogmatist offers a similar response.

Sceptical Argument 2

(9) Either you're not justified in believing that you're not in Winter Wonderland; or, if you are justified in believing this, it's because that justification rests in part on your justified belief that it will snow in Ithaca next winter.

(10) If you're to have justification for believing (1) on the basis of certain experiences or grounds *e*, then for every *q* which is ''bad'' relative to *e* and (1), you have to have antecedent justification for believing *q* to be false—justification which doesn't rest on, or presuppose any *e*-based justification you may have for believing (1).

(11) (5) is ''bad'' relative to any course of experience *e* you could have and (1).

C. You can't justifiably believe it will snow in Ithaca next winter on the basis of past experiences.

The dogmatist rejects (10), just as she rejects (7). I shall spend most of my time in the next two sections arguing for (10), returning to (7) only at the end. For it seems there are compelling reasons to accept (10),

dogmatism that empiricism is true. But it seems to be part of the point of the dogmatist position that we do not need to know a priori the truth of deeply contingent anti-sceptical propositions like the proposition that that we are not living in Winter Wonderland. So unless there are other reasons to believe in deeply contingent a priori propositions, it seems best to regard dogmatism as a form of empiricism. It is also a version of the kind of internalism discussed in n. 2, since according to the dogmatist seeming to see that *p* can be sufficient justification for belief in *p*. Pryor's preferred version of dogmatism is also internalist in the slightly stronger sense described in the text, but it seems possible that one could be a dogmatist without accepting that internalist thesis. One could accept, for instance, that seeming to see that *p* justifies a belief that *p*, but also think that seeming to see that *q* justifies a belief that *p* iff there is a known reliable connection between *q* and *p*. As I said, even the weaker version of internalism is sufficient to generate a conflict with anti-scepticism and empiricism, provided we just focus on the propositions that can be justifiably believed on the basis of introspective properties.

and hold that the problem with this argument is either with (9) or (11).[10]

2. DOMINANCE ARGUMENTS

The primary argument for (10) will turn on a dominance principle: if you will be in a position to justifiably believe p whatever evidence you get, and you know this, then you are now justified in believing p. This kind of reasoning is perfectly familiar in decision theory: if you know that one of n states obtains, and you know that in each of those states you should do X rather than Y, then you know now (or at least you should know) that you should do X rather than Y. This is a very plausible principle, and equivalent epistemic principles are just as viable. Dominance reasoning can directly support (10) and hence indirectly support (7). (As Vann McGee (1999) showed, the dominance principle in decision theory has to be qualified for certain kinds of agents with unbounded utility functions who are faced with a decision tree with infinitely many branches. Such qualifications do not seem at all relevant here.)

It will be useful to start with an unsound argument for (10), because although this argument is unsound, it fails in an instructive way. Before I can present the argument I need to make an attempt at formalizing Pryor's concept of ''badness''.

> q is **bad** relative to e and $p =_{\text{df}} q$ is deeply contingent, you know p entails $\sim q$, and for any possible evidence e' (that you could have had at the time your total evidence is actually e) there exists a p' such that you know p' entails $\sim q$ and you are justified in believing p' on the basis of e' if e' is your total evidence.

Roughly, the idea is that a bad proposition is one that would be justifiably ruled out by any evidence, despite the fact that it could turn out to be true.[11] Using this definition we can present an argument for

[10] Just which is wrong then? That depends on how ''bad'' is defined. On our final definition (8) will fail, but there are other sceptical arguments, using other sceptical hypotheses, on which (6) fails.

[11] Note that there's a subtle shift here in our conception of badness. Previously we said that bad propositions are those you allegedly know on the basis of your actual evidence (if you know p), even though they are logically consistent with that evidence. Now we say

rationalism. The argument will use some fairly general premises connecting justification, evidence, and badness. If we were just interested in this case we could replace q with (5), the proposition that I'm living in Winter Wonderland, r with the proposition that (5) is false, e with my current evidence, and e' with some evidence that would undermine my belief that (5) is false, if such evidence could exist. The intuitions behind the argument may be clearer if you make those substitutions when reading through the argument. But because the premises are interesting beyond their application to this case, I will present the argument in its more general form.

Rationalist Argument 1

(12) If you are justified in believing (1) (i.e. it will snow in Ithaca next winter) on the basis of e, and you know that (1) entails $\sim(q)$, then you are justified in believing $\sim(q)$ when your evidence is e.

(13) If you are justified in believing r (at time t) on the basis of e, then there is some other possible evidence e' (that you could have at t) such that you would not be justified in believing r were your total evidence e'.

(14) If you are justified in believing r, and there is no evidence e such that e is part of your evidence and you are justified in believing r on the basis of e, then you are justified in believing r a priori.[12]

that they are propositions you could rule out on *any* evidence, even though they are consistent with your actual total evidence. This is a somewhat narrower class of proposition, but focusing on it strengthens the sceptic's case appreciably.

[12] David Chalmers noted that (13) and (14) entail that *I exist* is a priori. He thought this was a bad result, and a sufficient reason to modify these premises. I'm perfectly happy with saying, following Kaplan, that *I exist* is a priori. I don't think this proves rationalism, because I think it's also deeply necessary that I exist. (It's not deeply necessary that Brian exists, but that's no objection to what I just claimed, because it's not deeply necessary that I'm Brian.) This position is controversial though, so I don't want to rest too much weight on it. If you don't think that *I exist* should be a priori, rewrite (14) so that it's conclusion is that you would be justified in believing the material conditional *I exist* $\supset r$ a priori. (Note that since I'm presupposing in the dominance argument that all the salient possibilities are ones in which I have some evidence, and hence exist, it's not surprising that *I exist* has a special status within the theory.) On a separate point, note that I make no assumptions whatsoever here about what relationship must obtain between a justified belief and the

(15) By definition, q is **bad** relative to e and p iff q is deeply contingent, you know p entails $\sim q$, and for any possible evidence e' (that you could have when your evidence is e) there exists a p' such that you know p' entails $\sim q$ and you are justified in believing p' on the basis of e' if e' is your total evidence.

(16) So, if q is bad relative to e and (1), and you are justified in believing (1) on the basis of e, then you are justified in believing $\sim q$ a priori.

(The references to times in (13) and (15) is just to emphasize that we are talking about your current evidence, and ways it could be. That you could observe Winter Wonderland next winter doesn't count as a relevant alternative kind of evidence *now*.)

Our conclusion (16) entails (10), since (10) merely required that for every bad proposition relative to e and (1), you have 'antecedent' justification for believing that proposition to be false, while (16) says this justification is a priori. ('Antecedent' justification need not be a priori as long as it arrives before the particular evidence you have for (1). This is why (16) is strictly stronger than (10).) So, if (10) is false, then one of these premises must be false. I take (15) to define ''bad'', so it cannot be false. Note that given this definition we cannot be certain that (5) is bad. We will return to this point a few times.

Which premise should the dogmatist reject? (12) states a fairly mundane closure principle for justified belief. And (13) follows almost automatically from the notion of 'basing'. A belief can hardly be based in some particular evidence if any other evidence would support it just as well. This does not mean that such a belief cannot be rationally *caused* by the particular evidence that you have, just that the evidence cannot be the rational *basis* for that belief. The dogmatist objects to (14). There is a prima-facie argument for (14), but as soon as we set it out we see why the dogmatist is correct to stop us here.

Consider the following argument for (14), which does little more than lay out the intuition (14) is trying to express. Assume r is such that for

evidence on which it is based. Depending on what the right theory of justification is, that relationship might be entailment, or constitution, or causation, or association, or reliable connection, or something else, or some combination of these. I do assume that a posteriori beliefs are somehow connected to evidence, and if the beliefs are justified this relation is properly called *basing*.

any possible evidence *e*, one would be justified in believing *r* with that evidence. Here's a way to reason a priori to *r*. On any possible evidence, the belief that *r* is true is justified. So I'm now justified in believing that *r*, before I get the evidence. Compare a simple decision problem where there is one unknown variable, and it can take one of two values, but whichever value it takes it is better for one to choose X rather than Y. That is sufficient to make it true now that one should choose X rather than Y. Put this way, the argument for (14) is just a familiar dominance argument.

Two flaws with this argument for (14) stand out, each of them arising because of disanalogies with the decision theoretic case. First, when we apply dominance reasoning in decision theory, we look at cases where it would be better to take X rather than Y in every possible case, *and this is known*. This point is usually not stressed, because it's usually just assumed in decision theory problems that the players know the consequences of their actions given the value of certain unknown variables. It's not obviously a good idea to assume this without comment in applications of decision theory, and it's clearly a bad idea to make the same kind of assumption in epistemology. Nothing in the antecedent of (14) specifies that we can know, let alone know a priori, that if our evidence is *e* then we are justified in believing *r*. Even if this is true, even if it is necessarily true, it may not be knowable.

Second, in the decision theory case we presupposed it is known that the variable can take only one of two values. Again, there is nothing in the antecedent of (14) to guarantee the parallel. Even if an agent knows of every possible piece of evidence that if she gets that evidence she will be justified in believing *r*, she may not be in a position to justifiably conclude *r* now because she may not know that these are all the possible pieces of evidence. In other words, she can only use dominance reasoning to conclude *r* if she knows *de dicto*, and not merely *de re*, of every possible body of evidence that it justifies *r*.

So the quick argument for (14) fails. Still, it only failed because (14) left out two qualifications. If we include those qualifications, and adjust the other premises to preserve validity, the argument will work. To make this adjustment, we need a new definition of badness.

q is **bad** relative to *e* and $p =_{\text{df}}$
(*a*) *q* is deeply contingent;
(*b*) *p* is known to entail $\sim q$; and

(*c*) it is knowable a priori that for any possible evidence e' there exists a p' such that p' is known to entail $\sim q$, and one is justified in believing p' on the basis of e'.

The aim still is to find an argument for some claim stronger than (10) in Sceptical Argument 2. If we can do that, and if as the sceptic suggests (5), the Winter Wonderland hypothesis really is bad, then the only anti-sceptical response to Sceptical Argument 2 will be rationalism. So, the fact that this looks like a sound argument for a slightly stronger conclusion than (10) is a large step in our argument that anti-scepticism plus internalism entails rationalism. (I omit the references to times from here on.)

Rationalist Argument 2

(12) If you are justified in believing (1) (i.e. it will snow in Ithaca next winter) on the basis of e, and you know (1) entails $\sim$ (5), then you are justified in believing $\sim (q)$ when your evidence is e.

(10′) If you are justified in believing r on the basis of e, then there is some other possible evidence e' such that you would not be justified in believing r were your total evidence e'.

(17) If you know you are justified in believing r, and you know a priori that there is no evidence e you have such that you are justified in believing r on the basis of e, then you are justified in believing r a priori.[13]

(18) By definition, q is *bad* relative to e and p iff q is deeply contingent, p is known to entail $\sim q$, and it is knowable a priori that for any possible evidence e' there exists a p' such that p' is known to entail $\sim q$, and one is justified in believing p' on the basis of e'.

(19) So, if q is bad relative to e and (1), and you are justified in believing (1) on the basis of e, then you are justified in believing $\sim q$ a priori.

This is a sound argument for (19), and hence for (10), but as noted on this definition of ''bad'' (11) may be false. If the Winter Wonderland

[13] Again, if you don't think *I exist* should be a priori, the conclusion should be that *I exist* $\supset r$ is a priori.

hypothesis is to be bad it must be a priori knowable that, on any evidence whatsoever, you'd be justified in believing it to be false. But as we will now see, although no evidence could justify you in believing the Winter Wonderland hypothesis to be true, it is not at all obvious that you are always justified in believing it is false.

3. HUNTING THE BAD PROPOSITION

A proposition is bad if it is deeply contingent, but if you could justifiably believe it to be false on the basis of your current evidence, you could justifiably believe it to be false a priori. If a bad proposition exists, then we are forced to choose between rationalism and scepticism. To the extent that rationalism is unattractive, scepticism starts to look attractive. I think Pryor is right that this kind of argument tacitly underlies many sceptical arguments. The importance of propositions like (5), the Winter Wonderland hypothesis, is not that it's too hard to know them to be false. The arguments of those who deny closure principles for knowledge notwithstanding, it's very intuitive that it's *easier* to know (5) is false than to know (1), that it will snow in Ithaca next winter, is true. So, why does reflection on (5) provide more comfort to the inductive sceptic than reflection on (1)? The contextualist has one answer, that thinking about (5) moves the context to one where sceptical doubts are salient. Pryor's work suggests a more subtle answer. Reflecting on (5) causes us to think about *how* we could come to know it is false, and prima facie it might seem we could not know that a priori or a posteriori. It's that dilemma, and not the mere salience of the Winter Wonderland possibility, that drives the best sceptical argument. But this argument assumes that (5) could not be known to be false on the basis of empirical evidence, that is, that it is bad. If it is not bad, and nor is any similar proposition, then we can easily deflect the sceptical argument. However, if we assume internalism, we can *construct* a bad proposition.

The prima-facie case that (5) is bad relative to (1) and our current evidence *e* (I omit these relativizations from now on) looks strong. The negation of (5) is (20), where *H* is a proposition that summarizes the relevant parts of the history of the world.[14]

[14] I assume *H* includes a 'that's all that's relevant clause' to rule out defeaters. That is, it summarizes the relevant history of the world *as such*.

(20) Either $\sim H$ or it will snow in Ithaca next winter.

Now, one may argue that (5) is bad as follows. Either our evidence justifies believing $\sim H$, or it doesn't. If it does, then it clearly justifies believing (20), for $\sim H$ trivially entails it. If it does not, then we are justified in believing H, and whenever we are justified in believing that the world's history is H, we can inductively infer that it will snow in Ithaca next winter. The problem with this argument, however, is fairly clear: the step from the assumption that we are not justified in believing $\sim H$ to the conclusion that we are justified in believing H is a modal fallacy. We might be justified in believing neither H nor its negation. In such a situation, it's not obvious we could justifiably infer (20). So, (5) may not be bad.

A suggestion by John Hawthorne (2002) seems to point to a proposition that is more plausibly bad. Hawthorne argues that disjunctions like (21) are knowable a priori, and this suggests that (22), its negation, is bad.

(21) Either my evidence is not e, or it will snow in Ithaca next winter.

(22) My evidence is e, and it will not snow in Ithaca next winter.

Hawthorne does not provide a dominance argument that (21) is knowable a priori. Instead he makes a direct appeal to the idea that whatever kinds of conclusions we can infer now on the basis of our evidence e we could have inferred prior to getting e as conditional conclusions. So, if I can now know it will snow in Ithaca next winter, prior to getting e, I could have known the material conditional *If my evidence is e, it will snow in Ithaca*, which is equivalent to (21). It's not clear this analogy works, since when we do such hypothetical reasoning we take someone to *know* that our evidence is e, and this may cause some complications. Could we find a dominance argument to use instead? One might be tempted by the following argument.

Rationalist Argument 3

(23) I know a priori that if my evidence is e, then I am justified in believing the second disjunct of (21).

(24) I know a priori that if my evidence is not e, then I am justified in believing the first disjunct of (21).

(25) I know a priori that if I am justified in believing a disjunct of (21), I am justified in believing the disjunction (21).
(26) I know a priori that my evidence is either *e* or not *e*.
(27) So, I'm justified a priori in believing (21).

The problem here is the second premise, (24). It's true that if my evidence is not *e*, then the first disjunct of (21) is true. But there's no reason to suppose I am justified in believing any true proposition about my evidence. Timothy Williamson (2000: ch. 8) has argued that the problem with many sceptical arguments is that they assume agents know what their evidence is. I doubt that's really the flaw in sceptical arguments, but it certainly is the flaw in the argument that (22) is bad.

The problem with using (22) is that the argument for its badness relied on a quite strong privileged access thesis: whenever my evidence is not *e* I am justified in believing it is not. If we can find a weaker privileged access thesis that is true, we will be able to find a proposition similar to (22) that is bad. And the very argument Williamson gives against the thesis that we always know what our evidence is will show us how to find such a thesis.

Williamson proposes a margin-of-error model for certain kinds of knowledge. On this model, X knows that p iff (roughly) p is true in all situations within X's margin of error.[15] The intuitive idea is that all of the possibilities are arranged in some metric space, with the distance between any two worlds being the measure of their similarity with respect to X. Then X knows all the things that are true in all worlds within some sphere centred on the actual world, where the radius of that sphere is given by how accurate she is at forming beliefs.

One might think this would lead to the principle B: $p \rightarrow \mathrm{K}{\sim}\mathrm{K}{\sim}p$, that is, if p is true then X knows that she does not know $\sim p$. Or, slightly more colloquially, if p is true then X knows that for all she knows p is true. (I use K here as a modal operator. K*A* means that X, the salient subject, knows that *A*.) On a margin-of-error model $p \rightarrow \mathrm{K}{\sim}\mathrm{K}{\sim}p$ is false only if p is actually true and there is a nearby (i.e. within the margin of error) situation where the agent knows $\sim p$. But if *nearby* is

[15] There's a considerable amount of idealization here. What's really true is that X is in a position to know anything true in all situations within her margin of error. Since we're working out what is a priori knowable, I'll assume agents are idealized so they know what they are in a position to know. This avoids needless complications we get from multiplying the modalities that are in play.

symmetric this is impossible, because the truth of p in this situation will rule out the knowability of $\sim p$ in that situation.

As Williamson points out, that quick argument is fallacious, since it relies on a too simplistic margin-of-error model. He proposes a more complicated account: p is known at s iff there is a distance d greater than the margin of error and for any situation s' such that the distance between s and s' is less than d, p is true at s'. Given this model, we cannot infer $p \rightarrow \mathrm{K}\sim\mathrm{K}\sim p$. Indeed, the only distinctive modal principle we can conclude is $\mathrm{K}p \rightarrow p$. However, as Delia Graff (2002) has shown, if we make certain density assumptions on the space of available situations, we can recover the principle (28) within this account.[16]

(28) $\quad p \rightarrow \mathrm{K}\sim\mathrm{KK}\sim p$

To express the density assumption, let $d(s_1, s_2)$ be the 'distance' between s_1 and s_2, and m the margin of error. The assumption then is that there is a $k > 1$ such that for any s_1, s_2 such that $d(s_1, s_2) < km$, there is an s_3 such that $d(s_1, s_3) < m$ and $d(s_3, s_2) < m$. And this will be made true if there is some epistemic situation roughly 'halfway' between s_1 and s_2.[17] That is, all we have to assume to recover (28) within the margin-of-error model is that the space of possible epistemic situations is suitably dense. Since the margin-of-error model, and Graff's density assumption, are both appropriate for introspective knowledge, (28) is true when p is a proposition about the agent's own knowledge.

To build the bad proposition now, let G be a quite general property of evidence, one that is satisfied by everyone with a reasonable acquaintance with Ithaca's weather patterns, but still precise enough that it is a priori that everyone whose evidence is G is justified in believing it will snow in Ithaca next winter. The internalist, remember, is committed to such a G existing and it being an introspective property. Now, consider the following proposition, which I shall argue is bad.[18]

[16] If we translate K as $\Box$ and $\sim \mathrm{K} \sim$ as $\Diamond$, (24) can be expressed as the modal formula $p \rightarrow \Box \Diamond \Diamond p$.

[17] Graff actually gives a slightly stronger principle than this, but this principle is sufficient for her purposes, and since it is weaker than Graff's, it is a little more plausible. But the underlying idea here, that we can get strong modal principles out of margin-of-error models by making plausible assumptions about density, is taken without amendment from her paper.

[18] If you preferred the amended version of (11) discussed in n. 12, the bad proposition is *I don't exist, or* (29) *is true*.

(29) I know that I know my evidence is *G*, and it will not snow in Ithaca next winter.

The negation of (29) is (30).

(30) It will snow in Ithaca next winter, or I don't know that I know my evidence is *G*.

It might be more intuitive to read (30) as the material conditional (30a), though since English conditionals aren't material conditionals this seems potentially misleading.

(30a) If I know that I know that my evidence is *G*, then it will snow in Ithaca next winter.

To avoid confusions due to the behaviour of conditionals, I'll focus on the disjunction (30). Assume for now that the margin-of-error model is appropriate for propositions about my own evidence. I will return below to the plausibility of this assumption. This assumption implies that principle (28) is always correct when *p* is a proposition about my evidence. Given this, we can prove (29) is bad. Note that all my possible evidential states either are or are not *G*. If they are *G*, then by hypothesis I am justified in believing that it will snow in Ithaca next winter, and hence I am justified in believing (30). If they are not, then by the principle (28) I know that I don't know that I know my evidence is *G*, so I can come to know (30), so I am justified in believing (30). So, either way I am justified in believing (30). It's worth noting that at no point here did I assume that I knew whether my evidence was *G*, though I do assume that I know that having evidence that is *G* justifies belief in snow next winter.

All of this assumes the margin-of-error model is appropriate for introspective properties. If it isn't, then we can't assume that (28) is true when *p* is a proposition about the introspective properties I satisfy, and hence the argument that (30) is knowable a priori fails. There's one striking problem with assuming a priori that we can use the margin-of-error model in all situations. It is assumed (roughly) that anything that is true in all possibilities within a certain sphere with the subject's beliefs at the centre is known. This sphere must include the actual situation, or some propositions that are actually false may be true throughout the sphere. Since for propositions concerning non-introspective properties there is no limit to how badly wrong the

subject can be, we cannot set any limits a priori to the size of the sphere. So, a priori the only margin-of-error model we can safely use is the sceptical model that says the subject knows that *p* iff *p* is true in all situations. For introspective properties the margin of error can be limited, because it is constitutive of introspective properties that the speaker's beliefs about whether they possess these properties are not too far from actuality. So, there seems to be no problem with using Williamson's nice model as long as we restrict our attention to introspective properties.[19]

If belief in (30) can be justified a priori, and it is true, does that mean it is knowable a priori? If we want to respect Gettier intuitions, then we must not argue directly that since our belief in (30) is justified, and it is true, then we know it. Still, being justified and true is not irrelevant to being known. I assume here, far from originally, that it is a reasonable *presumption* that any justified true belief is an item of knowledge. This presumption can be defeated, if the belief is inferred from a false premise, or if the justification would vanish should the subject acquire some evidence she should have acquired, or if there is a very similar situation in which the belief is false, but it is a reasonable presumption. Unless we really are in some sceptical scenario, there is no ''defeater'' that prevents our belief in (30) being an item of knowledge. We certainly did not infer it from a false premise, there is no evidence we *could* get that would undermine it, and situations in which it is false are very far from actuality.

Since there are no such defeaters, it is reasonable to infer we can *know* (30) a priori. The important premises grounding this inference are the anti-sceptical premise that we can know (1) on the basis of our current evidence, and the internalist premise that we used several times in the above argument. This completes the argument that the combination of empiricism, internalism, and anti-scepticism is untenable.

[19] There's a possible complication here related to the point made in n. 3 about the different ways of formulating the internalist claim. Even if internalism is true, it might be possible for an agent to be radically mistaken about the state of her evidence. For she might think internalism is false, that some extrospective property *F* is evidentially relevant, and be as mistaken as can be as to whether she instantiates *F*. By assuming that internalism is a priori knowable, we avoid that problem. For the agents we are discussing here are, as mentioned in n. 15, convenient idealizations who are aware of a priori facts like the truth of internalism.

4. HOW EXTERNALISM HELPS

It should be obvious how the rationalist can respond to the above argument—by simply accepting the conclusion. Ultimately, I think, that's the best response to this argument. As Hawthorne notes, rationalism is the natural position for fallibilists about knowledge to take, for it is just the view that we can know something a priori even though we could turn out to be wrong. In other words, it's just fallibilism about a priori knowledge. Since fallibilism about a posteriori knowledge seems true, and there's little reason to think fallibilism about the a priori would be false, if fallibilism about the a posteriori is true, the rationalist's position is much stronger than many have assumed.[20] The inductive sceptic also has an easy response: reject the initial premise that in my current situation I know that it will snow in Ithaca next winter. There are other responses that deserve closer attention: first, the inductive sceptic who is not a universal sceptic, and in particular is not a sceptic about perception, and second the externalist.

I said at the start that the argument generalizes to most kinds of scepticism. One kind of theorist, the inductive sceptic who thinks we can nonetheless acquire knowledge through perception, may think that the argument does not touch the kind of anti-sceptical, internalist, empiricist position she adopts. The kind of theorist I have in mind says that the objects and facts we perceive are constitutive of the evidence we receive. So given we are getting the evidence we are actually getting, these objects must exist and those facts must be true. She says that if I'd started with (31), instead of (1), my argument would have ended up claiming that (32) is bad for some *G*.

(31) A hand exists.
(32) A hand exists, or I don't know that I know that I'm perceiving a hand.

She then says that (32) is not deeply contingent, since in any situation where the first disjunct is false the second is true, so it cannot be bad. This response is correct as far as it goes, but it does not go far enough to deserve the name anti-sceptical. For it did not matter to the above

[20] As BonJour (1997) points out, rationalism has fallen into such disrepute that many authors leave it out even of surveys of the options. This seems unwarranted given the close connection between rationalism and the very plausible thesis of fallibilism.

argument, or to this response that (1) is about the future. All that mattered was that (1) was not *entailed* by our evidence. So had (1) been a proposition about the present that we cannot directly perceive, such as that it is not snowing in Sydney *right now*, the rest of the argument would have been unaffected. The summary here is that if one is suitably an externalist about perception (i.e. one thinks the existence of perceptual states entails the existence of the things being perceived) one can accept this argument, accept internalism, accept empiricism, and not be an *external world* sceptic. For it is consistent with such a position that one knows the existence of the things one perceives. But on this picture one can know very little beyond that, so for most practical purposes, the position is still a sceptical one.

The externalist response is more interesting. Or, to be more precise, the externalist responses are more interesting. Although I have appealed to internalism a couple of times in the above argument, it might not be so clear how the externalist can respond. Indeed, it may be worried that by exercising a little more care in various places I could have shown that everyone must accept either rationalism or scepticism. That is the conclusion Hawthorne derives in his paper on deeply contingent a priori knowledge, though as noted above he uses somewhat more contentious reasoning than I do in order to get there. To conclude, I will argue that internalism is crucial to the argument I have presented, and I will spell out how the externalist can get out of the trap I've set above.

One easy move that's available to an externalist is to deny that any facts about justification are a priori. That blocks the move that says we can find a *G* such that it's a priori that anyone whose evidence is *G* can know that it will snow in Ithaca next year. This is not an essential feature of externalism. One can be an externalist about justification and still think it is a priori that if one's evidence has the property *is reliably correlated with snow in the near future* then it justifies belief that it will shortly snow. But the position that all facts about justification are a posteriori fits well with a certain kind of naturalist attitude, and people with that attitude will find it easy to block the sceptical argument I've presented.

Can, however, we use an argument like mine to argue against an anti-sceptic, empiricist externalist who thinks some of the facts about justification *can* be discovered a priori? The strategy I've used to build the argument is fairly transparent: find a disjunctive a priori knowable

proposition by partitioning the possible evidence states into a small class, and adding a disjunct for every cell of the partition. In every case, the disjunct that is added is one that is known to be known given that evidence. If one of the items of knowledge is ampliative, that is, if it goes beyond the evidence, then it is possible the disjunction will be deeply contingent. But the disjunction is known no matter what.

If internalism is true, then the partition can divide up evidential states according to the introspective properties of the subject. If externalism is true, then such a partition may not be that *useful*, because we cannot infer much about what the subject is justified in believing from the introspective properties she instantiates. Consider, for example, the above partition of subjects into the *G* and the not-*G*, where *G* is some introspective property, intuitively one somewhat connected with it snowing in Ithaca next year. The subjects that are not-*G* know that they don't know they know they are *G*, because they aren't. Externalists need not object to this stage of the argument. They can, and should, accept that a margin-of-error model is appropriate for introspective properties. Since it's part of the nature of introspective properties that we can't be *too* badly wrong about which ones we instantiate, we're guaranteed to satisfy some reliability clause, so there's no ground there to deny the privileged access principle I defended above.

The problem is what to say about the cases where the subject is *G*. Externalists should say that some such subjects are justified in believing it will snow in Ithaca next winter, and some are not. For simplicity, I'll call the first group the reliable ones and the others the unreliable ones. If I'm *G* and reliable, then I'm justified in believing it will snow, and hence in believing (30). But if I'm *G* and unreliable, then I'm not justified in believing this. Indeed, if I'm *G* and unreliable, there is no obvious argument that I'm justified in believing *either* of the disjuncts of (30). Since this is a possible evidential state, externalists should think there is no dominance argument that (30) is a priori knowable.

Could we solve this by adding another disjunct, one that is guaranteed to be known if I'm *G* and unreliable? There is no reason to believe we could. If we're unreliable, there is no guarantee that we will *know* we are unreliable. Indeed, we may well believe we are reliable. So there's no proposition we can add to our long disjunction while saying to ourselves, ''In the case where the subject is *G* and unreliable, she can justifiably believe *this* disjunct''. If the subject is unreliable, she may not have *any* justified beliefs about the external world. But this is just to

say the above recipe for constructing bad propositions breaks down. Externalists should have no fear that anything like this approach could be used to construct a proposition they should find bad. This is obviously not a positive argument that this kind of anti-sceptical empiricist externalism is tenable, but it does suggest that such a position is immune to the kind of argument I have presented here.

REFERENCES

BonJour, Laurence (1997) *In Defense of Pure Reason* (Cambridge).

Chalmers, David (2004) 'Epistemic Two-Dimensional Semantics', *Philosophical Studies*, 118: 153–226.

Cohen, Stewart (1999) 'Contextualism, Skepticism, and the Structure of Reasons', *Philosophical Perspectives*, 13: 57–89.

Evans, Gareth (1979) 'Reference and Contingency', *The Monist*, 62: 161–89.

Graff, Delia (2002) 'An Anti-Epistemicist Consequence of Margin for Error Semantics for Knowledge', *Philosophy and Phenomenological Research*, 64: 127–42.

Hawthorne, John (2002) 'Deeply Contingent A Priori Knowledge', *Philosophy and Phenomenological Research*, 64: 247–69.

McGee, Vann (1999) 'An Airtight Dutch Book', *Analysis*, 59: 257–65.

Pryor, James (2000) 'The Skeptic and the Dogmatist', *Noûs*, 34: 517–49.

Williamson, Timothy (2000) *Knowledge and its Limits* (Oxford).

Yablo, Stephen (2002) 'Coulda, Woulda, Shoulda', in Tamar Szabó Gendler and John Hawthorne (eds.), *Conceivability and Possibility* (Oxford), 441–92.

INDEX