# THOUGHTS

# Thoughts

*Papers on Mind, Meaning, and Modality*

STEPHEN YABLO

OXFORD
UNIVERSITY PRESS

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

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

First published 2008

British Library Cataloguing in Publication Data
Data available

Library of Congress Cataloging in Publication Data
Data available

Typeset by Laserwords Private Limited, Chennai, India
Printed in Great Britain
on acid-free paper by
Biddles Ltd, King's Lynn, Norfolk

ISBN 978–0–19–926646–3

1 3 5 7 9 10 8 6 4 2

# *Preface*

This volume contains most of my published work on mind and modality, along with some related work on meaning. Papers in basic metaphysics—things, identity, causation, and the like—are left for a second volume, tentatively entitled *Things*.[1] The main omissions are technical work on truth and logic, and some muckraking, anti-ontological papers written in the last decade or so. One early fictionalist effort has been included, since it concerns the metaphysics of possible worlds.

David Lewis tells us in volume i of his *Philosophical Papers* that he set out to be "a piecemeal, unsystematic philosopher, offering independent proposals on a variety of topics" (p. ix). Unfortunately, "it was not to be". The story of Lewis's fall from unsystematicity can be told to a large extent in his own words. Already in volume i he admits to the existence of eight(!) "recurring themes that unify the papers in this volume" (p. xi) In volume ii, we get the full confession: he has been conducting "a prolonged campaign on behalf of the thesis I call 'Humean Supervenience' " (p. ix).

I too set out to be a piecemeal, unsystematic philosopher. I must say that so far I seem to be doing a better job of it. The papers collected here do not, to my knowledge, reflect any Humean Supervenience-like larger vision. I had thought that this would make the Preface easier to write. Rather than weaving the views expressed into a harmonious whole, it would be enough to show that they were not at war with each other.

I tried. I was going to explain why, if conceivability is such a good guide to possibility, the conceivability of zombies doesn't refute supervenience; and how, if disembodied pain is possible, pain can be a determinable of its physical underpinnings; and more of the same general sort. But, considered as a research topic, *Are these papers consistent*? has very little to recommend it, and I am now officially giving up. ●~~and I am now officially giving up~~. (Your counterpart in the nearest world where the boring introduction is completed would thank me on your behalf, if she could get a message through.) Problems there may be with these papers, but they're to do with relations to the world, not relations with volume-mates.

● Q1

Everyone mentioned in the footnotes: thanks again. Donald and George: you made this book possible. Thanks to Isaac and Zina for not making it impossible. (Here is the condensed version, just for them: blah blah, philosophy, blah blah, philosophy.) Utmost love and gratitude to Mrs Kulkarni.

[1] It may seem odd that the second paper here has "causation" in the title, while the second volume's first paper has "essence" in the title. But the "causation" is mental causation, and the "essence" paper is an attempt to make sense of contingent identity.

## Queries in Chapter 0

Q1. This sentence is repeating twice. Please check.

# *Acknowledgments*

I thank the editors and publishers who have granted permission to reprint the papers appearing in this volume. Dates and places of first publication are as follows.

"The Real Distinction Between Mind and Body," *Canadian Journal of Philosophy*, supp. vol. 16 (1990): 149–201

"Is Conceivability a Guide to Possibility?" *Philosophy and Phenomenological Research* 53 (1993): 1–42

"Textbook Kripkeanism and the Open Texture of Concepts." *Pacific Philosophical Quarterly* 81 (2000): 98–122

"Coulda, Woulda, Shoulda," T. Gendler and J. Hawthorne (eds.), *Conceivability and Possibility* (Oxford University Press, 2002): 441–492

"Beyond Rigidification" appears for the first time in this volume

caps

"No fool's cold: Notes on illusions of possibility," M. Garcia-Carpintero and J. Macia (eds.), *Two Dimensional Semantics* (Oxford University Press, 2006): 327–345

"How in the World?" *Philosophical Topics* 24 (1996): 255–286

"Mental Causation," *Philosophical Review* 101 (1992): 245–280

"Singling out Properties," *Philosophical Perspectives* 9 (1995): 477–502

"Wide Causation," *Philosophical Perspectives* 11 (1997): 251–281

"Causal Relevance," *Philosophical Issues* 13 (2003): 316–329

No Fool's Cold:
Notes on Illusions
of Possibility

# *Contents*

# 1

# The Real Distinction between Mind and Body

> . . . it [is] wholly irrational to regard as doubtful matters that are perceived clearly and distinctly by the understanding in its purity, on account of mere prejudices of the senses and hypotheses in which there is an element of the unknown.
>
> Descartes, *Geometrical Exposition of the Meditations*

## I. SUBSTANCE DUALISM

Substance dualism, once a main preoccupation of Western metaphysics, has fallen strangely out of view; today's mental/physical dualisms are dualisms of fact, property, or event. So if someone claims to find a difference between minds and bodies per se, it is not initially clear what he is maintaining. Maybe this is because one no longer recognizes 'minds' as entities in their own right, or 'substances'. However, *selves*—the things we refer to by use of 'I'—are surely substances, and it does little violence to the intention behind mind/body dualism to interpret it as a dualism of bodies and selves. If the substance dualist's meaning remains obscure, that is because it can mean several different things to say that selves are not bodies.

Any substance dualism worthy of the name maintains at least that

(1) I am not identical to my body;

and probably most dualistic arguments are directed at just this conclusion. But philosophers have been slow to appreciate how unimpressive non-identity theses can be. Assuming an unrestricted version of Leibniz's Law (the indiscernibility of

First

This paper is dedicated to the memory of George Myro; an early version was read at the George Myro Memorial Conference at UC Berkeley, and a later version at the University of Arizona. Thanks to George Bealer, Jonathan Bennett, Paul Boghossian, Janet Broughton, David Copp, Sally Haslanger, Keith Lehrer, Louis Loeb, Vann McGee, Joe Mendola, Sarah Patterson, Larry Sklar, Barry Stroud, William Taschek, Bruce Thomas, David Velleman, Ken Walton, and two anonymous referees for questions, discussion, and advice. Research supported by the National Endowment for the Humanities.

identicals), non-identity is established by *any* difference in properties, however slight or insignificant. If, as seems likely, my body will remain when I am dead, then that already shows that my body and I are not the same thing; and even if my body is not going to outlast me, such could have been the case, which again gives a difference entailing non-identity. You may say that this is dualism enough. But bear in mind that analogous considerations show equally that a statue is not identical to the hunk of clay which makes it up; and this is not normally taken as grounds for a dualism of statue and clay. On pain of insignificance, self/body dualism must mean more than just the non-identity of self and body.[1]

What more could be at issue? For all that non-identity tells us, I might still be necessarily *realized in*, or *constituted by*, my body. For this the obvious remedy is to strengthen (1) to

(2) I could have existed without my body.

But even (2) might mean only that I could have been constituted by a different body than actually: which leaves it open that I am necessarily always constituted by some body or other (as the statue is necessarily always constituted by some hunk of matter). Only with

(3) I could have existed in the absence of all bodies (= material objects),

it seems, do we assert a difference between self and body beyond that obtaining already between statue and clay.

Implying as it does that my existence is not essentially owing to the way in which the world's matter organizes itself, (3) approaches on a genuinely challenging form of dualism. Nevertheless the ambitious dualist will want more; for the possibility remains that I *am* in an extended sense essentially embodied, in that my existence depends on there being either bodies or entities *analogous* to bodies (say, ectoplasmic entities of some sort) whose behavior gives rise to my mental life.[2] Functionalists, for example, can allow that I could exist unaccompanied by anything material, as long as there was *something* present with the appropriate causal organization. But it would be a strange sort of dualism which insisted on my aptitude for existing in the absence of physical bodies, only to lose interest when non-physical 'bodies' were proposed in their place.

In the spirit of Descartes, let us speak of my 'thought properties' as all and only those properties which I am directly aware of myself as possessing.[3] To say that I

[1] Not that this has gone entirely unnoticed. Observing that not only modal but even temporal differences 'establish that a statue is not the hunk of stone, or the congery of molecules, of which it is composed', Kripke allows that 'mere non-identity . . . may be a weak conclusion' ('Identity and Necessity', 101). That is putting it mildly. That people were not identical to their bodies was supposed to be a powerfully antimaterialistic result; but in fact it is compatible with people being as closely bound up with their bodies as statues are with the hunks of matter which compose them!

[2] For the development of this possibility, see Shoemaker 1984*a*, *b*, *c*.

[3] '*Thought*. I use this term to include everything that is within us in such a way that we are immediately aware of it' (CSM II, 113; AT VII, 160). More needs to be said about 'immediate

am embodied in the extended sense seems at least to say that there is an entity, my 'body', which plays host to activities of which I am not directly aware, which activities somehow subserve my state of consciousness. Since these activities are not objects of direct awareness, they ought presumably to be reflected in properties which I possess in excess of my thought properties. So the truth of

(4) I could have existed with my thought properties alone,

should have the consequence that I am capable of existing not only without material things, but in a purely mental condition (i.e., without benefit of anything outside my consciousness). Indeed in a situation in which I possess my thought properties only, it would seem that I exist not just without benefit of anything outside my consciousness, but in the complete absence of any such thing. In recognition of this, we can strengthen (4) to

(5) I could have existed, in isolation, with my thought properties alone,

understood to mean that I could have existed with my thought properties alone and *in the company of no other particulars* (or at least none which are not part of me).

What more could be wanted? Notice that (4) and (5) speak only to how things *could* have been with me, not, or not directly, to how they *are*. In particular, (4) does not rule it out that *as matters stand*, I am constituted by my body, nor even that my body and I are, in the actual circumstances, exactly alike in every ordinary respect. Compatibly with (4) and (5), I might be indistinguishable from my body in point of size, shape, weight, etc., and my body might share all my feelings, thoughts, and desires.

Suppose we call a property *categorical* if its possession by a thing speaks exclusively to what it is like in the actual circumstances, irrespective of how it would, could, must, or might have been (naively, my thought properties are predominantly if not exclusively categorical, and so are most if not all of the traditional primary qualities); and *hypothetical* if it depends on a thing's liability to have been in a certain way different than it is actually (so dispositional, counterfactual, and modal properties, whether mental or physical, are hypothetical).[4] Then the difficulty with (4) and (5) is that they seem to express a merely *hypothetical* difference between myself and my body, whereas an ambitious dualism will want to find us *categorically* unlike. Either I do not possess my body's categorical physical properties, like that of taking up space; or my body does not possess my categorical mental properties, like that of experiencing pain; or both.

Beware of taking the point too far; no reasonable dualist believes that I have *no* categorical physical characteristics, or that my body has *no* categorical mental

awareness' to rule it out that I am directly aware, e.g., of whether my legs are crossed, but this is not a problem I take up here.

[4] For more on the categorical/hypothetical distinction, see Yablo 1987.

properties. Obviously we do. Even if I do not occupy space myself, I do have the physical property of coexisting, and presumably interacting, with something which does (my body); and my body, though perhaps not itself experiencing pain, coexists, and interacts, with something in which pain authentically resides (myself). Thus the claim must be that my categorical physical properties, and my body's categorical mental properties, are always *extrinsic* (*P* is *intrinsic* to *x* if *x*'s possession of *P* speaks exclusively to what *x* is like in itself, without regard to what may be going on outside of *x*, and *extrinsic* otherwise). From this it is a short step to

(6) All of my intrinsic, categorical, properties are mental rather than physical,

and

(7) All of my body's intrinsic, categorical, properties are physical rather than mental.

Assuming that my intrinsic, categorical mental properties are exactly my *thought* properties, the relation between (4) and (6) is as follows: where (4) postulates a counterfactual condition in which I exist with just my thought properties, (6) says that my *actual* condition is in all *intrinsic, categorical*, respects indiscernible from that counterfactual condition of pure disembodiment.

No doubt the exercise could be taken further. For example, (6) and (7) are somewhat overstated. Even the most extreme dualist will admit that she has (e.g.) her existence, and her duration, intrinsically; and these are not plausibly regarded as mental properties. But this is not something we need to bother about just now (see note 15). Another thing we will be leaving aside is the articulation of still stronger versions of dualism, for example the necessitations of (6) and/or (7).[5] What I want to ask now is whether dualism in any of these forms, but especially the fourth, fifth, sixth, and seventh, has any chance of being true.

Subject to correction by Descartes scholars, most of us suppose that Descartes maintained dualism in all the versions given. Unfortunately, his principal argument is nowadays seen as bordering on hopeless, and this on the basis of a single apparently decisive objection, roughly to the effect that de re conceivability is a defective guide to de re possibility.

In this paper, I want to pursue two ideas. The first is that Descartes' argument cannot be faulted simply for relying on an inference from de re conceivability

[5] Obviously I disagree with Bernard Williams when he says that it 'expresses the Real Distinction in its strongest form' to assert the necessitation of (1), i.e., to say that I am necessarily not identical with my body (1978, 117). Assuming that Leibniz's Law holds necessarily, the same can be said of a statue and the hunk of clay which makes it up; for necessarily the one has different modal properties, e.g., being essentially a statue, than the other. Since Kripke, most metaphysicians treat (non-)identity theses as equivalent to their necessitations; if they are right, then what Williams calls the strongest form of the real distinction is actually the *weakest* (or equivalent to it). Certainly it is far weaker than the claim that necessarily self and body have fundamental categorical differences (this is the necessitation of (6) and/or (7)).

to de re possibility; that inference is implicated in too many de re modal claims routinely accepted without qualm or question. So the standard objection needs refinement: even if some de re conceivability intuitions justify de re modal conclusions, others do not, and when the differences are spelled out, Descartes' argument emerges as unpersuasive. The paper's second idea is that, to the contrary, the more the differences are spelled out, the better Descartes' argument looks.

## II. STANDARD PROBLEMS WITH DESCARTES' ARGUMENTS

Descartes believed that he was importantly different from his body, and offered what looks like a variety of arguments for this conclusion. Some of these are less plausible than others. In *The Search After Truth*, there are indications of the much ridiculed 'argument from doubt': I am not a body, 'otherwise if I had doubts about my body, I would also have doubts about myself, and I cannot have doubts about that' (CSM II, 412; AT X, 518). Since I can doubt that my body exists, but not that I do, I am distinct from my body.

Whether Descartes intended precisely this argument or not, it is plainly fallacious, on any readily imaginable interpretation. Perhaps Descartes is reasoning as follows:

*Argument A*

(1) I can doubt that my body exists, but not that I do. (A)
(2) Therefore my body and I have different properties. (1)
(3) Therefore I am not identical with my body. (2)

However, (2) follows from (1) only if 'I can doubt that *x* exists' expresses a property of *x*; which, to judge by its admitted referential opacity, it appears not to do.

On the road to Descartes' true argument is a reading which replaces doubt with rational doubt:

*Argument B*

(1) It is not irrational for me to doubt that my body exists while believing that I do. (A)
(2) If I was identical to my body, this would be irrational. (A)
(3) Therefore I am not identical to my body. (1,2)

Again, there is a problem with the second step. Even if my self and body are identical, reason does not constrain me from feeling doubts about my body which I am unwilling to extend to myself, if I am unaware of their identity, and unaware more generally that it is impossible for the one to exist without the other.

Before I can draw any conclusions from the rational permissibility of doubting body but not self, I need assurances that my essential properties cannot but make themselves felt in my self-conception. Without these assurances, that I am not irrational in maintaining contrasting attitudes toward self and body is as likely due to my ignorance of my true nature as to anything else. Yet if the assurances are somehow obtained, then I already *have* my conclusion and the argument is no longer needed. For if I am unaware of being essentially accompanied by my body, then I am not; and so we are distinct.

Even if (as is sometimes alleged) the argument from doubt cannot fairly be attributed to Descartes, his other and more canonical arguments for the mind/body distinction appear to incorporate a similar fallacy. Thus the crucial assumption of the 'Sixth Meditation●' 's dualistic argument is that ● Q1

> the fact that I can clearly and distinctly understand one thing apart from another is enough to make me certain that the two things are distinct, since they are capable of being separated, at least by God. (CSM II, 54; AT VII, 78)

looks fine to me

Since I can understand, or conceive, myself clearly and distinctly apart from my body, I and my body 'are capable of being separated'; hence we are not identical. As an initial guess about what is going on here, consider:[6] FN:6

*Argument C*

| | | |
|---|---|---|
| (1) | I can conceive myself as existing without my body. | (A) |
| (2) | If I can conceive *x* as existing without *y*, *x* can exist without *y*. | (A) |
| (3) | So it is possible for me to exist without my body. | (1,2) |
| (4) | So I am not identical to my body. | (3) |

Before asking what might be wrong with this argument, notice an important respect in which it improves on the argument from doubt. All that that argument can hope to establish is that I am not identical to my body. But this goes hardly any distance towards justifying the grand claims of Descartes' dualistic metaphysics: that I am capable of existing without my body, that I am capable

[6] In interpreting the quoted passage, I follow the usual practice of disallowing any essential role to God's omnipotence. If we are to take seriously Descartes' doctrine of God's free creation of the eternal truths, God can create anything apart from anything, even *x* apart from *x*; and this without regard to what we may or may not find conceivable. Since that doctrine renders irrelevant conceivability considerations which Descartes clearly sees as crucial, and lends itself to the derivation of conclusions much stronger than he would accept, there is no option but to discount it in the present context. Having done so, the divine power to create *x* without *y* essentially converges on the metaphysical possibility of *x* without *y*. (Cf. Descartes' remark in the 'Geometrical Exposition of the Meditations' that '. . . I introduce the power of God as a means to separate mind and body not because any extraordinary power is needed to bring about such a separation but because the preceding arguments have dealt solely with God, and hence there was nothing else I could use to make the separation' (CSM II, 120; AT VII, 170), and in the 'Sixth Replies' that 'to occur "naturally" is nothing other than to occur through the ordinary power of God, which in no way differs from his extraordinary power—the effect on the real world is exactly the same' (CSM II, 293; AT VII, 435).)

of existing without *any* body, that I am unextended, and so on. Although argument (C) terminates in the non-identity thesis, it reaches it by way of the significantly *stronger* thesis that I am capable of existing without my body (and it would not significantly detract from the argument's plausibility if instead of 'my body,' we had written throughout 'any body'). So if it could be made to work, this argument might yield a dualism worth bothering about.

Nevertheless it seems not to work, and for essentially the same reason as before. According to (2), if I can conceive *x* as existing without *y*, then it *can* exist without *y*. But this is plausible only if I can be sure that I am not, in this act of conception, overlooking an essential property of *x* which renders its existence without *y* problematic or impossible. As Sydney Shoemaker expresses the point, the argument

> . . . involves a confusion of a certain sort of epistemic possibility with metaphysical possibility. In the sense in which it is true that I can conceive myself existing in disembodied form, this comes to the fact that it is compatible with what I know about my essential nature (supposing that I do not know that I am an essentially material being) that I should exist in disembodied form. From this it does not follow that my essential nature is in fact such as to permit me to exist in disembodied form.[7]

Absent prior assurances that his potential for independent existence is not obstructed by unappreciated necessary connections, Descartes is in no position to argue from separability in thought to separability in fact.

Because of difficulties like these, not many philosophers would concede Descartes' claim to have established even so much as his distinctness from his body, much less any *interesting* form of dualism. The problem with Descartes' approach is supposed to be one of principle rather than detail, with the result that most philosophers would now be gravely suspicious of *any* epistemic argument for dualistic conclusions.

## III. THE INDISPENSABILITY OF CONCEIVABILITY

Then what kind of argument *is* available to the dualist? Encouraged by recent advances in modal semantics and metaphysics, modern dualists prefer to base their conclusions in *modal* rather than epistemic premises.

No doubt this is an advance of some sort, but it has worrisome aspects. For one, it ignores that the modal premises stand themselves in need of support, which typically they find in conceivability considerations of the sort that Descartes is faulted for having taken seriously. Insofar as they suppress the role of conceivability in modern-day modal arguments, today's dualists let themselves off the hook on which they hoisted Descartes. Second, once the indispensability

[7] Shoemaker 1984*b*, 155.

of conceivability intuitions is allowed, explanations will be required of how it is that *some* such intuitions may be relied on, even if others cannot. Thus grant that the ancients' ability to conceive (say) heat without motion should not have been taken, even by them, to establish that this was possible. Even so, that I can conceive of *myself* existing without the Washington Monument, *does* seem prima facie to indicate that the one could have existed without the other (or else how do I know that it could?). Presumably there are some unobvious principles at work here that would explain why the one intuition may be relied on, though the other may not. And so far, nothing rules out that when the operative principles are discovered, Cartesian conceivability intuitions will be vindicated.

As already explained, the usual charge against Descartes' argument from his ability to conceive *x* as existing without *y*, to the conclusion that *x* can exist without *y*, is that it seems just to take it for granted that *x*'s essential properties do not go beyond those of which Descartes is aware. Objections of this kind were put to Descartes repeatedly, most notably by Caterus in the *First Objections* and by Arnauld in the *Fourth*. Arnauld asks,

> How does it follow, from the fact that he is aware of nothing else belonging to his essence, that nothing else does in fact belong to it? (CSM II, 140; AT VII, 199)

complaining that

> if the major premise of this syllogism [that the conceivability of *x* without *y* shows the possibility of *x* without *y*] is to be true, it must be taken to apply not to any kind of knowledge of a thing, nor even to clear and distinct knowledge; it must apply solely to knowledge which is adequate. (CSM II, 140; AT VII, 200; interpolation mine),

where here 'adequate knowledge' of a thing is knowledge that embraces all the thing's properties (or at least all its essential properties).[8]

Undeniably this looks like an extremely strong objection, maybe even decisive. How wonderful then that Descartes had the chance to hear it and respond. But before looking at what he says, it's important to see that the problem, if there is one, is extremely general. To be consistent, Arnauld should hold that *all* de re conceivability intuitions are suspect, unless the ideas employed are certifiable in advance as adequate, i.e., as embracing all properties, or at least all essential properties, of their objects. What is not often noticed is that if he is right in this, then an enormous part of our de re modal thinking falls under suspicion.

Distinguish two types of de re modal claim: *positive* claims, to the effect that something *x* has a property *Q* essentially; and *negative* claims, to the effect that something *y* has a property *R* only inessentially or accidentally. Naturally it is the positive claims which have attracted all the attention (e.g., natural kinds have their deepest explanatory features essentially, artifacts have

[8] Certainly this is how Descartes read Arnauld's use of 'adequate', and most modern commentators have agreed. However, true Arnauldian adequacy may be a subtler affair than Descartes appreciated (see Bruce Thomas, 'Conceivability and the Real Distinction').

their original matter essentially, etc.). But it is sometimes just as important if something has a property only accidentally (if, for example, people have their personalities, or their genders, only accidentally); and even where it is not important, it is often true, and often, apparently, *known* to be true. No one would doubt of herself that (e.g.) she could have been born on a different day than actually; and outside of philosophy, no one would question that we know such things. But how do we know them, if not by way of conceiving ourselves without the relevant properties, and finding no difficulty in the conception?

What gives this question its force is the specter of an Arnauldian skeptic who argues from the possible inadequacy of my self-conception, to the conclusion that I am in no position to rule out even such obviously absurd essentialist hypotheses as that I am essentially born on September 30, 1957. If I might, unbeknownst to myself, be essentially accompanied by my body, however clearly I seem to be able to conceive myself without it, why might I not equally be essentially born on that day, however clearly I seem to be able to conceive myself born a day earlier or later? In both cases, the skeptic continues, I have no basis to question the deviant hypotheses unless I have prior assurances that my self-conception embraces all my essential properties. Yet how could I?

In a curious way, this sort of objection reverses a more familiar challenge to *positive* de re modal claims. Suppose I assert that something *x* has some property *Q* essentially, e.g., that this bit of water essentially contains hydrogen. Of course, I might be wrong in supposing that this, or any, water contains hydrogen at all. But now I am interested in the allegation that I might be wrong in another way: I am right that this, like all, water *actually* contains hydrogen, but wrong that it could not have been hydrogen-free. In possible worlds very like this one, it is agreed, it does contain hydrogen; but it is alleged that there may also be worlds in which it contains only oxygen and helium, and yet other worlds in which it contains only helium and aluminum, or helium and aluminum and lead.

Naturally you complain that no grounds have been given for thinking this possible; but then no grounds have been given for thinking it impossible either, and the claim was only that it was possible for all you know. After all, once you have picked *x* out, what essential properties it has is no longer in your hands, but depends entirely on what sorts of counterfactual changes *x* can as a matter of objective modal fact tolerate. How could anything in your way of conceiving *x* rule out that the thing *in itself* is capable of more extreme departures from its actual condition than you had imagined?

Postpone for now the question whether this is a cogent thought, and notice the parallel with Arnauld. Where the present objection is that one cannot rationally exclude that the object of thought has *fewer* essential properties than contemplated, Arnauld contends that one cannot rationally exclude that it has *more* essential properties than contemplated. To answer either objection would be to explain what licenses us in reasoning from premises about what we can conceive of a thing to conclusions about what is possible for it. But let us concentrate on the Arnauldian worry that what I seem able to conceive regarding *x* provides no firm basis for *excluding* properties from *x*'s essence.

Actually, there is a certain irony in Arnauld's position. Leibniz, in his correspondence with Arnauld, alleges that the essence of a thing *x* embraces all of *x*'s properties whatsoever. Since Adam is such that Peter denied Christ some thousands of years after his death, this holds *essentially* of Adam, who would accordingly not have existed had Peter not gone on to be disloyal:

> if in the life of some person and even in this entire universe something were to proceed in a different way from what it does, nothing would prevent us saying that it would be another person or another possible universe that God would have chosen. *It would thus truly be another individual*. . . . (LAC, 60; my emphasis)

Unsurprisingly Arnauld objects:

> . . . I find in myself the concept of an individual nature, since I find there the concept of myself. I have only to consult it, therefore, to know what is contained in this individual concept. . . . I can think that I shall or shall not take a particular journey, while remaining very much assured that neither one nor the other will prevent my being myself. So I remain very much assured that neither one nor the other is included in the individual concept of myself. . . . (LAC, 32–3)

Within limits, it seems obvious, we share Arnauld's assurance. Nobody seriously imagines that it is essential to Arnauld to take, or essential to him not to take, the journey. Still it is hard to see what entitles *him* to the assurance that 'neither one nor the other will prevent me from being myself'. How does Arnauld know that his idea is adequate, i.e., that he is aware of *all* of his essential properties?[9]

Take the Arnauldian skeptic to be the one who questions Descartes' right to reason from separability in conjecture to separability in fact, on the basis that our concepts may for all we know be inadequate; and take the Arnauldian believer to be the one who maintains, against Leibniz, that properly conducted thought experiments can support de re inessentialist conclusions. If the skeptic's doubts are allowed to stand, then it is not obvious how the believer can hope to refute Leibniz's suggestion that my essence takes in all my properties whatsoever! Yet surely we side here with the believer. Even without an answer to the skeptic, I think we feel that he *must* be wrong. Somehow or other, I *must* be in a position to refute the suggestion that I am essentially born on the day of my actual birth, or, even more unbelievably, essentially surrounded by the entire course of actual history.

[9] To complete the irony, something uncomfortably like this Arnauldian point is put to Arnauld by Leibniz himself: '. . . although it is easy to judge that the number of feet in the diameter is not contained in the concept of a sphere in general, it is not so easy to judge with certainty . . . whether the journey which I plan to take is contained in the concept of me, otherwise it would be as easy to be a prophet as to be a geometer . . .' (LAC, 59). Leibniz thinks that individual concepts frameable by finite minds are rarely adequate, much less certifiably adequate. From this it seems to follow that we must view all our conceivability intuitions with extreme suspicion. Nevertheless, Arnauld confidently asserts that he knows that he might not have taken the journey.

## IV. THE CONCEIVABILITY ARGUMENT

What I want to investigate is whether Descartes had even the beginnings of an answer to the Arnauldian skeptic. For this the natural starting point is Descartes' historical controversy with Arnauld, which centers on the conceivability/possibility principle that

If I can conceive of *x* as lacking some property *S*, then it is possible for *x* to exist without *S*.

For such a principle to be valid, Arnauld thinks, it 'must be taken to apply not to any kind of knowledge of a thing, nor even to clear and distinct knowledge; it must apply solely to knowledge which is adequate' (CSM II, 140; AT VII, 200). In response, Descartes appears willing to grant that the mere conceivability, even the clear and distinct conceivability, of *x* as lacking some property *S* is not itself convincing evidence of *S*'s inessentiality. As Arnauld suggests, *x* must be conceived in a suitably comprehensive manner:

> a real distinction cannot be inferred from the fact that one thing is conceived apart from another by an abstraction of the intellect when it conceives the thing inadequately. It can be inferred only if we understand one thing apart from another completely, or as a complete thing. (CSM II, 155; AT VII, 220)

(Cf. also CSM II, 86; AT VII, 121.) But in Descartes' view, Arnauld is wrong to think that our conception needs to be certifiable in advance as 'adequate' (CSM II, 155; AT VII, 220). Admittedly, he may have given a contrary impression when he said that a real distinction could not be inferred by 'an abstraction of the intellect when it conceives a thing inadequately'; but he

> did not think this would be taken to imply that *adequate* knowledge was required. . . . All I meant was that we need the sort of knowledge that we have not ourselves made *inadequate* by an abstraction of the intellect. (CSM II, 155–6; AT VII, 221)

To the question, 'what manner of conception is required if we are to be able to rely on the inference from conceivability to possibility?' Descartes therefore answers that we should conceive *x* 'completely, or as a complete thing'; to which it appears to be a corollary that our conception of *x*, even if not adequate in Arnauld's sense, is free at least of that specific type of inadequacy engendered by intellectual abstraction.

In his day as in our own, Descartes' readers have sensed a confusion in his writings between (i) a conception of myself *in which I do not credit myself with corporeal features*, and (ii) a conception of myself *as lacking in corporeal features*. Sometimes it is said that only the former conception is claimed by, or even available to, Descartes; though it is the latter he needs to argue for the possibility of disembodiment. But Descartes could hardly be clearer that he possesses a self-conception of type (ii); and his repeated insistence on the importance of 'complete conception', and the avoidance of 'abstraction', is, as we will see, directed against just the confusion to which he is so often thought to have succumbed.

To conceive something in a complete manner, Descartes explains, he 'must understand the thing well enough to know that my understanding is *complete*'; and his understanding of a thing *x* is called 'complete' if and only if he understands *x* 'to be a complete thing' (CSM II, 156; AT VII, 221). On its face, this could hardly look less enlightening; but let us pursue it. In general, Descartes calls a thing *complete* if and only if it is a *substance*, that is, it is capable of existing on its own (or, since nothing can exist without God's concurrence, capable of existing unaccompanied by anything but God).[10] Intriguingly, though, he here gives a more elaborate explanation, in which epistemological considerations come strikingly to the fore:

> . . . by a "complete thing" I simply mean a substance endowed with the forms or attributes which *enable me to recognize that it is a substance*. (CSM II, 156; AT VII, 221; emphasis mine)

From this it appears that a complete thing is a substance taken together with a set of its properties meeting some further epistemological condition. And the condition is, that those properties should enable him to recognize their bearer as a substance.

Initially, at least, this is extremely puzzling. In Descartes' view, substances are never directly apprehended, but only by way of their properties (CSM II, 124; AT VII, 176); and whenever we apprehend a property, we may infer that there is a substance in which it inheres (CSM I, 210; AT VIIIA, 25). So when Descartes speaks of 'forms or attributes which enable me to recognize that it is a substance', he cannot, on pain of triviality, mean simply 'forms or attributes which convince me that there is a substance *about*' (*all* properties do that much). Instead, the properties with which the substance is to be thought of as endowed should *present* to me the substance in a way that allows me not merely to recognize that a substance is there, but also *that it is a substance*. Since to be a substance is to be capable of solitary existence, the obvious thought is that *x* is recognizable *as* a substance, if and only if it is presented by way of properties which reveal to me *how it is that x is capable of existing by itself.* In other words, the properties by which *x* is presented are such that *I find it intelligible that it should exist with those properties alone, in the absence, specifically, of any further properties such as might*

[10] Cf. CSM I, 210; AT VIIIA, 25, for the problem about God. Even putting that problem to one side, the definition of substance in terms of capability for unaccompanied existence is still misleading, since a substance is always accompanied by its primary attribute and modes thereof. The natural remedy is to define a substance as an entity which can exist without other substances; however, that would be circular. Some have suggested using another Cartesian notion of substance, that of property-bearer, to give a non-circular definition of the first notion (cf. Loeb 1981, 94).

*require the existence of some other substance*. If and only if *x* is thus presented, do I conceive it in a complete manner, or as a complete thing.[11]

Separability in conjecture does not argue for separability in fact if 'one thing is conceived apart from another by an abstraction of the intellect which conceives the thing inadequately . . . [but] only if we understand one thing apart from another completely, or as a complete thing' (CSM II, 155; AT VII, 220). Thus complete conceivers 'need the sort of knowledge that we have not ourselves made *inadequate* by an abstraction of the intellect' (CSM II, 156; AT VII, 221).[12] Intellectual abstraction is explained in a letter to Gibieuf; it

> . . . consist[s] in my turning my thought away from one part of the contents of [a] richer idea the better to apply it to another part with greater attention. . . . I can easily recognize this abstraction afterwards when I look to see whether I have derived the idea . . . from some richer idea within myself, to which it is joined in such a way that although one can think of the one without paying any attention to the other, it is impossible to deny one of the other when one thinks of both together. (K, 123)

Abstraction, then, consists in *prescinding* from some aspect of an idea, such that one cannot *deny* the ignored aspect 'when one thinks of both together'. Thus it is important that Descartes thinks that he can avoid this with the ideas of himself and his body:

> If I said simply that the idea which I have of my soul does not represent it to me as being dependent on a body . . . this would be merely an abstraction, from which I could form only a negative argument, which would be unsound. But I say that this idea represents it to me as a substance which can exist even though everything belonging to body be excluded from it; from which I form a positive argument, and conclude that it can exist without the body. (K, 152)

Evidently Descartes sees the reliability of his modal intuition as hinging on his avoidance of abstraction in favor of exclusion; and, as we know, he attaches a similar significance to his employment of a complete idea of self. Unsurprisingly, then, the completeness of his self-conception as a thinking thing is strongly associated with his ability to exclude his bodily aspects therefrom:

> . . . the idea of a substance with its extension and shape is a complete idea, because I can conceive it alone, and deny of it everything else of which I have an idea. Now it seems to

[11] Understanding complete things in this way sheds some light on Descartes' otherwise enigmatic remarks about 'incomplete substances', e.g., 'a hand is an incomplete substance when it is referred to the whole body of which it is a part; but it is a complete substance when it is considered on its own. And in just the same way the mind and the body are incomplete substances when they are referred to a human being which together they make up. But if they are considered on their own, they are complete' (CSM II, 157; AT VII, 222). On the other hand, pursuing the Aristotelian resonances of this and similar passages, one might well arrive at a richer notion of 'complete thing' than that suggested here, e.g., entity with an 'internal principle of activity' (see, for example, *Metaphysics* VII. 10 and *De Anima* II. 1).

[12] In these remarks about abstraction, I am greatly indebted to Bruce Thomas's 'Abstraction and Complete Things'.

> me very clear that the idea which I have of a thinking substance is complete in this sense, and that I have in my mind no other idea which is prior to it and joined to it in such a way that I cannot think of the two together while denying the one of the other; for if there was any such within me, I must necessarily know it. (K, 124)

(Cf. also K, 109.) So when Descartes tells us that in conceiving himself as a thinking thing, his idea of himself is complete, he means (at least) that he is capable not only of *prescinding* from thoughts of body in conceiving of himself, but of conceiving himself as *lacking* in bodily aspects.

Now we should ask, exactly how is this supposed to contribute to the reliability of Descartes' modal intuition? Abstraction is not, for Descartes, always and everywhere a bad thing. In *Rules for the Direction of the Mind*, he emphasizes the beneficial effects of freeing our conception of a question 'from every superfluous conception' (CSM I, 51 ff.; AT X, 430 ff.). Nevertheless, abstraction can sometimes lead us astray. Indeed in its most extreme form, where one prescinds in thought from *all* the attributes by which a thing is recognized, abstraction is always problematic. Since 'we do not have immediate knowledge of substances', prescinding in thought from all of a thing's properties leaves us without any proper grasp of what it is that we are thinking about (CSM II, 156; AT VII, 222).

To avoid extreme abstraction, we must conceive our object in terms of some suitable selection of its properties; presumably *which* properties depends on the nature of the investigation. Then what if the investigation is into what is possible for a thing? Given Descartes' rejection of the Arnauldian adequacy requirement, not *all* the thing's properties are needed. But it would seem that we *do* risk a problematic act of abstraction if we prescind in thought from such, or so many, properties that our object cannot be understood as lacking the properties prescinded from (CSM II, 276–7; AT IXA, 216). For this might tempt us into thinking that *x* could exist with no properties other than those included in our conception, when in fact the hypothesis of *x* without those further properties was not fully intelligible. In some such cases, the distinction between *x* and some particular omitted property is merely 'conceptual':

> a *conceptual* distinction is a distinction between a substance and some attribute of that substance without which the substance is unintelligible. . . . Such a distinction is recognized by our inability to form a clear and distinct idea of the substance if we exclude from it the attribute in question. . . . (CSM I, 214; AT VIIIA, 30)

In others, one assumes, what is 'unintelligible' is not *x* without some particular omitted property (e.g., the wax without extension), but *x* as lacking each of a class of omitted properties (e.g., the wax with no particular shape). Quite generally, though, the complete conceiver must take pains not to exclude from her conception of a thing such, or so many, properties that the thing is 'unintelligible' without them. Drawing on the discussion above, we take this to mean that one avoids problematic abstraction by thinking of *x* in terms of properties such that the supposition of its existing with them alone is not repugnant to reason.

Avoidance of abstraction, so understood, is necessary, but not quite sufficient, for complete conception. Remember that complete conception requires knowledge of a thing sufficient to let us know that it is *complete*, and a complete thing is described as 'a substance endowed with the forms or attributes which enable me to recognize that it is a substance' (CSM II, 156; AT VII, 221). Thus complete conception additionally requires that the possibility of *x*'s possessing the indicated properties alone reveals it *as* a substance, i.e., as something that can exist on its own. Gathering these threads together, *x* is conceived as a complete thing, if and only if by way of properties **P** such that

*Containment Condition*: *x* is clearly and distinctly conceivable as possessing the properties in **P** to the exclusion of all others.

*Isolation Condition*: For *x* to possess the properties in **P** to the exclusion of all others is for *x* to exist alone (so that its capability to possess the **P** properties exclusively shows that *x* is a substance).[13]

To *be* a complete thing is accordingly to be a substance *x* taken together with properties **P** in terms of which it is completely conceivable (there is no distinction between being completely conceivable in terms of **P**, and being complete, qua possessor of **P**).

Applying this account to the case of interest, to conceive *myself* as a complete thing is to conceive myself in terms of a set **P** of properties such that I am clearly and distinctly conceivable as possessing **P** alone, where to exist with **P** alone is to exist unaccompanied by any other substance.

Does Descartes think that he can conceive himself as a complete thing in this sense? Indications are that he does think that he can do this, by conceiving himself in terms of what I have called his *thought properties*. Indeed, I suggest that he finds, in the fact that he conceives himself, qua possessor of his thought properties, as a complete thing, all that he needs to reach the conclusion that he could have existed, in isolation, with his thought properties alone. Assuming

[13] Notice how this account preserves the distinction, on which Descartes so much insisted (CSM II, 155–6; AT VII, 220–1), between understanding something adequately, that is, in terms of 'absolutely all the properties which are in the thing', and understanding it completely. To understand *x*, qua **P**, in a complete manner, is not to know *everything* about it, but only enough so that, at least from the subjective perspective, *x* does not appear to need more than what you know about in order to exist. Thus if adequate ideas embrace all of a thing's properties, then complete ideas need not be adequate. (From the definition of completeness it admittedly follows that, at least from the subjective perspective, no property outside of **P** is essential to *x*, so that **P** will *seem* to give an upper bound on the set of *x*'s essential properties. If to be adequate an idea needs only to include *x*'s essential properties, then a complete idea of *x* will at least appear to the thinker to be adequate. Notice though that the thinker need not yet have any views about which of the properties in **P** are essential to *x* and which accidental.)

that by 'that of which I am aware', he means his thought properties, Descartes indicates by his statement that

> . . . it may be that there is much within me of which I am not yet aware . . . [but] *that of which I am aware is sufficient to enable me to subsist with it and it alone*. . . . (CSM II, 155; AT VII, 219; emphasis added)

his satisfaction that his idea of himself as thinking thing meets the containment condition on complete conception.[14] On no further basis than this, he concludes that

> I am certain that I could have been created by God without having these other attributes of which I am unaware. (CSM II, 155; AT VII, 219)

In other words, God could have created him with his thought properties alone. Since he finds nothing in his thought properties to suggest the existence of any other substance (CSM I, 213; AT VIIIA, 29), circumstances in which he has them 'without these other attributes of which I am unaware' will be circumstances in which he exists in isolation (this is the isolation condition). Hence he is entitled to conclude that he can exist, in isolation, as a purely thinking thing. And this completes the argument.

*Argument D*

(1) Qua possessor of my thought properties, I am a complete thing. (A)
(2) I am clearly and distinctly conceivable as possessing my thought properties to the exclusion of all other properties. (1)
(3) If *x* is clearly and distinctly conceivable as possessing exactly the **P** properties, then *x* can exist with exactly the **P** properties. (A)
(4) I can exist with exactly my thought properties. (2,3)
(5) For me to exist with exactly my thought properties is for me to exist in isolation. (1)
(6) I can exist, in isolation, with exactly my thought properties. (1,3)

Here (1) is the claim of completeness, (2) and (5) are the containment and isolation conditions on complete conception, and (3) is the conceivability/possibility

[14] Although there may be a question whether Descartes is fully *consistent* in finding himself clearly and distinctly conceivable as possessing his thought properties exclusively. his idea of God, together with certain principles revealed by the 'natural light', proves God's existence as a non-deceiver, which implies in turn the reliability of his senses; whence his experience as of material objects outside himself guarantees their existence. But then how can he clearly and distinctly conceive himself with his actual thought properties, but *without* the properties that he possesses in virtue of his relations to the external material objects which sense reveals, e.g., the property of having a body? Perhaps the answer is that he conceives himself with no other *intrinsic* properties than his actual thought properties (additional *extrinsic* properties are allowed); or that he conceives himself in sole possession of thought properties *other* than those he possesses in actuality. But there is little textual basis for either suggestion, and both sit poorly with the quoted passage (among others). Thus Descartes' dualistic arguments and his antiskeptical arguments appear to be in some tension; and I am forced to ignore the latter in favor of the former.

caps

principle by which Descartes hopes to infer his aptitude for solitary mental existence from his thinkability in that condition.

## V. THINKING THINGS AS COMPLETE THINGS

Evidently argument (D) is formally valid, so its soundness depends on the acceptability of its premises: the claim (1) that I am, qua possessor of my thought properties, a complete thing, and the conceivability/possibility principle (3) which enables me to conclude, on that basis, that I can exist with my thought properties to the exclusion of all others.

To say that I am, qua possessor of my thought properties **T**, a complete thing, is to make two claims: that I am clearly and distinctly conceivable as possessing the properties in **T** to the exclusion of all others; and that to possess the properties in **T** to the exclusion of all others is to exist in isolation. Now the second of these claims is extremely plausible. If I am not isolated, then there is something *y* outside myself, in virtue of my relations to which it seems inevitable that I should possess properties in excess of my thought properties. But the first claim raises, to begin with, an interesting technical difficulty of which Descartes may not have been explicitly aware.

Is it really conceivable that I should possess my thought properties to the exclusion of *all* others?[15] If we understand the word 'property' so that the class of properties is closed under complementation, then nothing *x* can have the properties in a set **P** to the exclusion of all others, unless for *each* property *S*, **P** contains either *S* or its complement not-*S* (*proof*: if it contains neither, then *x* possesses neither, which is absurd). Yet when Descartes claimed he could have the properties of which he was aware but 'without . . . these other attributes of which I am unaware', he certainly did not suppose that for every property *S*, he was aware of himself either as possessing *S*, or as possessing not-*S* (e.g., he didn't think of himself either as extended or as unextended). For present purposes, then, Descartes would not, or should not, have understood the set of properties as closed under complementation. As it happens, he observed a distinction, between *positive* and *negative characteristics*, or *genuine properties* and *mere privations*, which will secure the needed result, if in the definition of a complete thing we read 'property' as signifying genuine properties only.[16]

[15] Immediately one sees that what Descartes called the 'transcendental' or 'common' attributes (existence, duration, unity, etc.) will have to be allowed as exceptions. For I am not readily conceivable as, e.g., lacking duration. Henceforth, 'property' means non-transcendental property.

[16] Specifically, if **P** is a set of genuine properties (= positive characteristics), then *x* is a complete thing, qua possessor of **P** iff (a) *x* is clearly and distinctly conceivable as possessing the (genuine) properties in **P** to the exclusion of all other (genuine) properties; and (b) only if it possesses (genuine) properties beyond those in **P** can *x* fail to be alone.

Not to minimize its difficulties, several things may be said in defense of the revised definition of a complete thing. For one, it relies on a distinction which is, for all its obscurities, important to Descartes, both in his metaphysics (the cosmological proof of God's existence) and in his epistemology (his doctrine of simple natures and materially false ideas). Second, what Descartes is looking for in a complete thing is a substance fitted out with properties sufficient to render it 'intelligible' as a self-standing entity; and intelligibility is aided not by the accumulation of negative characteristics, but of positive. Third, the old definition leads to results which Descartes clearly does not intend. Consider the negative characteristic $U$ of being unextended: since $U$ is not a member of $\mathbf{T}$, to conceive myself as possessing $\mathbf{T}$ exclusive of all other characteristics is to conceive myself as lacking $U$, and thereby as possessing corporeal properties after all! Fourth, that Descartes never himself contemplates conceivability arguments which trade on negative characteristics such as $U$, suggests that he implicitly understood completeness in terms of positive characteristics. Fifth and lastly, by restricting ourselves to positive characteristics in the definition of a complete thing, we do not limit the definition's generality so much as lessen its redundancy. Let $S$ be positive, so that not-$S$ is negative; then whatever not-$S$ might have accomplished by its presence in $\mathbf{P}$, is accomplished anyway by $S$'s (presumed) absence. So much, at any rate, is to the credit of the revised definition. On the minus side, the revised definition inherits all the obscurity of the distinction between positive and negative characteristics. But let us see where it takes us.

Somewhat tentatively, I propose that to conceive it as possible that *p* is to enjoy the appearance that *p* is possible, by intellectually envisaging a more or less determinate situation in which *p* is understood to obtain.[17] FN:17 Clarity and distinctness come in as follows: I conceive *p*'s possibility *clearly* in proportion as I possess a comprehensive, explicit, and determinate~~,~~ intellectual vision of what the contemplated situation is like, and how it verifies the condition that *p*; and I conceive it *distinctly* in proportion as whatever is not contemplated as pertaining to the envisaged situation may consistently be understood *not* to pertain (equivalently, nothing which is not contemplated as pertaining is rationally required by factors which are contemplated as pertaining).

correct:

takes *possibly*, *p* as its

[17] Read this not as an analysis, but only a partial explication, of conceiving; the idea is to give some indication of what my *conceiving* it as possible that *p* adds to its merely *seeming to me as if* it was possible that *p* (as it might if I was reliably informed that *p* was possible). Among the many questions which I leave open are: what is the precise relation between conceiving (it as possible) that *p*, and believing that it is possible that *p*? and, is conceiving to be understood as a non-modal attitude which (sometimes) takes *possibly*,● *p* as its propositional content, or an intrinsically modal attitude which takes *p* as propositional content? Without prejudice to this latter question, we use 'conceive that *p*' and 'conceive it as possible that *p*' as synonyms; both indicate an act that is veridical if, and only if, it is possible that *p* (analogously, we can agree that the denial that *p* is correct iff it is not the case that *p*, without settling whether denying that *p* is believing that it is not the case that *p*). In this respect, our usage may differ from that of Descartes, who seems willing, at times, to distinguish between conceiving that *p*, and conceiving that *p* is possible (CSM I, 299; AT VIIIB, 351–2). ● Q2

Assuming that my conception of a situation in which I exist in a purely mental condition is not *manifestly* incoherent, the role of distinctness is to show that it harbors no *latent* incoherence, i.e., nothing that would generate manifest incoherence if its consequences were followed out; and the role of clarity is to show further that the conception is free of saving unspecificities which, however FN:18 resolved, would result in incoherence.[18]

Start with distinctness. Nowadays we are familiar with a range of arguments purporting to show that there *is* a latent and unobvious incoherence in the idea of myself existing with my thought properties alone. Arguments like this are associated with Kant and Wittgenstein, and more recently with Ryle, Strawson, behaviorism, and externalist theories of mental content. Those unaware of, or unconvinced by, the considerations offered may claim to find it conceivable that they should exist with only their mental properties; but if those considerations are finally cogent, then they expose all contrary conceptions as incoherent. Obviously Descartes gave little thought to (e.g.) the Private Language Argument; but the general problem of unobvious entailments and the attendant risk of latent incoherence is one to which he was very much alive. As he observes in several places, '. . . there are many instances of things which are necessarily conjoined, even though most people count them as contingent, failing to notice the relation between them' (CSM I, 46; AT X, 422). Nevertheless, Descartes is convinced that his conception of himself with only his thought properties *is* FN:19 relevantly distinct, and so deeply coherent if superfically so.[19] Speaking of his idea of himself as a thinking substance, he claims that he can

> conceive it alone, and deny of it everything else of which I have an idea . . . [I have] no other idea which is prior to it and joined to it in such a way that I cannot think of the two together while denying the one of the other; for if there was any such within me, I must necessarily know it. (K, 124)

Of course, this is the very claim that Kant, Wittgenstein, and the others would want to question (could he deny external objects, if he understood their role in internal time-consciousness, or public language, if he appreciated its connection to the normativity of thought?). Since Descartes understands the distinctness

[18] Two remarks. First, on this reading, there is little real prospect of an *absolutely* clear and distinct conception of the possibility that *p*, but only of a conception ~~appropriately and~~ sufficiently clear and distinct to allay anxieties about incoherence (notice that Descartes regularly treats clarity and distinctness as matters of degree, e.g., at CSM II, 22, 24; AT VII, 33, 35). Second, Descartes' view that there can be clarity without distinctness, but not conversely (CSM I, 208; AT VIIIA, 22), fits naturally with our account; an unclear conception, because it is silent about how certain matters stand, must be indistinct, since it would be incoherent to suppose that they stood in *no* way. Nevertheless, it is convenient to follow Descartes in treating clarity and distinctness as separate requirements.

[19] Admittedly there is a question, already alluded to, how Descartes hopes to reconcile this conviction with his argument for an external world; despite their enormous differences, Descartes, no less than Kant, thinks he sees an unobvious entailment from his subjective condition to external material objects (see note 14).

claim as central to his argument, the issues they raise are exactly those on which he would, or should, have thought the matter rested. Unless we want to speculate on Descartes' response to the Refutation of Idealism, Private Language Argument, etc., the question cannot be pursued much further here. Suffice it to say that there *is* a question, and that anyone who champions Descartes' reasoning has got to assume that it will ultimately be answered in the negative.[20]

To clearly conceive of a situation in which I enjoy purely mental existence is to have a full, explicit, and determinate conception of what that situation would be like, in particular a conception free of saving unspecificities which however resolved would result in incoherence. At one time, I suppose I found it conceivable that there should be a town whose resident barber shaved all and only the town's non-self-shavers.[21] But this conception escaped inconsistency only by remaining unclear; once the barber's shaving habits were specified, the contradiction became obvious. Is my conception of myself as a purely mental being likewise saved from incoherence only by its inexplicitness?

Usually when we are asked to conceive a situation contrary to the actual, we are working to highly partial specifications. Sometimes this leads to trouble, as in the barber case above; but trouble is the exception rather than the rule (which is why nobody complains if my conception of a situation in which Humphrey is President is silent on questions with no apparent bearing on Humphrey's office, e.g., the outcome of the Indian Mutiny). Thus it is all the more striking that when I am asked to conceive myself with exactly my thought properties, this comes very near to providing me with a complete specification of the situation intended: namely, one in which I possess all of the properties which I am in the actual situation directly aware of myself as possessing, and *no more*. Since the properties with which I credit myself in this conception are fixed by my actual state of consciousness, it is not easy to imagine where the problematic indeterminacy could be thought to reside. (Perhaps it goes too far to claim that my conception is fully explicit on every point; certainly, though, it compares extremely well with the competition.)

Tentatively, then, I conclude that I am, qua possessor of my thought properties, a complete thing, and specifically that I can clearly and distinctly conceive myself in a purely mental condition. Postpone for a moment the question whether this

[20] Keep in mind that, unless Descartes can be faulted for not anticipating revolutionary developments to come, it was not *unreasonable* for him to claim distinctness for his self-conception as purely thinking thing; and also that there is no consensus, even among contemporary philosophers aware of those developments, that Cartesian solipsism is latently incoherent. In any case, the usual charge against Descartes' argument is *not* that he was wrong, or irresponsible, to claim consistency for his self-conception as purely thinking thing, but that he was wrong to think that such a claim could bear in any convincing way on his aptitude for purely mental existence.

[21] Sometimes 'conceivable' is used 'factively', so that from *p*'s conceivability as possible, its possibility follows. On this usage, I did not conceive it as possible that the town's barber should shave all and only non-self-shavers; I only seemed to do so. As I use 'conceivable', that *p* is conceivable amounts roughly to its *seeming* to be conceivable in the first sense.

is enough to justify me in believing that I could exist in that condition; and ask instead, does it show, at least, that there can be no justification for *doubting* that I could? That depends on a subtle issue of modal epistemology. Descartes thinks that modal opinions are generated by reason; and this faculty he credits with a certain sort of priority relative to the other faculties: its deliverances are correctable only through the further exercise of reason, never by imagination or sense.[22] Thus correction is impossible if the grounds of our opinion are free of 'internal' difficulties, i.e., difficulties in principle disclosable through the exercise of reason. Insofar as clarity and distinctness are the ultimate 'internal' virtues, that my self-conception as purely thinking thing possesses these virtues would seem to show that nothing could justify me in *doubting* that purely mental existence is possible for me.

For Descartes, 'internal' deficiencies provide the *only* basis on which a modal opinion can be criticized as inaccurate. In recent years, through the work of Saul Kripke, an entirely different basis for criticism has come to light. What Kripke saw, and established beyond reasonable doubt, is that modal opinions *are*, Descartes notwithstanding, correctable through the exercise of sense (e.g., unaided reason finds no difficulty in the conception of a situation involving heat but not motion, but empirical research has turned up facts given which this is seen to be impossible). As a result, purely 'internal' virtues like clarity and distinctness are no longer enough to secure modal intuitions against attack; the most conscientious and clear-headed conceiver can be refuted in a moment by the dullest observer of the passing scene. Obviously this raises new problems, unimagined by Descartes, for the inference from conceivability to possibility, and indeed transforms the issues on which that inference depends in the profoundest way. Nevertheless, the essential lines of his thinking continue to hold up, or so I shall maintain.

## VI. CONCEIVABILITY AND POSSIBILITY

Whether *p*'s lucid conceivability makes it irrational to *doubt* that *p* is possible is one question; whether it rationalizes the *belief* that *p* is possible is another. Why should what I am able to conceive of as occurring be any sort of guide to what can actually occur? Specifically, why should the possibility of my existing in a purely mental condition be thought to follow from my conceivability as existing in that condition? Here there are really two questions, one about why Descartes thought it followed, the other about why we should think so. On the first, I have nothing much to add to what Descartes says himself. In Descartes' view, all of his faculties are the handiwork of an all-powerful, all-knowing, and undeceiving God; and such a God 'surely did not give me the kind of faculty which would

[22] See Loeb 1990.

ever enable me to go wrong while using it correctly' (CSM II, 37–8; AT VII, 54). Not only his general faculty of judgment, but also its specific application to matters of mind and body, is said to be authorized by the veracity of God. To Gibieuf he writes that

> . . . I do not deny that there can be in the soul or the body many properties of which I have no ideas; I only deny that there are any which are inconsistent with the ideas that I do have . . . for otherwise God would be a deceiver. . . . (K, 125)

Not that Descartes supposes that divine veracity entirely precludes erroneous judgments about these topics. Through carelessness, inattention, or failure of imagination, unobvious consequences of my self-conception may escape my notice, with the result that I credit as possible a state of affairs which could never arise. But what apparently cannot happen, compatibly with God's veracity, is that the impossibility of this state of affairs should be forever undetectable, i.e., that what I conceive as possible is not possible, though there is no appreciable defect or difficulty in the conception. Mistakes can indeed arise, but when he reflects carefully on the fact 'that God is not a deceiver, and the consequent impossibility of there being any falsity in my opinions which cannot be corrected by some other faculty supplied by God', Descartes sees that none of these are mistakes which he lacks the means to put right (CSM II, 55–6; AT VII, 80; see also K, 124).

So much for Descartes; why should *we* accept the inference to my possibly existing in a purely mental condition from its conceivability as possible? Strange as it may seem in view of his appeal to God's veracity, Descartes' account contains the seeds of a solution that may find favor even today. Two Cartesian ideas will be important. First, Descartes believes that it is *only* by way of our ideas that we can attain knowledge of what is possible; so that if these ideas are unreliable, then modal knowledge must remain out of reach. Insofar, then, as we credit ourselves with modal knowledge, there is no alternative but to take our ideas as a guide to the modal facts. Already this is hinted at by the continuation of his remark to Gibieuf, quoted above; he says that soul and body cannot have properties inconsistent with his ideas, or else 'God would be a deceiver, *and we would have no rule to make us certain of the truth*' (K, 125; my emphasis). But the point recurs throughout the letter to Gibieuf, intricately interwoven with the appeal to divine veracity that was featured above:

> [you may object that] . . . although I conceive the soul and body as two substances which I can conceive separately, and which I can even deny of each other, I am not certain that they are in reality such as I conceive them to be. Here we have to recall the principle already stated, that we cannot have any knowledge of things except by the ideas we conceive of them; and consequently, that we must not judge of them except in accord with these ideas, and we must even think that whatever conflicts with these ideas is absolutely impossible and involves a contradiction. (K, 124)

In effect, Descartes is saying that we have no other option than to rely on what we find conceivable in drawing conclusions about what can, and what cannot, happen. To be sure, God sees to it that this procedure will not lead us too far wrong. But it is a completely separate point that the vehicle of modal knowledge, if that knowledge can be obtained at all, must be our ideas.[23]

That modal intuition *must* be accounted reliable if we are to credit ourselves with modal knowledge, is a point that retains its plausibility even for those who disagree with Descartes about how that reliability should be accounted for. Unless we are willing to give up our claims to knowledge about what could have happened, though it did not, it seems unavoidable that we treat conceivability as a respectable, if not an infallible, guide to possibility. No doubt we are unhappy with Descartes' attempt at a justification for this policy, and hope to find another, but that is a separate question.[24] The point for now is simply that this *is* our policy; *within limits*, what we are able to conceive as possible, it is our practice to admit as possible. Simple consistency obliges us to consider whether my conception of myself existing with my thought properties alone falls within these limits.

At this point another Cartesian idea becomes important, that we can never reach false conclusions, about modal matters or matters of any other kind, except through the misuse of our faculties. According to the usual story, Descartes claims certain knowledge of this principle on the basis of his certain knowledge of God's veracity. Lacking that recourse, I can't pretend to the same knowledge. Nor do I even believe the principle as stated. What I *do* think is that something *like* a 'no gratuitous error' claim is implicit in our daily practice, in the form of a ban on gratuitous *attributions* of error. Not that doubts must always be backed up by a story about how the thinker has misused her faculties; obviously

[23] Cf. K, 123: 'I am certain that I can have no knowledge of what is outside me except by means of the ideas I have within me. . . . But I think *also* that whatever is to be found in these ideas is necessarily also in the things themselves (emphasis added). Notice too that Descartes considers the argument from his ideas of self and body to be acceptable by *ordinary* standards even *without* the invocation of God's veracity: '. . . had I not been looking for greater than ordinary certainty, I should have been content to show in the Second Meditation that the mind can be understood as a subsisting thing despite the fact that nothing belonging to the body is attributed to it, and that, conversely, the body can be understood as a subsisting thing despite the fact that nothing belonging to the mind is attributed to it. I should have added nothing more in order to demonstrate that there is a real distinction between the mind and the body, since we commonly judge that the order in which things are mutually related in our perception of them corresponds to the order in which they are related in actual reality' (CSM II, 159; AT VII, 226). (See also CSM II, 272 ff.; AT IXA, 207 ff.)

[24] Traditional conceptualism about modal truth might provide one such justification, but other forms of anti-realism could also serve. Neither is anti-realism forced on us; there are options in the theory of knowledge as well. The account in the text is meant to be neutral between these various possibilities, and indeed to allow that none of them is finally convincing. The problem of justifying reliance on our faculties is quite general, and the potential solutions similar, and similarly unsatisfying, across faculties (e.g., perception, memory, logical and mathematical intuition). Obviously it is not, and could not be, our policy to postpone assent to a faculty's deliverances until its reliability is philosophically assured. (In any case, the complaint against Descartes has always been that his appeal to conceivability involves certain *specific* errors, in light of which the proposed conclusion cannot be drawn *in this case*; it should not be allowed to degenerate into a general modal skepticism according to which we are never justified in relying on conceivability considerations, and so never justified in regarding the non-actual as possible.)

it is possible to reach a false conclusion through no fault of one's own. But the suspicion that a judgment, modal or otherwise, is erroneous *does* ordinarily need to be grounded in a reason to think that error in *this* case was significantly likely.

Such a claim is of course commonplace as regards perception (the analogy with perception is meant to be suggestive, not probative). Absent specific and overriding grounds for doubt, perception affords a (defeasible, but that goes without saying) basis for belief. Doubts are of course legitimate if we have independent reason to think that the facts are not as reported, or not of the right kind to be perceived; or that the observer is reckless, or incompetent; or that even competent observers are, on this occasion, liable to go astray. Quite often we can cite some prior error or oversight, which explains the appearances even better than the hypothesis that the facts are as maintained. But what plainly *cannot* be used to justify incredulity is the abstract possibility of error. Obviously this is not meant to constitute any sort of answer to skepticism. The point is only that doubts not backed up in these ordinary ways *are* skeptical doubts; and where skepticism is not at issue, perceptual reports not subject to any but skeptical doubts are accepted, and I will suppose acceptable, as prima facie accurate.

~~Not to minimize their differences, conception~~ Conceiving seems analogous to perception in this respect: absent specific grounds for doubt, *p*'s conceivability as possible prima facie justifies me in the belief that *p* is possible. Outside of philosophy, this would hardly require argument. Imagine that you claim to be able to conceive of a situation in which you exist, but the Washington Monument does not. Assuming that we ourselves find no difficulty in the conception, are we still in a position seriously to question the possibility of yourself without the Monument? Only, it seems obvious, if we can point to some complicating factor of a kind not yet envisaged (imagine your reaction if we said, 'nevertheless, we wonder whether it is really possible', though no further complication suggested itself!). Unless we have it in mind to play the skeptic, and dissent from received standards of evidence, to resist now, without grounds for doubt or the prospect of them, would simply be to reveal ourselves as ignorant of what counts as sufficient reason for belief in cases like this.

With these lessons in mind, return to my conception of myself in a purely mental condition. Naturally I wonder whether this conception is veridical, i.e., whether it is the conception of a real possibility. Presumably this is because I have heard of cases of ~~falsidical conception~~ "mis-conception", cases where people conceived something as possible which was not in fact possible; and I wonder whether my own case might not be like that. For example, I suppose that the ancients had no difficulty in conceiving it as possible that Hesperus should have existed without Phosphorus. From this they might erroneously have concluded that the contemplated situation could have obtained (erroneously, because Venus cannot exist without Venus). Maybe I am making an analogous mistake when I conceive myself as a purely thinking thing, and conclude that this is truly possible for me.

But is the analogy a good one? Remember that the ancients found it conceivable that Hesperus should have existed without Phosphorus, only because they falsely believed that Hesperus and Phosphorus were distinct. What is the mistaken belief which accounts for my erroneous intuition, as the ancients' misjudgment belief that Hesperus was not Phosphorus accounts for theirs?

Reflection on the ancients' mistake points toward the following model of modal error. First, I conceive it as possible that *p*, although *p* is necessarily false. Second, that *p* is necessarily false emerges from the truth of some proposition *q*. Third, I do not realize this, believing instead either that *q* is false, or that it is false that if *q*, then *p* is impossible; and that is how I am able to conceive, erroneously, of a situation in which *p*. Thus:

(a) *q*;
EXTRA # (b) if *q*, then $\Box{\sim}p$; and
(c) my ability to conceive it as possible that *p* is explained by my denial of (a), or else by my denial of (b).

('$\Box p$' means: necessarily, *p*.) Subject to a qualification to be mentioned presently, FN:25 every instance of erroneous conception that I am aware of fits this pattern.[25] For example, the ancients could conceive it as possible that Hesperus should exist without Phosphorus (that *p*) only because they denied the truth (*q*) that they were identical; if some contemporary philosophers, aware of this identity, find themselves capable of the same conception, that must be because they deny the conditional truth that if the identity holds, then Hesperus is impossible without EXTRA # Phosphorus (that *q* only if $\Box{\sim}p$). Similarly, Oedipus may suppose that he could have been King even if Jocasta had never lived (that *p*). But that is because he believes that he is not her son (that *q*); and if he persists in his error, that is because he denies, what for argument's sake we assume to be true, that if she is his mother, then he could not have existed unless she had (that *q* only if $\Box{\sim}p$). EXTRA #
Examples are easily multiplied, but let us return to the case of interest.

Conceivings are prima facie veridical; so I am prima facie entitled to think that I am capable of purely mental existence. The question is whether this prima facie entitlement can be defeated along the lines just indicated. For my modal intuition is erroneous, if there is a proposition *q* such that

(a) *q*;
(b) if *q* then □ (I possess more than my thought properties); and
(c) my ability to conceive it as possible that I should possess no more than my thought properties is explained by my denial of (a), or of (b).

Certainly it would establish that my modal intuition was erroneous if someone was able to *prove* that it could be explained away in the manner indicated. But

[25] Although such a claim might well be correct, I do *not* claim that all modal error whatsoever fits the model (indeed, I leave it open that there might, in principle, be absolutely undetectable modal errors, to which, a fortiori, the model would not apply); my concern is more with the assertability, than the truth, of '*x* is mistaken in conceiving it as possible that *p*'.

so much is not required. To raise legitimate *doubts* about the intuition, it ought to be enough to find a proposition $q$ for which there is good reason to think that the model *may* apply (for in that case, the intuition is potentially explicable on some other basis than that it is true). Call $q$ a *defeater* if there is, plausibly, a significant chance that (a), (b), and (c). Then the objector's challenge is to find a proposition which defeats my intuition of the possibility of purely mental existence.

Admittedly, it may be difficult for the objector to present me with a subjectively convincing example of a defeater. For no proposition $q$ will strike *me* as a defeater unless I can be brought to recognize that I deny something (that $q$, or that $q$ only if $\Box(\sim p)$) that is not improbably true. And this is not something I am likely to admit.[26] But this complication need not detain us for long. For I ought to be able to recognize a proposition $q$, if there is one, such that it is because I deny that $q$, or that if $q$ then $\Box\sim p$, that I am able to conceive it as possible that $p$. Having done so, I must admit that if, contrary to what I suppose, it is true that $q$, and that $q$ only if $\Box\sim p$, then what I find conceivable is not in fact possible. Whether the objection succeeds must now depend on whether the propositions that $q$, and that $q$ only if $\Box\sim p$, possess credibility sufficient to overcome the presumptive reliability of modal intuition.

Certainly there are very many propositions $q$ such that I deny that $q$ is a truth which shows me to be incapable of purely mental existence; for example, I deny this of the proposition that I was born on the planet Neptune. Most such denials are irrelevant, since there is no significant chance that they are in error. But when we turn to propositions $q$ such that it is *not* wildly improbable that $q$ is a truth given which purely mental existence is impossible for me, e.g., that I possess more than my thought properties, or that my mental life is grounded in my physical condition, or that I necessarily possess more than my thought properties, or that I am identical to my body, we are met with a certain difficulty. Going into my thought experiment, *I do not deny* that these are truths which rule out the possibility of my purely mental existence; rather, I *come* to these denials as a *result* of the thought experiment. In some cases, the thought experiment leads me to deny $q$'s truth, in others its tendency is to show that I am incapable of purely mental existence. But in all cases, the conception precedes, and so cannot be explained by, the denial.

To illustrate, it cannot be said that I am able to conceive myself with my thought properties alone only because I initially deny that I possess physical properties, or that my mental life is grounded in my physical nature; or because I initially deny that if these things are true, then I am incapable of purely mental

[26] Perhaps this is why the gap between conceivability and possibility can seem so hard to appreciate from the first-person point of view. Intuitively, 'I can conceive it, but it isn't really possible' has something in common with 'I believe it, but it isn't really true'. If the assertability of '$x$ can conceive that $p$, but it isn't possible that $p$' is connected, as I am suggesting, with the assertability (for some $r$) of '$x$ believes that not-$r$, but $r$', then the reasons for the analogy become clearer.

existence. When I attempt my conception, I *acknowledge* that I possess more than my thought properties, and acknowledge too that my mental life is grounded in my physical nature. And even if I do not acknowledge that these facts reveal me as essentially unfit for purely mental existence, neither do I deny it; indeed, I attempt the thought experiment in order to discover whether denying it would be unreasonable. Similarly I acknowledge that *if* I am identical to my body, then purely mental existence is impossible for me; and although I do not antecedently acknowledge, neither do I antecedently deny, that I am identical to my body. That is what the thought experiment tells me. So far, then, my conception is not in danger of being explained away.[27] FN:27

Someone might object as follows. To erroneously conceive it as possible that *p*, why should I have to go so far as to *deny* the proposition *q* given which *p* is impossible, or to *deny* the proposition that *p* is impossible if *q* is true? Isn't it enough if I am simply ~~*ignorant*~~ unaware that *q*, or ~~*ignorant*~~ unaware that if *q* is true, then *p* is impossible? Thus consider a less demanding model of how erroneous conception can arise: there is a proposition *q* such that

(a) *q*;
(b) if *q* then □ ∼ *p*; and
(c) that I can conceive it as possible that *p* is explained by my ~~ignorance~~ unawareness of the fact that (a), or else by my ~~ignorance~~ unawareness of the fact that (b).

Perhaps the ~~ignorance~~ 'unawareness' model does do a certain justice to cases which the 'denial' model leaves unaccounted for. Imagine, for example, that the medievals, rather than denying that whales were mammals, simply had no opinion either way. Mightn't they still have conceived it as possible, erroneously of course, that they should have been something else (say, fish)? If so, then this gives an example of a ~~falsidical~~ false conception whose explanation lies not in the fact that *q* is denied, but in the fact that it is not believed. Or take the stock example of the conceivability of Goldbach's conjecture, on the assumption that it is, unbeknownst to anyone, false; then it is not because I deny, but because I am ~~ignorant,~~ unaware, that some even number is not the sum of two primes, that I can conceive it as possible that the conjecture holds.[28] FN:28

[27] Following Kripke, many philosophers believe that (K) for all *z*, if *z* is the zygote from which I actually derive, then I am necessarily derived from *z*. George Bealer observes that if (K) is independently credible, the proposition *q* that I derive from *z* (my actual zygote) looks like a defeater of my modal intuition; for *q* is independently credible, and given the independent credibility of (K), so, apparently, is the conditional proposition that if *q*, then □ (I posess more than my thought properties). The problem is avoided if by 'I could have existed in a purely mental condition', I mean only that I could have existed in that condition over some considerable part of my life. Admittedly this response is superficial, if, as may appear, I am now open to a second 'reduplication' argument of the sort typically offered for (K). But that argument, or so I claim, proves difficult to formulate.

[28] Not everyone agrees that I can conceive it as possible that Goldbach's conjecture is false. Some will see me as confusing conceivability as metaphysically possible with some sort of epistemic possibility, e.g., it is not known, or not knowable a priori that, not-*p*; and others will claim to find

Now we have a less demanding, and perhaps (see the last note) a more realistic, model of how modal intuition goes wrong. The objector's challenge is to identify a proposition $q$ for which there is a significant chance that the model applies. ~~Now~~ you may say that nothing could be easier. Let $q$ be the proposition that I am incapable of purely mental existence; then as long as my intuition is still sub judice, there might seem to be a significant chance that (a) $q$ is true, (b) if $q$, then I am incapable of purely mental existence (this is obvious), and (c) my ignorance of (a) explains my ability to conceive myself in a purely mental condition.

Nevertheless, I take it that it gives me *no* real reason not to trust my intuition that I am capable of purely mental existence, to be told that that intuition might be due, in part, to my ignorance of what might, for all I know, be the fact that I am *in*capable of purely mental existence. After all, it could equally be said that I am able to conceive it as possible that I should have had a different birthday, only because of ignorance about the necessity of my actual birthday. In either case, the most that can be claimed is that *if* the alleged defeater is true, and, e.g., it *is* necessary that I am born on September 30, then if I had not been ignorant of that fact, I would not have found any earlier birthday conceivable. And that is hardly an *objection*; no more than it is an objection to the veridicality of my perceptual impression that there are ducks present, that if I am wrong, and they are decoys, then my ignorance of that fact would figure in the explanation of how I was able to suppose that they were ducks.

Relating this intuitive response to the formal model takes some care; two points need to be distinguished. Even if we allow there is a significant chance that I am incapable of purely mental existence, there seems little chance that my ignorance of this fact could constitute the *explanation* of how I was capable of a contrary conception; the explanation must cite some *other* error or oversight to which my mistaken conception can then be attributed. But that is not the important

a confusion between the conceivability of $p$, and its not being inconceivable (van Cleve 1983). To the former, let me say that although 'conceivable' *can* be used to indicate epistemic possibility, what I mean by it is 'conceivable as metaphysically possible'. To the latter, my response is to question the existence of any sharp or principled distinction between its being conceivable, and its not being inconceivable, that $p$. Practically all conception is in some degree vulnerable to defeat; as the vulnerability increases, and our consciousness of it grows, we back off the 'conceivability' claim and incline more and more to the 'not inconceivable' formulation. But we do this in response to the gradual intensification of a concern that is never wholly absent, the concern that our intuition is liable to defeat by eventualities which we are not yet in a position to rule out. In the example given, this concern is deeply felt, and that accounts for our admitted hesitation in calling it conceivable that Goldbach's conjecture should be false. But I submit that I feel the same *sort* of hesitation, to a lesser degree, in claiming the conceivability of a situation in which I exist but my car does not (skeptics should consult their TV listings for reruns of the situation comedy 'My Mother the Car'). Having said that, I agree that in the Goldbach example we feel so *much* hesitation that the conceivability claim is at least tendentious. If anything, this strengthens my argument: the 'ignorance' model extends the 'denial' model only in cases where I simply cannot tell whether ($q$ & (if $q$ then $\Box\sim p$)); but in those cases, I am presumably reluctant, anyway, to claim that $p$ is conceivable. Thus it is mainly in connection with *uneasy* conceivability intuitions that the ~~'ignorance'~~ model opens up new possibilities for criticism (this is a point I return to). EXTRA # 'unawareness' point; for even if a more informative explanation is constructed, it carries little force if its plausibility depends on the *prior* concession that my conception is not improbably mistaken ~~falsidical~~ (this would be like explaining away my perception as of ducks by saying that they were not improbably decoy ducks, decoy ducks being the usual explanation of misleading ~~falsidical~~ duck appearances). If there is any point to saying that the faculty of modal intuition is presumptively reliable, it is that one may not *assume* that a given intuition is untrustworthy, in making the case that it should not be trusted. Only if there is some basis *independent of the issue under dispute* to suspect that my refusal of some relevant proposition *s* really does put me out of touch with the facts, does the allegation that *s* provide a reason for doubt.[29] FN:29

To summarize, the objector's challenge is to identify a proposition *q* for which there are *independent* grounds to suspect that my conceivability as a purely thinking thing is explained by my ignorance of the following fact: that *q* is a truth which shows that this is impossible for me. To see some of the difficulties involved, compare our imaginary medievals' intuition that whales could have been other than mammals, with my own intuition that I am capable of existing in a purely mental condition. Believing (as I will suppose) that whales might after all *turn out* to be mammals, and that if so they are mammals necessarily, these medievals should at least have felt some considerable *uneasiness* about their conception of whales as possibly not mammals. After all, they knew of a hypothesis *q*, amenable to straightforward empirical verification, whose truth would, by their own lights, reveal their conception as not veridical. However I know of no empirical hypothesis *q*, for which it is antecedently at all probable that if *q* is true, then I couldn't have existed as a purely mental being (which is why I do not feel the same sort of mistrust of my modal intuition as I am supposing that the medievals must have felt of theirs). Insofar, indeed, as *q* is an empirical hypothesis with some reasonable chance of coming out true, it is antecedently highly unlikely that if *q*, then I couldn't have existed as a purely mental being. And something like this holds more generally, I claim, of proposed defeaters *q* of my modal intuition: the better the chances are that *q* is true, the

[29] Someone might object that any consideration with the power to exhibit my unacceptance of the proposition that *q*, or that *q* only if *p* is impossible, as putting me out of touch with the facts, is, eo ipso, *not* 'independent of the issue under dispute' (since that issue is whether or not *p* is possible). But for *s* to be credible independently of the issue of whether it is possible that *p* does *not* mean that *s*, if credible, cannot confer credibility on the thought that *p* is impossible; it means that *s*'s credibility is not owing to the *prior* credibility of that thought. Undoubtedly the distinction here alluded to raises fascinating and difficult problems, but its reality seems unmistakable. For example, observation gives me evidence that this swan is black, and this then confers credibility on the thought that not all swans are white. But the fact that 'this swan is black' would not be credible, if 'not all swans are white' were not also credible, has no tendency whatever to show that the former owes its credibility to the latter; and it would be absurd to complain, on the ground that my observation is misleading if all swans are in fact white, that I have failed to supply a reason 'independent of the issue' of whether all swans are white, to think that this swan I am now looking at is black. So I see no in principle difficulty about finding reasons independent of the issue of whether *p* is possible, for propositions which, if credible, would call *p*'s possibility into question.

worse the chances for truth of the conditional proposition that if *q*, then purely mental existence is impossible for me.

Let us consider cases. Maybe *q* is the proposition that I have physical properties, where these may be either intrinsic or extrinsic. Since there is independent reason to think that I possess at least extrinsic physical properties, *q* is independently probable. But I am not aware of *any* independent reason to think that if I possess physical properties, *even if only extrinsic ones*, then I am incapable of purely mental existence. Someone might claim that there is independent reason to suspect that I have *intrinsic* physical properties, specifically, extension; and that there is independent reason to think that if so, then I am extended necessarily, and therefore cannot exist in a purely mental condition. About the second half of this, I am extremely doubtful. Like most people, I regard it as significantly likely that I *am* extended; somehow, though, this does not seem to inhibit me in conceiving myself as a purely thinking thing. But then I need positive argument that this intuition of being possibly-but-not-actually unextended is accountable to some prior error, before I can accept that any independent credibility attaches to the conditional hypothesis stated; otherwise, the objection comes to nothing more than the unsubstantiated allegation that my intuition may be wrong.[30] Of course, the conditional hypothesis becomes virtually certain if we let *q* be the proposition that I am *necessarily* extended. But now it is *q* itself which wants independent evidence.

Better, then, to look for a proposition *q* which, though not itself modal in character, has modal consequences (specifically, that I am incapable of purely mental existence). Perhaps there are independent grounds to suspect that I am *the same thing* as my body; and that if so, I am incapable of existing with my thought properties alone. (Certainly we seem to have an awful lot in common: shape, size, mass, and so on.) But what is meant by 'same thing'? If it means 'identical', then the first conjunct needs some reason to believe it. However *categorically* similar my body and I may be, this gives grounds to suspect only that we are *coincident*, not that we are identical.[31] Evidence that we were moreover *identical* would presumably be evidence that my body and self agreed on a wide range of *non*-categorical or *hypothetical* properties, specifically on those for which the agreement is not readily accounted for in terms of our admitted categorical similarity. Counterfactual and dispositional properties are therefore of limited importance, and evidence of identity must to a large extent be evidence of *modal*

[30] Some philosophers may find it tempting to argue as follows: whatever is extended is some sort of body; and whatever is a body is necessarily so, and so necessarily extended. But this reasoning is vitiated by an ambiguity in 'is a body'. If it means 'is of the metaphysical kind << body >>', then it is not antecedently plausible that whatever is extended is a body; if it means 'has the categorical properties of something of that kind, e.g., extension, mass, solidity . . . ,' then it is not antecedently plausible that bodies are necessarily bodies (see section VII).

[31] See Yablo, 1987. Assume for the sake of the objection that there are no temporal differences between my self and my body~~, e.g.,~~ my body doesn't antedate me, nor is it going to outlast me.

similarity; yet this can only come from conceivability considerations, which seem in fact to argue the other way! If 'same thing' is understood so as to require sharing of categorical properties only, the problem is merely relocated. Now the apparent categorical similarity of my self and my body *does* give independent grounds for suspecting that we are the 'same thing'; only there is no longer any reason to think that our being so rules out the possibility of my purely mental existence.

So there seems to be at least this much difference between our imaginary medievals' intuition of the possibility that whales are not mammals and my intuition of the possibility of purely mental existence: unlike the medievals, I am not aware of any independently credible hypothesis whose truth might be supposed, on independent grounds, to have the consequence that my intuition is incorrect. Surely it would be absurd and irrational for me to defer, in these conditions, to the abstract possibility that I am in error?

Maybe not. To this point, I have been pretending that the medievals were aware of certain *specific* issues (e.g., are whales mammals?), amenable to independent investigation, whose unfortunate resolution would, by their own lights, have exposed their modal intuition as incorrect. But it may be truer to the normal progress of our dialectic that the conceiver is *not* specifically aware of her conception's vulnerability to its eventual defeater, until the defeater comes along and does its work. Before the discovery of genes, for example, the thought may not have been readily available that scenarios in which animal life was organized along some non-genetic basis risked exposure as not only false, but impossible, by the progress of science. None of this is to deny that the concept of an animal must *somehow* 'prepare the ground' for the eventual recognition that (e.g.) animals necessarily propagate their kind by way of genes. But it is striking how unaware it is nevertheless possible to be of the vulnerability of one's modal intuition to what emerges, in the end, as its defeater. And now the objection comes, can't that be how it is with my intuition of the possibility of purely mental existence?

Ideally lucid conception, were it obtainable, would anticipate, I suppose, every possible scenario for defeat (even before Mendel, ideally lucid conceivers would have realized that such-and-such discoveries would rule it out that animal life could be organized on a non-genetic basis). So understood, ideally lucid conception is not within our powers; but what we are being asked to consider is how very far short of ideal lucidity our conceivings can fall, and how risky it therefore becomes to assume that no defeater would come into view, if it were somehow obtained. Of course this risk cannot be generally prohibitive, or *no* modal intuition would be trustworthy; so the idea must be that there is something in the nature of the thought that I exist in a purely mental condition to encourage the suspicion that in *this* case, if ideal lucidity were achieved, defeat would follow.

What might that something be? Recent work in the theory of content has turned up a variety of cases in which there is a significant gap between

FN:32 grasping a thought content, and appreciating the truth-conditions it induces.[32] Misidentifying a cunningly groomed shrub as Brendan Sullivan, I entertain the content that *that* individual is not a potted plant; although I consider that I have thought something true in just those worlds in which Sullivan is not a potted plant, I am mistaken: it is the *shrub's* (actual and counterfactual) condition that matters. In this example, indexicality appears to be the culprit. But the same phenomenon can arise with contents that are not on their face indexical: for example, contents involving natural kind concepts; or concepts sensitive to community consensus regarding the use of their standard linguistic expression; or, more generally, concepts whose contribution to truth-conditions is affected by factors potentially unavailable to the thinker.

Since conceivability is a matter of thought content, and possibility a matter of truth-conditions, contents for which this gap is especially large (call them 'schematic' contents) seem peculiarly apt to figure in delusive conception. Continuing the example above, I experience no difficulty in conceiving as possible a situation in which *that* individual employs dishonest methods, because I fail to see that that must be a situation in which the *shrub* does this. And now the objector argues that if my conception of myself in a purely mental condition is similarly schematic, then that should provoke concern about its accuracy. For as content grows more schematic, it constrains truth-conditions less and less; and the risk accordingly grows that the truth-conditions present difficulties to which the content offers no clue. Defeat is therefore to be expected, in the form of a proposition *q* spelling out the worldly facts which guide the transition from (benign) content to (malignant) truth-conditions.

However, I will need an argument before I concede that my I-thoughts are dangerously schematic. Remember that conceivability intuitions vary in subjective insecurity, according to how seriously one regards the threat of defeat. Ideally the potential defeaters have been identified, and then our confidence depends on the probability we attach to their being truths incompatible with the intuited possibility. But even when the threat is open-ended, subjective insecurity continues to track expectation of defeat, via our sensitivity to schematic elements FN:33 in the content entertained.[33] Other things being equal, one would expect the perilously schematic character of my I-thoughts to express itself in a pronounced insecurity about my intuition of disembodied existence. Then why do I not feel this insecurity? Various explanations may be possible; but the natural explanation is that my I-thought is not perilously schematic after all.

[32] By 'thought content' I mean something peculiarly suited to the classification of the thinker's subjective condition or internal point of view; and by 'truth-conditions' I mean something which determines truth-values over all possible worlds. Depending on context, a given content can induce a variety of truth-conditions; the larger this variety, the larger the gap referred to in the text.

[33] Thus I feel far more confident of my intuition that there could have been a planet without mountains, than of my intuition that there could ^be ~~have been (e.g.)~~ a force proportional to the ^square root of the mass of the object it acted on ~~(as gravity was supposed to be)~~.

Even if true, the allegation that my I-thoughts are schematic could hardly be decisive. After all, an I-thought serves as the content of my intuition that I could have existed without Margaret Truman; yet this does not suffice to call the intuition into serious doubt. Thus the abstract possibility of trouble en route from content to truth-conditions, unsupplemented by a plausible scenario about how that possibility might be finding expression in the actual case, seems not to be enough. ~~But when~~ we will need a proposition spelling out how the envisaged complications are supposed to arise (maybe I *am* Margaret Truman); and this proposition would seem to be none other than a defeater, whence a defeater is required in any case. To consider the obvious example, someone might believe that my I-concept picked out the entity, whatever it was, activities in which constituted the ultimate basis of *these thoughts*; and she might attempt to explain my modal intuition away by citing my failure to bear in mind that: I am the entity so described, and the entity so described is my body (but for this, I would see that the possibility of my enjoying purely mental existence is ruled out by my body's inability to do the same). But this is just to offer as defeater the proposition *q* just formulated, which must then be subjected to the same scrutiny FN:34 as any other proposed defeater. Like them, it is found wanting.[34]

caps

Vague and circumstantial worries about its potential for defeat cannot overcome the prima facie credibility of the Cartesian intuition. Pending the discovery of a specific defeater, I propose to acquiesce in the intuition, and to conclude that purely mental existence is possible for me.

## VII. CATEGORICAL DUALISM

Maybe you think that this conclusion is in order, or maybe you think it goes too far; in either case, it is important to remember that the full-blooded Cartesian dualist maintains something even stronger. At the outset I distinguished between the *hypothetical* dualism which asserts the *separability* of selves from bodies and the *categorical* dualism which claims to find fundamental categorical differences between self and body, such as would imply their *separation in fact* (as statue and clay are *not* separate in fact). In the *Meditations*, at least, Descartes betrays little appreciation of this crucial distinction:

> [because] I know that everything which I clearly and distinctly understand is capable of being created by God so as to correspond exactly with my understanding of it . . . the fact

[34] As I see it, no independent credibility attaches to *q*'s first conjunct *r*: I am the entity activities in which constitute the ultimate basis for these thoughts. Like many people, I acknowledge that my thoughts are owing to occurrences in my body; yet this does not inhibit me in conceiving myself in a disembodied state. Absent positive argument for *r*, to offer it as independently credible is simply to forget the presumptive reliability of this modal intuition. Perhaps the needed credibility is thought to flow from an (acknowledged?) a priori equivalence between my I-concept and the descriptive condition given. But if this a priori equivalence obtained, then ~~presumably~~ I ought to know it; and not-*r* ought accordingly to strike me as a priori false (which I submit it does not).

> that I can clearly and distinctly understand one thing apart from another is enough to make me certain that the two things are distinct. (CSM II, 54; AT VII, 78)

If by 'distinct', Descartes means non-identical, then from the premise that *x* and *y* 'can be made to exist in separation', it does indeed follow that they are distinct. But if by 'distinct' he means *categorically* unlike, then he simply does not explain how this is supposed to follow from mere separability. Thus Descartes' argument for hypothetical dualism, even if accepted, is far from establishing the *categorical* dualism which asserts actual separation on the basis of fundamental categorical dissimilarity.

Now Descartes does of course believe that there are important categorical differences between mind and body, in particular that *minds are not extended*, and that *bodies do not think*.[35] To be sure, the situation is somewhat complicated by his contention that there is something—the mind/body union, sometimes called the 'human being' or the 'man'—which is both thinking and extended. But this latter doctrine should not distract us from Descartes' repeated assertions that the components of this union are in categorical respects utterly disparate; the body is extended and unthinking ('I have never seen or perceived that human bodies think; all I have seen is that there are human beings, who possess both thought and a body' (CSM II, 299; AT VII, 444)), and the mind is unextended and thinking ('I deny that true extension as commonly conceived is to be found in God or in angels or in our mind or in any substance which is not a body' (K, 239)). As for the man, he is thinking and extended *only* in the sense that he has disjoint parts of which one is an unextended thinker and the other unthinking and extended:

> . . . the question is whether we perceive that a thinking thing and an extended thing are one and the same by a unity of nature. That is to say, do we find between thought and extension the same kind of affinity or connection that we find between shape and motion, or understanding and volition? Alternatively, when they are said to be "one and the same" is this not rather in respect of unity of composition, in so far as they are found in the same man, just as bones and flesh are found in the same animal? *The latter view is the one I maintain*. . . . (CSM II, 286; AT VII, 424; emphasis added)

Apparently, then, no single thing is both thinking and extended, in the way that triangularity and rectilinear motion *can* jointly inhere in a single thing. Adapting Descartes' terminology slightly, we can say that the mind thinks by *nature*, the man by *composition*, that is, by inheritance from a proper part which thinks by nature; similarly the body is extended by nature, the man by composition. Then Descartes' view is that *the thing which thinks by nature is not extended, and the*

[35] This is not to say that the 'real distinction,' as Descartes conceives it, expresses a categorical dualism; indeed in its canonical statements (e.g., CSM I, 213; AT VIIIA, 28–9) it sounds decidedly hypothetical. What I do think is that, first, Descartes *was* a categorical dualist, second, he was *seriously* unclear about how far categorical dualism outreaches hypothetical, and, third, he had *some tendency* to read his arguments for the real distinction as having established categorical dualism inter alia.

*thing which is by nature extended does not think*. Using ‘I’ for the thing which thinks by nature, and ‘my body’ for the thing which is by nature extended, Descartes maintains that I am not extended, nor does my body think.[36] FN:36

Given the centrality of these ideas in his thought, it is little short of astonishing that our problem here is not so much to evaluate his reasoning, as to discover what his reasoning could have been. In a work as late as the *Principles of Philosophy* (1644), Descartes still shows a tendency to slide over from separability into separateness:

> . . . even if we suppose that God has joined some corporeal substance to . . . a thinking substance so closely that they cannot be more closely conjoined, thus compounding them into a unity, they nonetheless remain really distinct. For no matter how closely God may have united them, the power which he previously had of separating them, or keeping one in being without the other, is something he could not lay aside. . . . (CSM I, 213; AT VIIIA, 29)

From this one surmises that Descartes takes mind’s separability from body to indicate that even in the actual circumstances, soul and body are at best ‘closely conjoined’. If ‘conjoined’ is understood so as to permit overwhelming categorical similarity (i.e., if statue and clay are ‘conjoined’), then the conclusion follows, but has no tendency to show that mind is actually unextended, or that body does not think. But if, as seems enormously likelier, ‘conjoined’ entities are categorically unlike, then it needs an argument to show that my separability from my body entails that we are, as matters stand, at best ‘conjoined’. #

Most reconstructions of Descartes’ reasoning make appeal here to the premise that whatever is embodied is necessarily so.[37] If accidental embodiment is impossible, then from my possible disembodiment, my actual disembodiment evidently follows. Whether Descartes takes the impossibility of accidental embodiment as a premise or not, in the present context its plausibility owes entirely to a confusion between (a) being *a body*, in the sense of belonging to the kind <<body>>, and (b) being *embodied*, in the sense of being categorically (almost) indiscernible from something of that kind. Admittedly bodies are necessarily bodies (and so necessarily embodied); thus if embodiment implies being a body, nothing can be embodied without being necessarily so. But to assume that FN:37 EXTRA #

[36] Three remarks. First, someone might question whether Descartes would assent to ‘I am unextended’, on the ground that ‘I’ refers not to the mind but to the man. Actually, Descartes’ usage is unclear on this point, but even if I were the man, it would remain that I was categorically distinct from my body, for I think, and my body does not. In the text, we use ‘I’ for the thing which thinks by nature; on that usage, Descartes does of course think that he is unextended. Second, Descartes does sometimes allow that mind can be in a very weak sense ‘extended’, simply by being in union with body (we might say that mind can be extended ‘by union’); however, he makes it very clear that extension by union is not extension in any real or familiar sense (K, 119, 143). Third, when *x* is said to possess an attribute *P* ‘by nature’, this does not mean that *P* is a *nature* of *x*, and in particular it does not mean that *P* is a property that *x* cannot exist without, or a basis for its other properties. (For example, it is by nature that the plank is warped, but being warped is not the plank’s nature.)

[37] See van Cleve 1983; Hooker 1978; and Schiffer 1976.

only bodies can be embodied is simply to beg the question against the categorical monist who alleges that what I am is not a body but an embodied *person*, whose categorical properties are (approximately) those of a certain thinking body, but with modal characteristics all its own.

Nothing remains for Descartes but a last-ditch appeal to the idea that thought *excludes* extension, i.e., that nothing can possess both 'by nature'.[38] Since I undoubtedly think, it would follow that I am not extended. Some slight evidence that Descartes is attracted to this reasoning comes from his response to a 1647 pamphlet published by his former disciple Regius. Regius remarks that:

> . . . if we are to follow some philosophers, who hold that extension and thought are attributes which are present in certain substances, as in subjects, then since these attributes are not opposities but merely different, there is no reason why the mind should not be a sort of attribute co-existing with extension in the same subject, though the one attribute is not included in the concept of the other. . . . (CSM I, 294–5; AT VIIIB, 342–3)

Descartes replies that if we are talking about

> . . . attributes which constitute the natures of things, it cannot be said that those which are different, and such that the concept of the one is not contained in the concept of the other, are present together in one and the same subject; for that would be equivalent to saying that one and the same subject has two different natures—a statement that implies a contradiction, at least when it is a question of a simple subject (as in the present case) rather than a composite one. (CSM I, 298; AT VIIIB, 350)

(Cf. also CSM II, 159; AT VII, 227.) Apparently, then, whatever is both thinking and extended *must* be composite:

> A composite entity is one which is found to have two or more attributes, each one of which can be distinctly understood apart from the other. For, in virtue of the fact that one of these attributes can be distinctly understood apart from the other, we know that the one is not a mode of the other, but is a thing, or attribute of a thing, which can subsist without the other. A simple entity, on the other hand, is one in which no such attributes are to be found. . . . [Hence] that which we regard as having at the same time both extension and thought is a composite entity, namely a man—an entity consisting of a soul and a body. (CSM I, 299; AT VIIIB, 350–1)

Ignore as irrelevant the question why a composite of soul and body should be expected to inherit thought and extension, strictly understood, from its thinking and extended parts (or why, if it did, its unthinking and unextended parts should not equally confer on it thoughtlessness and unextension!). Our problem is much more basic. If by a 'composite' entity, Descartes means a subject of distinctly comprehensible attributes, then that reduces his complaint against Regius, that whatever has distinctly comprehensible attributes is composite, to the triviality

[38] Notice that this assumption, if Descartes were prepared to make it, would render his subtle conceivability argument entirely superflouous. But then why bother with the conceivability argument at all?

that whatever has distinctly comprehensible attributes, has them. To restore the complaint's substance, 'composite' needs to be returned to its original meaning, namely 'divisible into disjoint parts'. But now the same old worries recur. How does Descartes know that only what is divisible into disjoint parts can possess both thought and extension? What is the argument which rules it out that *some* things, for example, *people*, are thinking and extended *by nature*, that is to say, FN:39 *otherwise* than by separate inheritance from categorically disparate components?[39]

Obviously it would be disappointing if Descartes had to resort here to a neo-scholastic prejudice according to which every undivided entity *must* be characterized by a single fundamental nature, of which all of its other (non-transcendental) properties are modes. For positive argument, he seems driven back on his apparent conviction that nothing is conceivable as thinking and extended, except by postulating a separation of that thing into purely extended and purely thinking parts; in a word, that nothing is conceivable as thinking and FN:40 extended *by nature*.[40] Whatever the precise bearing may be of inconceivability on impossibility (this is something we have not discussed), the problem with this lies elsewhere: it is simply not obvious, if it ever was, that nothing is conceivable as by nature both thinking and extended. In the *Essay Concerning Human Understanding*, Bk. 4, Ch. 3, Part 6, John Locke ~~proposes~~ suggests that it is

> not much more remote from our comprehension to conceive that God can, if he pleases, superadd to matter a faculty of thinking, than that he should superadd to it another substance with a faculty of thinking.

From the ensuing controversy, it emerges that Locke was at any rate not simply FN:41 *wrong* about this, even for his own time.[41] Even if subsequent discussion has done little to relieve the obscurity of bodily thought, it has tended to confirm Locke's judgment that the combination is not strictly inconceivable. But then I am still without a reason to believe that I am not extended, or that my body does not think.

[39] Admittedly, the clear and distinct comprehensibility of thought without extension does establish that the former is not, in Descartes' sense, a mode of the latter (for modes are not intelligible without their associated attributes (CSM I, 210–1; AT VIIIA, 25)). But on an intuitive level, that we can understand thought without extension shows at most that thought is *not necessarily* a way of being extended, not that it is *necessarily not* a way of being extended. Thus there is room, which the categorical monist may want to take up, for the view that it is in fact by being appropriately extended that one thinks, though thinking can in principle proceed on some other basis, or on no basis at all (thought remains an attribute, since it does not *presuppose* extension). On such a view, we do indeed possess distinctly comprehensible attributes by nature. But it is equally open to the categorical monist to say that we possess thought and extension both by nature, although thinking is *not* a way of being extended, nor conversely.

[40] Perhaps Descartes' 'incompatibilist' remarks in the 1647 *Notae*, and his 1648 statement to Burman that we possess clear conceptions of mind and body 'as two substances which not only do not entail one another but are actually *incompatible*' (CB (28); emphasis added), reflect a belated recognition of the gap between his premises and his conclusion.

[41] See Yolton 1983, for a detailed history of the debate Locke provoked by ~~his innocent~~ this remark.

## REFERENCES AND ABBREVIATIONS

Adam, C. and Tannery, P. (eds.) (1964–76). *Œuvres de Descartes*. Paris: Vrin/C.N.R.S. (= AT).

page #s correct ● Q3

Cleve, J. van (1983). 'Conceivability and the Cartesian Argument for Dualism', *Pacific Philosophical Quarterly* **64**, ●pp. 35–45

Cottingham, J. (ed.) (1976). *Descartes' Conversation with Burman* (Oxford: Clarendon Press) (= CB).

Cottingham, J., Stoothoff, R., and Murdoch, D. (eds.) (1985). *The Philosophical Writings of Descartes* (I, II) (Cambridge: Cambridge University Press) (= CSM).

Hooker, M. (1978). 'Descartes' Denial of Mind–Body Identity'. In Hooker (ed.), *Descartes: Critical and Interpretive Essays*, Baltimore: Johns Hopkins, pp. 171–85

Kenny, A. (ed.) (1981). *Descartes: Philosophical Letters* (Minneapolis: University of Minnesota Press, ~~1981~~) (= K).

Kripke, S. (1977). Identity and Necessity'. In Schwartz (ed.), *Naming, Necessity, and Natural Kinds*. (Ithaca, NY: Cornell University Press), ●pp. 66-101

● Q4

● Q5

Locke, J. (1996) *Essay Concerning Human Understanding* ●(Indianapolis, IN: Hackett Publishing Company)

Loeb, L. (1981). *From Descartes to Hume: Continental Metaphysics and the Development of Modern Philosophy*. Ithaca, NY: Cornell University Press, ~~pp. 3–43.~~

_____ (1990). 'The Priority of Reason in Descartes', *Philosophical Review* **99**, pp. 3-43

Mason, H. T. (ed.) (1967). *The Leibniz–Arnauld Correspondence*. Manchester: Manchester University Press (= LAC).

Schiffer, S. (1976). 'Descartes on his Essence', *Philosophical Review* **85**, pp. 21–43.

Shoemaker, S. (1984*a*). 'Embodiment and Behaviour'. In Shoemaker, *Identity, Cause, and Mind*, New York: Cambridge University Press, pp. 113–138.

_____ (1984*b*). 'Immortality and Dualism'. In Shoemaker, *Identity, Cause, and Mind*, pp. 139–158.

_____ (1984*c*). 'On an Argument for Dualism'. In Shoemaker, *Identity, Cause, and Mind*, pp. 287–308.

Thomas, B. 'Abstraction and Complete Things' (unpublished manuscript, University of Michigan).

_____ 'Conceivability and the Real Distinction' (unpublished manuscript, University of Michigan).

Williams, B. (1978). *Descartes: The Project of Pure Enquiry*. Atlantic Highlands, NJ: Humanities Press ~~pp. 293–314.~~

Yablo, S. (1987). 'Identity, Essence, and Indiscernibility'. *Journal of Philosophy* **84**. pp.293-314

Yolton, J. (1983). *Thinking Matter*. (Minneapolis: University of Minnesota Press.

## Queries in Chapter 1

Q1. Please check and confirm the author correction here.

Q2. Author edit is not clear. Please check.

Q3. Please check and confirm whether the page numbers inserted by us is fine or not.

Q4. Page number is missing here, Please check.

Q5. This reference seems incomplete, Please check.

# 2

# Is Conceivability a Guide to Possibility?

> . . . because I find absence of incompatibility, because, that is, I am without a certain perception, I am to call my idea compatible. On the ground of my sheer ignorance, in other words, I am to know that my idea is assimilated, and that, to a greater or lesser extent, it will survive in Reality.
>
> F. H. Bradley, *Appearance and Reality*

## I. INTRODUCTION

Some propositions are "possible": the way they represent things as being is a way things metaphysically *could* have been. Other propositions are not in this sense possible. How do we tell the difference? Or more particularly, of the possible propositions, how do we tell *that* they are possible?[1] Hume's famous answer is that it is

> an establish'd maxim in metaphysics, *That whatever the mind clearly conceives, includes the idea of possible existence*, or in other words, *that nothing we imagine is absolutely impossible*.[2]

And if there is a seriously alternative basis for possibility theses, philosophers have not discovered it. So it is disappointing to realize that Hume puns on

Research for this paper was supported by the National Endowment for the Humanities and the Social Sciences and Humanities Research Council of Canada. For discussion and comments I am grateful to Paul Boghossian, Jim Brown, Jim Conant, John Devlin, Graeme Forbes, Hannah Ginsborg, Danny Goldstick, Sally Haslanger, David Hills, Eileen John, Gideon Rosen, Larry Sklar, William Taschek, David Velleman, Ken Walton, Nick White, Crispin Wright, Catherine Wright, and audiences at Davidson College, Queens University, Wayne State University, and Ohio State University.

[1] Sometimes, of course, this is easy. If a proposition $p$ is true, and known to be, then its possibility can be inferred from $p$ itself. The problem is to find grounds for thinking a proposition possible which is *not* known to be true, most obviously because it is false.

[2] Hume 1968, p. 32. The maxim seems to say that conceivability *suffices* for possibility. This is implausibly strong, so I propose to (mis)interpret Hume as claiming only that the conceivable is *ordinarily* possible and that conceivability is *evidence* of possibility.

"establish'd". What the maxim *is*, is entrenched, perhaps even indispensable. But our *entitlement* to it has often been questioned.[3]

Doubts about Hume's maxim have a variety of historical sources. Some date back as far as Descartes's claim that, since he can conceive himself in a purely mental condition, his essence is only to think. "How does it follow", Arnauld asks, "from the fact that he is aware of nothing else belonging to his essence, that nothing else does in fact belong to it?"[4] Others are as recent as the discovery by Kripke and Putnam of necessary truths knowable only *a posteriori*:

> we can perfectly well imagine having experiences that would convince us . . . that water *isn't* $H_2O$. In that sense, it is conceivable that water isn't $H_2O$. It is conceivable but it isn't logically possible! Conceivability is no proof of logical possibility.[5]

Between times we find Reid and Kneale warning that if a proposition is true "for all you know", then you will find it conceivable whether it is possible or not. More than can be appreciated from a few examples, though, pessimism about conceivability methods has been a consistent theme in philosophy. When Mill says that

> our capacity or incapacity of conceiving a thing has very little to do with the possibility of the thing in itself; but is in truth very much an affair of accident, and depends on the past history and habits of our own minds,[6]

he sums up the position of many authors and the instinctive assumption of many more.

Yet throughout this complicated history runs a certain schizophrenia in which, the theoretical worries forgotten, conceivability evidence is accepted without qualm or question. Hume's own famous applications of his maxim are a case in point. There is nothing necessary about the uniformity of nature, he says, for

> We can at least conceive a change in the course of nature; which sufficiently proves, that such a change is not absolutely impossible.[7]

Causes are not strictly necessary for their effects, because the latter are conceivable as uncaused; nor are they sufficient, since it is always conceivable that the effect should not ensue. Whatever our other differences with Hume, these arguments are normally credited with a good deal of persuasive force. Or consider a case from the philosophy of language. As everyone knows, 'Alexander's teacher' is not a rigid designator. How, though, does everyone know this? Well, we imagine a counterfactual situation in which Aristotle refuses Phillip's call, or

[3] Arthur Pap writes that "there is no objection to the imaginability criterion simply because there is no alternative to it" (1958, p. 218). As the advice not to abandon a leaky lifeboat, this has its points. As factual observation, though—well, such objections are extremely common.

[4] CSM II, p. 140.

[5] Putnam 1975, p. 233. See Putnam 1990, pp. 55–7, for second thoughts.

[6] Mill 1874, book II, chapter V, section 6.

[7] Hume 1968, p. 89.

dies of dysentery on the way to Macedonia. Such imaginings would be irrelevant to the rigidity of 'Alexander's teacher' if they gave no evidence of possibility.

In the actual conduct of modal inquiry, our theoretical scruples about conceivability evidence are routinely ignored. Double-think, though, is not the method of true philosophy. Those of us willing to be persuaded of *p*'s possibility by our ability to conceive it (and that is most of us, most of the time) should face the issue squarely: is this procedure ill-advised? There will be just one constraint on the discussion. Because the topic is not knowledge in general but knowledge of possibility, we will confine ourselves to problems or supposed problems *peculiar* to conceivability arguments. Such arguments have been charged, for instance, with trading on a confusion between two senses of 'could'; with implicit circularity; and with misclassifying most or all *a posteriori* impossibilities as possible.

Other, more sweeping, objections have also been raised. Two in particular deserve mention now, if only to put them aside for purposes of this paper. First is the traditional skeptical lament that

> (S) No independent evidence exists that conceivability is a guide to possibility—no evidence obtainable without reliance on the faculty under review.

indent like (A) on p50

True enough. But there is no independent evidence either that perception is reliable about actuality; and if the worst that can be said about conceivability evidence is that it is as bad as perceptual evidence, that may be taken as grounds for relief rather than alarm. Now though comes the objection from naturalism:

> (N) Granted the unavailability of any philosophically satisfying *reason* to think that perception is adequate to its task, we see at least how it *could* be. In fact perception itself brings word of sensory mechanisms seemingly hard at work monitoring external conditions. By contrast "we do not understand our own must-detecting
> FN:8 faculty".[8] Not only are we *aware* of no bodily mechanism attuned to reality's
> FN:9 modal aspects, it is unclear how such a mechanism could work even in principle.[9]

indent like (A) on p50

Taken in a suitably flat-footed way, these claims are again true enough. But the same could be said about various other faculties, ~~notably~~ such as logical and mathematical intuition; and to judge by our reaction there, they constitute a reason less for mistrusting the faculty than for reconsidering either the nature of the target facts
FN:10 or the nature of our access to them.[10]

So much for the grand-scale objections. Ultimately they are going to require answers, but answers of a kind that the experience of philosophy has accustomed us to doing without. At any rate they are not the objections that concern me, or, I think, Arnauld, Reid, Kneale, etc. Two differences seem important. First, these

[8] Blackburn ~~1986, p. 119~~ 1993, p.52. [9] Cf. Wright 1986, pp. 206–7.

[10] For a sense of the possibilities, see Coppock 1984; Forbes 1985, chapter 9; Bealer 1987; Sidelle 1989; and ~~Yablo, forthcoming~~ Yablo 1992.

philosophers seem prepared to bracket worries that arise also with other accredited ways of knowing, the better to focus in on what might be *specially* problematic about conceivability. Second, rather than simply deploring the *absence* of reason to think that conceivability *is* a guide to possibility, Arnauld and company offer *positive* evidence that it is *not* a guide. If the problem with conceivability methods was only that we could not prove, or explain, their reliability, then maybe we could live with that. But the problem is supposed to be that they are demonstrably *unreliable*.

## II. CONCEIVABILITY AND THE MODAL-APPEARANCE TEST

What conceivability is, is a question I hope to put off as long as possible. For now we can get by on a single assumption, one perhaps implicit in Hume's remark quoted above:

this is a quote and should be in small type as on p39

> whatever the mind clearly conceives, includes the idea of possible existence, or in other words, . . . nothing we imagine is absolutely impossible.

As often when Hume takes himself to be saying the same thing twice, he seems here to be saying two quite *different* things:[11] FN:11

(a) what we imagine or conceive is *presented* as possible;
(b) what we imagine or conceive *is* possible.

Where (b) claims for conceivability a certain *external* relation with possibility, (a) looks more like a partial *analysis* of conceivability, namely, that to conceive or imagine that *p* is *ipso facto* to have it seem or appear to you that possibly, *p*. Without suggesting that Hume would go quite so far, I take the idea to be that conceiving is in a certain way analogous to *per*ceiving. Just as someone who perceives that *p* enjoys the appearance that *p* is *true*, whoever finds *p* conceivable enjoys something worth describing as the appearance that it is *possible*.[12] FN:12 In slogan form: *conceiving involves the appearance of possibility*.

[11] The classic example: "we may define a cause to be *an object, followed by another, and where all the objects similar to the first are followed by objects similar to the second*. Or in other words *where, if the first object had not been, the second never had existed*" (Hume 1963, section VII, part II).

[12] Two notes about terminology. First, here and below I use 'conceive that *p*' and 'find *p* conceivable' essentially interchangeably. (But see note 59.) Second, 'conceive' has a *factive* sense—in which I don't find *p* conceivable unless it is possible—and 'perceive' is *normally* factive—I don't perceive that *p* unless *p*. In *this* paper, both terms are to be understood *non*factively. Thus 'I perceived that *p* but it wasn't true' and 'although I found *p* conceivable, it turned out to be impossible' are perfectly in order. Out of order, though, will be the following: 'I *veridically* perceived that *p*, but *p* wasn't true' and 'although I *veridically* conceived that *p*, it turned out to be impossible'.

Before trying to make the slogan clearer, let me stress that I am *not* touting the "appearance of possibility" as all there is to conceiving,[13] or the only thing conceiving can ever be. Far from trying to give the notion's one true meaning, my aim right now is ~~only~~ to distinguish conceiving in the sense that matters from various *other* cognitive operations doing business under the same name. ~~For~~ as I will be interpreting it, the question whether conceivability is a guide to possibility concerns the kind of conceivability that *advertises* itself as such a guide. This means that if there are kinds of conceivability that do *not* portray *p* as possible—and there are—then for my purposes *it will not matter* if their modal guidance should prove unreliable.

caps

FN:13

FN:14

Following in the tradition of Brentano, Husserl, and most recently Searle,[14] suppose we take seriously the idea that many intentional states and acts—beliefs, desires, and perceptual experiences, for instance—have *satisfaction* conditions. And let us agree that these satisfaction conditions are at least in some cases the conditions under which the state in question is *true* or *veridical.* So, your belief that DeGaulle liked cheese is true just in case he did, and my perceptual impression that rain is falling is true just in case rain *is* falling.

FN:15

From examples like these, one obvious conjecture would be that the *truth conditions* of an intentional state (assuming it has some) are a function of its *content.*[15] But consider someone who, rather than *believing* that DeGaulle liked cheese, inwardly *denies* that he did. This person's state has the same content as the believer's, yet unlike the believer's state it is correct just in case DeGaulle did *not* like cheese. So, *the truth conditions of an intentional state cannot be read off its content alone*; as the examples of denial, expectation, and memory show, the state's psychological mode or manner is also relevant. This is crucial because *one* thing I will be taking "conceivability involves the appearance of possibility" to mean is that the truth conditions of an act of conceiving that *p* include, not the condition that *p*, as in perception, but the condition that *possibly p*. From now on I will express this by saying that *p*'s possibility *representatively appears* to the conceiver.

Maybe the analogy with perception can be carried a little further. Perceiving that *p* has in general the effect of *prima facie* justifying, to the subject, the belief that *p*, and thereby *prima facie* motivating that belief. Here the parenthetical "to the subject" is to indicate that the perceiver need only *feel* himself to be *prima facie* justified, that is, to cancel any suggestion that he is *prima facie* justified in fact. Thus someone convinced that he can judge sexual orientation at a glance might feel justified, on the basis of casual inspection, in believing a neighbor

[13] Later I'll suggest that the conceiver enjoys this appearance *in a certain way*—by imagining a more or less determinate situation of which *p* is held to be a correct account.

[14] See Dreyfus 1982, "Introduction" and *passim*; and Searle 1983.

[15] Thus Searle: "To know the [representative content of an intentional state] is already to know [its satisfaction conditions], since the representative content gives us the conditions of satisfaction, under certain aspects, namely those under which they are represented" (Dreyfus 1982, p. 266).

to be heterosexual, yet without possessing the slightest real evidence that this is so. That his neighbor is heterosexual epistemically appears to this person, even though his feeling of justification is quite misplaced. To have a word for this, let's say that *p epistemically* appears to me when some representative appearance I enjoy *prima facie* motivates me to believe that *p*, by making that belief seem to FN:16 me *prima facie* justified.[16]

That our two readings of "appears" are compatible should be clear; the state that moves me to *believe* that rain is falling can surely be one with the *truth conditions* that rain is falling. Perhaps it could even be argued that the representative reading *entails* the epistemic one, for instance, that a visual experience with the truth conditions that *p* cannot help but move the experiencer FN:17 to believe that *p*.[17] However that may be, the readings are distinct, for the converse entailment fails: for me and I assume for others, it is only *epistemically* that the bull looks as though it is about to charge, or the car sounds like it's FN:18 not going to make it through the winter.[18] (Suppose that your car *does* make it through the winter. Then your experience has tempted you into a false belief, but it's not as though you were the victim of a sensory illusion!)

Back to the slogan "conceivability involves the appearance of possibility", should "appearance" here be taken in the representative sense or the epistemic one? *Both* senses are intended. Just as to perceive that *p* is to be in a state that (i) is veridical only if *p*, and that (ii) moves you to believe that *p*, to find *p* conceivable ^that is to be in a state^ ~~which~~ (i) is veridical only if possibly *p*, and^ (ii) moves you to ^that believe that *p* is possible.

With this background I can state my position. When we look at the standard objections to Hume's maxim, we find that they presuppose conceivability-notions that are neither mandatory nor particularly natural relative to the purposes at hand. Not natural, because none of them involves the *appearance* of possibility. Not mandatory, because there is an alternative notion, *philosophical* conceivability, that *does* involve this appearance and that sustains Hume's maxim

[16] For brevity, I'll speak simply of being *moved* to believe that *p*. (Why not define epistemic appearance in purely motivational terms? Because I do not want to say that *p* epistemically appears in cases where my motive for believing it is nonepistemic. Suppose I enjoy a representative appearance of someone offering to settle my debts if I will agree that *p*; this might tempt me to believe that *p*, but *p* does not epistemically appear to me.)

[17] *Objection*: Someone confronted with the Müller–Lyer diagram enjoys the representative appearance that the lines are unequal, but unless the diagram is completely new to her, she does not *believe* that they are unequal. *Reply*: What epistemically appears to a subject turns not on her beliefs but on what she is *moved* to believe. And why speak of a Müller–Lyer *illusion* if typical observers aren't *moved* to believe the lines unequal?

[18] Admittedly it is hard to draw a definite line between representative and (merely) epistemic appearances. Experts (matadors and mechanics) can enjoy representative appearances which to most of us are available only epistemically. But expertise is acquired gradually, and on the road to it there will be appearances not happily classified either way. For our purposes the indeterminacy doesn't matter; what will matter is the contrast between cases where *p* appears in both senses and those where it appears in neither.

against the objections. So the story has a negative part (sections III–IX) and also a positive one (sections X–XIV). At the end (section XV) I draw some tentative morals for the issue of realism vs. antirealism about modality.

## III. THE CONFUSION OBJECTION

Strangely prevalent in philosophy is the idea that to find a proposition conceivable is to find that it *is true for all you know*. Since Reid explained conceiving *p* as "giving some degree of assent to it, however small",[19] the idea has been repeated by many authors; to choose a source almost at random, William Kneale says or implies that to find *p* conceivable is to "have in mind no information which formally excludes" that *p* is true.[20] Ignoring minor differences of formulation, suppose we let the proposal be that *p* is conceivable iff it is *not unbelievable*, or for short *believable*.[21] (Remember that this is not to say that we see *p* as particularly *likely*, but just that we feel unable to rule it out.)

From an ordinary language perspective, the proposal is hard to argue with. Writing in the spring of 1990, Elizabeth Drew observed that German reunification had "become conceivable only in the last few months".[22] Anyone reading this would take it to mean, not that our powers of imagination had suddenly improved, but that reunification could no longer be regarded as out of the question. Likewise if I call it inconceivable that there is a largest prime number, but conceivable that there is a largest *twin* prime, I am saying that although it is certain that the primes are infinite in number, with the twin primes, things are not so clear.

Suppose I find *p* conceivable in the sense of believable. Does this give me reason to think that *p* is metaphysically possible? In other words, do I acquire evidence in *favor* of a proposition's possibility, by finding myself *without* evidence against its truth? That would be very strange, to say the least. Among other things it would have the result that there is a necessary limit on how bad my epistemological position can get: the *poorer* my evidence for *p*'s truth, the *better* my evidence for its possibility.[23] (In the limit of perfect ignorance about *p*'s truth, its possibility

[19] Reid 1969, essay IV, chapter III. This isn't Reid's preferred account. Usually he says that to "conceive a proposition . . . is no more than to *understand distinctly its meaning*" (*loc. cit.*). Since one can distinctly understand the meaning of a contradiction, this is an obvious nonstarter as an analysis of the kind of conceivability which purports to discover possibilities. (For early discussion of the "some degree of assent" theory, see Mill 1874, book II, chapter V, section 6, and Mill 1868, vol. I, chapter VI.)

[20] Kneale 1949, p. 213.

[21] Cf. Pap 1958, pp. 37–8, and van Cleve 1983, p. 37.

[22] *New Yorker*, March 19, 1990, p. 104. (At the time of writing reunification was far from a sure thing; to everyone's surprise it occurred just a few months later.)

[23] Compare Bradley's sarcastic remark that "merely because I do *not* find any relation between my idea and the Reality, I am to assert, upon this, that my idea is compatible". The epigraph is in a similar vein: "On the ground of my sheer ignorance . . ." (Bradley 1969, pp. 345–6).

would be absolutely assured!) Yet the fact is that I can be *completely* in the dark about truth and possibility simultaneously, as for example with the twin prime conjecture.

Apart though from the sheer oddity of arguing from ignorance to substantive modal conclusions, how reliable are such arguments? Already in Reid we find the only plausible answer:

> will it be said, that every proposition to which I can give any degree of assent is possible? This contradicts experience, and therefore [Hume's] maxim cannot be true in this
> FN:24 sense.[24]

Reid doesn't say what sort of "experience" he has in mind, but perhaps he was thinking of something he mentions later:

> Mathematics afford[s] many instances of impossibilities in the nature of things, which no man would have believed if they had not been strictly demonstrated [that is, their
> FN:25 impossibility would not have been believed if it had not been proved].[25]

So propositions to which people *once* gave "some degree of assent", say, the axioms of naive set theory, have often turned out *later* to be impossible. As an example of Kneale's shows, it is not always necessary to wait. Speaking of Goldbach's conjecture that every even number is obtainable as the sum of two primes, Kneale
FN:26 says that although it "looks like a theorem, . . . it may conceivably be false".[26] Likewise it may conceivably be true. But if true, it is necessarily true, and if false, necessarily false. Thus either the conjecture or its denial is a conceivable, that is to say a believable, impossibility. And the gimmick generalizes: we get a present-tense counterexample to the possibility of the believable *whenever* a proposition's truth-value is necessary but still unknown.

As a guide to possibility, then, conceivability *qua* believability is unreliable in the extreme. The fact that *p* might, for all I know, be true in the actual world, is just irrelevant to the issue whether it is true in some possible world or other. This leaves a puzzle, however: if the argument is as bad as that, why does there so much as *seem* to be an evidential connection? The answer is supposed to be that terms like "could" and "might" are *ambiguous*, which leads us into a certain confusion. Neglecting the distinction between what could be so in the sense that one is in no position to rule it out, and what could be so in the sense that it is metaphysically possible, we jump straight from the one to the other. According to the confusion objection, once this equivocation is noticed the appearance evaporates that conceivability argues for possibility.

● Q1 [24] Reid 1969, ~~T. and Hamilton, W. and Walker, J. (1855): *Essays on the Intellectual Powers of Man.* JC Derby~~ ● essay IV, chapter III.

[25] *Ibid.*

[26] Kneale 1949, p. 80.

## IV. BELIEVABILITY

Without a doubt, sliding from epistemic to metaphysical "could" is something we sometimes do, though we really should not. But, could a mix-up this elementary[27] really be *all there is* to the conceivability maxim? FN:27

caps ~~Probably~~ the *locus classicus* of the supposed confusion is Descartes's argument in the *Meditations* for the possibility of disembodied existence. Finding in the "First Meditation" that there might, for all he knows, be no material things, he suggests in the "Second" that he can exist without them. Isn't Descartes reasoning here that since he "could" in the believability sense exist without benefit of matter, he "could" do it in the metaphysical sense as well?

Part of the problem with such an interpretation is just that the attributed argument is so awful. But never mind that: if Descartes is attracted to this sort of argument, why does he not use it more often? At this point in the *Meditations*, remember, Descartes finds virtually *everything* believable, including for instance that he is essentially a body, and that God does not exist. Shouldn't he then conclude that these other things are possible as well? To answer that he doesn't conclude that they are possible, because he doesn't *believe* that they are possible, treats Descartes as rather more arbitrary than his position requires. Surely it would be better if we could make him out to mean something *other* than "believable" by "conceivable", such that he does *not* find it conceivable, in the sense *he* means, that he is essentially a body, or that God does not exist.[28] FN:28

[27] Among the many who have noticed it are Moore 1966, pp. 228 ff.; Sellars 1963, pp. 76 ff.; and Kripke 1980, p. 141.

[28] Consider in this connection Michael Hooker's challenge to Descartes's argument: Existence in the absence of bodies is no more conceivable than existence in the absence of persons not identical to bodies. On his own principles, then, Descartes *could* have been identical to a body. But whatever is possibly a body is a body essentially; so, although Descartes's actual position is that he can exist without bodies, he could equally have concluded that he is essentially a body (my précis of Hooker 1978, section II).—But why think that Descartes finds it conceivable that he should have been identical to a body? The only evidence Hooker offers is that "he does not know at this point in his inquiry that there are any disembodied minds", and that if "reflective consideration . . . leads one to doubt that *p*, then the truth of not-*p* is at least conceivable" (p. 181). However, this is just to say that (reflective) believability suffices for Cartesian conceivability, which is exactly what I deny. Hooker might counter that it is still mysterious why existing *as* a body should be any less conceivable than existing *without* bodies. Here is a suggestion: If all possible bodies are essentially bodies, and Descartes knows this, then to conceive himself identical to a body will be to imagine a world relative to which he is a body in *every* world. But how is Descartes to tell whether he can imagine a world like that without first attempting to imagine worlds in which he is *not* a body? Finding that he *can* imagine such worlds, Descartes is unable to conceive himself identical to a body. (Analogy: asked to think of a number such that all numbers are prime, you first consider whether you know of any *non*prime numbers. Realizing that you do, you find numbers of the first type unthinkable.)

Or take the example of our finding it conceivable, in the sense of believable, both that Goldbach's conjecture holds and also that it fails. If the inference from epistemic "could" to metaphysical "could" were so inviting, then it ought to seem strange that not a single author has concluded that although in some possible worlds, every even number is the sum of two primes, in others one or more of them *stops* being the sum of two primes.[29] Was it just that they knew that, in this case, such a conclusion would be counterintuitive? Again, a more sympathetic interpretation would be that conceivability, in the sense relevant to possibility, is a different thing from believability, and that neither Goldbach's conjecture nor its negation is conceivable in the relevant sense.

Earlier I agreed that "conceivable", as it occurs in daily conversation, usually *does* mean "believable". In fact more is true. As G. E. Moore noticed in an early paper,[30] not only "conceivable" but even "*possible*" normally indicates believability. Suppose, for example, that I tell you "it is possible that I was born on the moon". ~~Assuming~~ Given that I metaphysically *could* have been born on the moon, why does my statement sound so incredible? The reason is that "it is possible that *p*", where the embedded sentence is in the *indicative* mood, expresses uncertainty that *p* is false.[31] Thus "it is possible that I was born on the moon" says, not that this *could* have happened although it didn't, but that I am not entirely convinced I was born on Earth. (To assert genuine possibility, I must use the subjunctive mood: "it is possible that I should have been born on the moon.")

None of this is really very interesting except as a reminder that philosophers sometimes use words differently from other people. In metaphysics, for example, "possible" is often used for something *other* than believability, and this whether the subjunctive mood is used or not. Mightn't something similar be true of "conceivable"? The view I called strangely prevalent above is not that "conceivable" *ever* means believable, but that this is what it *always* means, including in conceivability arguments. For the truth is that in conceivability arguments, or at least competent ones, "conceivable" rarely if ever means believable.

There are two directions to this: conceivable propositions need not be believable, and believable propositions need not be conceivable. The easy direction is the first. An old Jewish saying runs: "Life is so full of misery and woe; how much better it would have been never to have existed at all; yet how many of us are that lucky?" Thinking about this, I find it conceivable that I should never have existed. Never for a moment, though, do I find it believable

[29] ~~Compare~~ Reid: "I have never found that any mathematician has attempted to prove a thing to be possible because it can be conceived . . ." (Reid 1969, essay IV, chapter III).

[30] "Certainty" (Moore 1966).

[31] To a first approximation, anyway. See DeRose 1991 for a more sophisticated treatment.

that I *have* never existed. So here is an example of a conceivable proposition that isn't believable.[32] Notice the underlying point: if conceivability entailed believability, then whenever one was *certain* that something was not the case, one would be unable to conceive it even as a possibility! This being absurd, the entailment does not go through.

Of believable propositions that aren't conceivable, it is difficult to give a pure example, if this means a believable proposition which is positively *in*conceivable.[33] After all, if $p$ is believable, then the *actual* world might for all I know be a $p$-world. So I am unlikely to have it appear to me that $p$ cannot be true in *any* possible world.

Perhaps there can be an impure example though. Sometimes when we find ourselves unable to conceive a proposition, we don't find it inconceivable either; its modal status is *undecidable* on the available evidence.[34] Despite what you often hear, this is how it is with Goldbach's conjecture. No thought experiment that I, at any rate, can perform gives me the representational appearance of the conjecture as possible *or* as impossible, or the slightest temptation to believe *anything* about its modal character. So this is already an example of a believable proposition that is not conceivable. But let me suggest some more interesting cases.

According to legend, the queen of Sheba tested Solomon's wisdom by challenging him to distinguish a flower from a wax facsimile thereof constructed in the royal workshop. As an aid to thought, suppose that she introduces these look-alikes to Solomon as Jacob and Esau—without, of course, telling him which is the artifact and which the flower. Then initially, before he determines, with the help of a bumblebee from the garden, that Jacob is the waxen artifact, Solomon finds it *believable* that Jacob should sprout new petals. Does he find this conceivable, though, in the sense relevant to possibility? Not if the stories about his wisdom are correct; he finds it undecidable on the available evidence. "If I assume that Jacob is a flower," Solomon might reflect, "then I *can* conceive it sprouting new petals; and if I assume that it is an artifact, then this becomes inconceivable for me. As it is, though, the petal hypothesis is neither conceivable nor inconceivable." Another story has Solomon ruling on a maternity case: is Mary, or Martha, the mother of this baby? Eventually he resolves the issue in Mary's favor, by offering to saw the baby in half. But initially, when Solomon found it believable both that Mary was the mother and that Martha was, did it appear to him that the baby's ancestry was metaphysically *contingent*? Only

[32] This gives, incidentally, another reason not to interpret Descartes as meaning "believable" by "conceivable". Probably there is nothing that Descartes finds more unbelievable than that he does not exist; yet for every created thing, Descartes finds it conceivable that it should not have existed. (Thanks to John Devlin for the next two sentences and the next note.)

[33] Although consider Tertullian: "Credo quia absurdum est."

[34] van Cleve 1983 distinguishes in a similar vein between *strong* and *weak* conceivability—"seeing" that $p$ is possible vs. not "seeing" that it is impossible—and he describes Goldbach's conjecture as only weakly conceivable.

please put the two subscripted 'b's back in (as on middle of p63) they are crucial!!!!!

if such an appearance were compulsory could one maintain that believability entailed conceivability.

b Two senses of "conceivable" have been distinguished: the believability sense (call it *conceivability*$_b$) and the philosopher's sense, the one that involves the appearance of possibility. Where the objector goes wrong is in failing to appreciate this distinction. Having uncovered a confusion about "could" in the

b argument from *conceivability*$_b$ to possibility, he falls into a confusion of his own when he offers this as a refutation of *conceivability* arguments.

this is correct as is, sans subscript

## V. SOME CIRCULARITY OBJECTIONS

Suppose that we are careful to keep believability and conceivability apart, and that we conclude to *p*'s possibility only when *p* is conceivable. Even this would be bad procedure, if it could be shown that

conceivability is a guide to possibility *only as constrained by prior modal information tantamount to the information that p is possible.*

This is roughly what the circularity objection alleges. Because the objection is easily misunderstood, let me consider some things it had better *not* be saying before working up to what I think it is saying.

Even the staunchest defender of Hume's maxim would not insist that the conceivable was *always* possible, or that *p*'s conceivability *proved* its possibility. Everyone is well aware of cases where impossible propositions have been found conceivable notwithstanding. The position to be defended, then, is only the following: that what is conceivable is typically possible, and that *p*'s conceivability justifies one in believing that possibly *p*.[35] Objection (A) does little more than reiterate these concessions in an accusing tone: FN:35

(A) Since your argument is by admission fallible, you yourself recognize that it *might* fail in any given case. Therefore you should refuse to draw the conclusion, until you get prior assurances that it won't fail in *this* case. And that means: prior assurances that *p* is possible. So the argument becomes circular.

What is unconvincing here is the move from "the conclusion might be false, compatibly with the truth of the premise" to "you should refuse to draw the conclusion until you're sure that it is *not* false". Arguments like this *usually* lead from truth to truth, so unless there is reason to think that truth is *not* preserved, it makes sense to suppose that it is.

[35] Further only *prima facie*, or *defeasible*, justification is claimed. Again, everyone knows of cases where additional evidence turns up that convinces us, or ought to, that *p* was not possible after all.

Do conceivability arguments have a deeper problem than ordinary fallibility? Maybe there is something special about their failures. If we think of an argument's premises as stating the *evidence* for its conclusion, it is an initially unsettling fact about conceivability arguments that when they fail, the evidence's very *existence* can be due to the conceiver's ignorance of the fact that her conclusion was false. So, Aristotle might not have been able to conceive matter as indefinitely divisible, if he had known that it could be divided only so far; "contingent identity" theorists like J. J. C. Smart might not have found mental and physical phenomena conceivable as distinct if they had realized that they were identical as a matter of necessity; and so on. For evidence to be in this sense *fragile* is hardly the usual thing. When Russell's chicken, for example, concludes from having been fed for months that he will be fed tomorrow, his evidence would *still* have existed even had he known his true fate. All the more striking, then, that when I conceive something in fact impossible, if I had appreciated its impossibility, ~~then~~ the misleading evidence might not have been:

(B) For all you know, you would not have found $p$ conceivable if you had been better informed, specifically, if you had known that $p$ was impossible. But evidence that might, for all you know, be dependent on ignorance is inherently untrustworthy. To be sure that your evidence is *not* thus dependent, you need to know that $p$ is possible. But then your argument becomes circular: you must already know that $p$ is possible, before you can conclude that it is from your ability to conceive it.

Now, it is a difficult question how fragile conceivability evidence really is. Whether foreknowledge of $p$'s impossibility would have prevented me from conceiving it seems to depend on how fully I grasp the *reasons* why $p$ is impossible, and how revealing those reasons are. But let's assume, for argument's sake, that *whenever* I find an impossibility conceivable, I would not have done so, had I but realized the proposition's impossibility. The problem is to see why this should reduce my confidence that *this* conceivable proposition is possible. After all, I draw the modal conclusion because I take it that given my evidence, it's probably true. And how is that probability affected, if I agree that in those occasional cases where my conclusion is false, my evidence would not have existed if I'd somehow fastened on the truth beforehand? Such a circumstance makes my errors more *embarrassing*, perhaps, but it doesn't seem to make them any more

FN:36 *common*.[36]

Some of the propositions I find conceivable are (I suppose) impossible, though of course I don't usually realize this in particular cases. Objection (B) tried to

rom

amount

caps caps [36] ~~Note that~~ a *certain* ~~degree of~~ fragility is only to be expected with arguments of the *it appears that p / therefore p* variety. ~~For instance,~~ the dishes displayed outside some Japanese restaurants stop *looking* like food when you are told that they're plastic models. ~~So~~ it is not just conceivability appearances that sit uneasily with a full and proper appreciation of their deceptiveness. caps find a problem in the fact that my not realizing it is a *necessary* condition of my finding them conceivable. Maybe this gets things backwards, however. Maybe the problem is that my ignorance of these propositions' impossibility would *sufficiently* explain my ability to conceive them:

(C) How can you infer to *p*'s possibility before you have ruled out alternative explanations of its conceivability? Since for *p* to be unbeknownst to you impossible would sufficiently account for your ability to conceive it, this is one of the alternative explanations you need to rule out. To rule it out, though, you need to know that *p* is possible, thus rendering the argument circular.

What is true in the objection is that when you base a claim on such-and-such evidence, the claim can be challenged by pointing to alternative explanations of the evidence which you are unable to exclude. They may have *looked* like ducks in the pond, but if there are known to be convincing decoy ducks about, you cannot *assume* that they were ducks unless you have something to say against the decoy hypothesis. There are limits, though. You are *not* required to rule out the alternative "explanation" that although they for some reason looked like ducks, in fact they were not, that is, that your evidence was somehow misleading. For one thing, this can hardly be considered an *explanation* of your evidence at all; for another, it is so far just allegation without the slightest reason to believe it. But how is objection (C) any better? The suggestion is that *perhaps* I had it appear to me that *p* was possible only because I somehow missed the fact that *p* was not possible. In short: perhaps my evidence is misleading. Perhaps it is, but don't I need a reason to think so before taking the idea seriously?

## VI. THE CIRCULARITY OBJECTION

Actually, the last two objections were bound to fail. For notice a feature they have in common: they propose accounts of such conceivability errors *as in fact occur* but without addressing the issue of whether their occurrence is at all to be *expected*. When you *do* conceive an impossibility, they say, a necessary and/or sufficient condition for this is that you did not realize that it was impossible. But this is compatible with your conceiving impossibilities rarely or never. To make the case that you conceive them often, the premise the objector needs is not that ignorance of impossibility is all it takes to *explain* a conceivability error, assuming it made, but that such ignorance is all it takes to *make* one. This stronger premise can be motivated by looking at a second alleged fallacy in Descartes's argument for dualism—this one rather more interesting than the last.[37]

FN:37

[37] See also Yablo 1990.

From his conceivability as existing without a body, Descartes concludes that disembodied existence is possible for him. The fallacy is said to lie in the fact that he simply takes it for granted that he has no essential properties beyond those that he knows of.

Objections like this were put to Descartes repeatedly, most notably by Arnauld in the "Fourth *Meditation*". Arnauld's view is that

> if the major premise of this syllogism [that the conceivability of *x* without *y* shows the possibility of *x* without *y*] is to be true, it must be taken to apply not to *any* kind of knowledge of a thing . . .; it must apply solely to knowledge which is adequate.[38]

By *adequate* knowledge of a thing, Arnauld means knowledge of all of its essential properties. Although what is possible for Descartes depends on his essence in its entirety, what he can conceive of himself is constrained by just that portion of his essence that he knows of. Unless his self-knowledge is adequate, then, his capacity for incorporeal existence might, for all the thought experiment tells him, be obstructed by unappreciated necessary connections. Here is Shoemaker in the same spirit:

> In the sense in which it is true that I can conceive myself existing in disembodied form, this comes to the fact that it is compatible with what I know about my essential nature . . . that I should exist in disembodied form. From this it does not follow that my essential nature is in fact such as to permit me to exist in disembodied form.[39]

What concerns me here is not the viability of Descartes's specific argument, or the truth of its conclusion, but the strategy which Arnauld's (Shoemaker's) objection represents. To be consistent, Arnauld should hold that *no de re* conceivability intuitions are trustworthy, unless the ideas employed are certifiable in advance as adequate—as embracing every essential property of their objects. But then an enormous part of our modal thinking falls under suspicion.

No one would doubt of herself that (e.g.) she could have been born on a different day than actually, or lived in different places; and outside of philosophy, no one would question that we know such things. But how do we know them, if not by attempting to conceive ourselves with the relevant characteristics and finding that this presents no difficulties?

What gives this question its force is the specter of an Arnauldian skeptic who holds that, given the possible inadequacy of my self-knowledge, I am in no position to oppose even such patently absurd essentialist hypotheses as that I am essentially born on September 30, 1957. If I might, unbeknownst to myself, be essentially accompanied by my body, however clearly I seem able to conceive myself without it, why couldn't I also be essentially born on that day, however clearly I seem able to conceive myself born a day earlier or later? Equally open to question are conceivability intuitions about objects *other* than oneself, like my

[38] CSM II, p. 140; my interpolation and emphasis.

[39] Shoemaker 1984, p. 155.

intuition that Humphrey could have been born on a different day or that the Eiffel Tower could be painted yellow; for here too the adequacy of my ideas has not been demonstrated. Really, the skeptic says, I have no basis to quarrel with *any* essentialist hypothesis about *any* object—even the superessentialist hypothesis that it could not have been different *in any way*—until I get assurances that none of the object's essential properties are hidden from me.[40]

At this point the restriction to *de re* propositions begins to seem artificial. If ignorance of an *individual*'s essential properties can generate modal error, why not ignorance of a *property*'s essential properties? Imagine that my grasp of a property *S* fails to reflect the fact that it is essentially uninstantiable (*S* might be the property of being sodium-free salt). Nothing to prevent me, then, from conceiving it as possible that *S*s should exist: a *de dicto* conceivability error rather than a *de re* one. Likewise the *de dicto* impossibility that some *Q*s are *R*s will be conceivable, if my understanding of *Q* omits its essential property of having no *R*s in its extension. Probably there is *no* proposition for which a worry like this cannot be raised. In skeptical moods, Arnauld will always be able to point to a potential gap in my modal information that would enable me to find *p* conceivable despite its impossibility. This suggests one final generalization of his objection to Descartes:

(D) If all it takes to find a proposition conceivable is to be unaware that it is impossible, then since impossibilities go unappreciated *all the time*, they are just as often conceivable. Before relying on conceivability evidence in any specific instance, then, you need a reason to think that in *this* case, *p*'s conceivability signifies that it is possible rather than that, although it is impossible, you are unaware of this. That is, you need a reason to *deny* that

(*) although you are unaware that *p* is impossible, *p* is impossible.

Because (*)'s first conjunct is true, and known to be—you *are* unaware that *p* is impossible—you can be reasonable in denying (*) only if you are in a position to deny its second conjunct. But its second conjunct is that *p*

[40] This brings out a seeming historical irony in Arnauld's position. Leibniz, in his correspondence with Arnauld, proposes that *none* of a thing's properties is accidental to it. Since Adam is such that Peter denied Christ some thousands of years after his death, this holds essentially of Adam, who would accordingly not have existed had Peter not gone on to be disloyal. Arnauld objects: "I find in myself the concept of an individual nature, since I find there the concept of myself. I have only to consult it, therefore, to know what is contained in this individual concept. . . . I can think that I shall or shall not take a particular journey, while remaining very much assured that neither one nor the other will prevent my being myself" (Mason 1967, pp. 32–3). Within limits, we share Arnauld's assurance, but it is hard to see what entitles *him* to it. How does he know that his self-conception is adequate, i.e., that he is aware of *all* of his essential properties? To complete the irony, something uncomfortably like this Arnauldian point is put to Arnauld by Leibniz himself: ". . . although it is easy to judge that the number of feet in the diameter is not contained in the concept of a sphere in general, it is not so easy to judge with certainty . . . whether the journey which I plan to take is contained in the concept of me, otherwise it would be as easy to be a prophet as to be a geometer . . ." (*op. cit.*, p. 59).

> is impossible! So you must already know that *p* is possible before you can conclude that it is from its conceivability.

(D) is the strongest form I know of the circularity objection; my only doubts are about its opening sentence. That conceivability arguments are *fallible* is of course admitted. But all the Humean need claim is that they are reliable *enough* for me to say: I'm justified, because *probably*, if my evidence holds, then so does my conclusion. Have conceivability arguments really been shown to be *so* fallible that this can no longer be said?

Without claiming to know exactly how fallible that is, I use the word "often" so that if impossibilities are *often* conceivable, then conceivability evidence is not *per se* justifying. Here is the opening lemma spelled out more fully:

(E1) Almost always, when I am unaware that *p* is impossible, I find it conceivable.
(E2) Often, when *p* is impossible, I am unaware that it is impossible.

---

(E3) Often, when *p* is impossible, I find it conceivable.

The first sign of trouble is that (E)'s logical form

(F1) Almost all Bs are Cs.
(F2) Many As are Bs.

---

(F3) Many As are Cs.

is deductively invalid. From the premises we know only that there is a high concentration of Cs among Bs, and a significant concentration of ~~Cs~~ Bs among As; what we don't know is whether these two concentrations *line up* to any significant extent. Thus it might be that although half of all As are Bs, only 1% of the Bs are As, and it is the *other* 99% of the Bs which make it the case that nearly all Bs are Cs. More generally, the Bs which are also As might form a small enough fraction of the total B-population as to be subsumable under the allowable exceptions to the general rule that almost all Bs are Cs. This is illustrated by argument (G):

(G1) Almost all swimmers are fish. (say, 95%)
(G2) Many mammals are swimmers. (say, 50%)

---

(G3) Many mammals are fish. (0%)

The conclusion is false because the *mammalian* swimmers—the ABs—are one and all exceptions to the generalization that swimmers are usually fish—that almost all Bs are Cs.

As a rough but workable guide to when this kind of trouble arises, an argument of form (F) is acceptable just in case premise (F1) can be rewritten as

(F1*) Almost all Bs, *whether they are As or not*, are Cs without loss of plausibility. Argument (G) is bad because, when we rework the first premise as indicated, we get something false:

(G1$^*$) Almost all swimmers, *whether mammals or not*, are fish.

Applying the rule to argument (E) yields

(E1$^*$) Almost whenever I am unaware that *p* is impossible, *whether it is impossible or not*, I find it conceivable.

The question, then, is whether unawareness of impossibility is *uniformly* conducive to conceivability—whether the relation holds *regardless* of *p*'s modal status.

Take first propositions such that I am unaware that they are impossible *and they are possible*. Surely I *do* find a great many of *these* conceivable, including almost every possibility I claim knowledge of: that I could have been taller, for example, or a better dancer, or born on a different day.[41] But the critical claim is that this generalizes to the *im*possible propositions:

(E1$^{\times}$) Almost always, when I am unaware that *p* is impossible, and it *is* impossible, I find it conceivable.

Because (E1$^{\times}$)'s antecedent says that I fail to appreciate the fact that *p* is impossible, this can be simplified to: unappreciated impossibilities are almost always conceivable.

Dialectically, at least, (E1$^{\times}$) is in a rather weak position. Remember that the objector is trying to convince someone not initially convinced of it that

(E3) Often, when *p* is impossible, I find it conceivable.

But anyone doubtful of (E3) will be doubly suspicious of (E1$^{\times}$), for understandable reasons. No one supposes that impossibilities *appreciated* as such are often conceivable; so to be doubtful that impossibilities are often conceivable is already to be doubtful that *unappreciated* impossibilities often are. And anyone doubtful that they are *often* conceivable will hardly be in a mood to concede (E1$^{\times}$)'s claim that they are *almost always* conceivable!

However, the problem is more than dialectical. The objector makes a statistical hypothesis: that almost whenever you fail to appreciate a proposition's impossibility, you find it conceivable. Normally such hypotheses are advanced on the strength of confirming instances. Why not now? Part of the reason might be that hardly any exist. At least, almost every unappreciated impossibility one *knows* of—Goldbach's conjecture (or its denial), Jacob's sprouting new petals, Martha's maternity, etc.—is not conceivable but undecidable. Rather than enumerating cases, though, I issue a challenge: if we are as prone as the objector suggests to

[41] Do I find conceivable *almost every* possibility such that I am not aware that it is impossible? Hardly—there are infinitely many unobvious arithmetical truths to the contrary—but let that pass.

conceiving unappreciated impossibilities, I would like to know what some of them are.[42]

## VII. BELIEVABILITY OF POSSIBILITY

Where does the objector get his confidence that unappreciated impossibilities are almost always conceivable? Perhaps for him this is not a statistical hypothesis at all, but a consequence of what he means by conceivability.

To see what his definition might be, look again at Arnauld's complaint against Descartes: "how does it follow, from the fact that he is aware of nothing else belonging to his essence, that nothing else does in fact belong to it?" What is striking here is Arnauld's assumption that Descartes *thinks* it follows. After all, Descartes's premise is not that he is *unaware* that he is essentially embodied, it is that he can *conceive* himself in a disembodied condition. That Arnauld puts the one premise for the other suggests that at some level, he takes them to say the same: a conceivable proposition is just one not known to be impossible. Shoemaker is more straightforward:

> in the sense in which it is true that I can conceive myself existing in disembodied form, this comes to the fact that it is compatible with what I know of my essential nature . . . that I should exist in disembodied form.

Apparently both authors equate conceivability, at least of the kind they find in Descartes, with what I will call *conceivability*$_{bp}$: the believability of *p is possible*.

Now, on *this* interpretation of conceivability, (E1$^{\times}$) looks awfully plausible. In fact it becomes something on the order of a conceptual truth: namely, that someone who doesn't realize that *p* is impossible will find its possibility believable. But if (E1$^{\times}$) is true on the new interpretation, then the critique of the last section no longer applies. What is my response to the circularity objection read in terms of conceivability$_{\text{bp}}$?

What response? I *share* the objector's doubts about conceivability$_{\text{bp}}$ arguments. In fact let me throw in some additional doubts of my own. To find a proposition conceivable$_{\text{bp}}$ is to find oneself unable to rule its possibility out. But you do not acquire justification for *believing* that something is possible simply through lacking justification for *denying* that it is. Otherwise, there could be no such thing as a person completely in the dark about *p*'s modal status; the less she knew *against p*'s possibility, the better her grounds would be for concluding that it *was* possible. (Recall that the argument from straight believability to possibility was criticized on similar grounds. If that argument was bad, the one from the believability of possibility is worse, for the new premise is strictly weaker than the old.)

[42] Bearing in mind that *not* to find a proposition *inconceivable* is not yet to find it conceivable.

So nothing as complicated as the circularity objection is needed to see that a proposition's possibility is not inferable from its conceivability$_{bp}$. But the objection's real problem is rather this: it *makes no difference* to Hume's maxim whether the inference goes through, for conceivability$_{bp}$ fails the modal appearance test on both counts. Thus suppose that I have *no idea* whether *p* is possible (*p* might be Goldbach's conjecture). Then I find *p* conceivable$_{bp}$—it is possible for all I know—but I have no inclination whatever to think it possible, nor have I misrepresented anything should it turn out not to be. In the end, then, the seemingly deeper circularity objection comes down to the same sort of misunderstanding as its predecessor: except that where the one mistook conceivability for the believability of truth, the other mistakes it for the believability of possibility.[43] FN:43

## VIII. THE *A POSTERIORITY* OBJECTION

Up to now we have been looking at *traditional* criticisms of Hume's maxim. But some may feel that the really decisive difficulty came to light only recently, with the discovery by Kripke and Putnam of *a posteriori* necessary truths: that cats are animals, that Hesperus is identical to Phosphorus, and so on.[44] FN:44 This would be strange if true, since for their own part these authors use conceivability methods all the time. But that is a separate issue; what is the problem that *a posteriori* necessary truths can *seem* to raise for the conceivability maxim?

Take any *a posteriori* necessity and negate it; the result is a necessary falsehood whose falsity is knowable only through experience, for instance, that cats aren't animals, or that water is distinct from $H_2O$. But, if it takes experience to show that these propositions are false, there ought to be *alternative* courses of experience that would have revealed them as true:

this should be in smaller type; it's a direct quote

we can perfectly well imagine having experiences that would convince us (and that would make it rational to believe) that water is not $H_2O$. In that sense, it is conceivable that water isn't $H_2O$.[45] FN:45

Putnam's conclusion is only that conceivability is no *proof* of possibility, but there is a more damaging result in prospect:

[43] This is not to say that Descartes's argument goes through. Perhaps Shoemaker is right that it is only in the believability-of-possibility sense that Descartes can conceive himself as disembodied. (Yet I assume that Descartes, for his part, would claim conceivability in a stronger sense; and so far we have no reason to doubt him.)

[44] See, for one, Teller 1984. By an *a posteriori* necessary truth I mean a necessarily true proposition whose truth is knowable *a posteriori* but not *a priori*; an *a posteriori* impossibility is the denial of an *a posteriori* necessary truth, in other words a metaphysical impossibility whose falsity is knowable only *a posteriori*.

[45] Putnam 1975, p. 233. ~~For discussion purposes,~~ I assume that water *is* necessarily $H_2O$.

(G1) Whenever *p* is *a posteriori* false, I find it conceivable whether it is possible or not.
(G2) Often, a posteriori falsehoods are impossible.

---

(G3) So *a posteriori* falsehoods are often found conceivable despite their impossibility.

This objection doesn't purport to embarrass *all* conceivability arguments, notice, only those where the conceived proposition is *a posteriori* false. But that is bad enough. For example, I should not argue from the conceivability of my sleeping late this morning, to the conclusion that this could really have happened. Even if it was *not* possible for me to sleep late, still I was going to find it conceivable that I should do just that.

## IX. EPISTEMIC POSSIBILITY

To conceive a proposition, in Putnam's sense, is to imagine acquiring evidence that justifies you in believing it: call this *conceivability$_{ijb}$*. But the definition is silent on a crucial point.

Distinguish three subtly different ways in which the thought experiment might go. Either the evidence is imagined to be disclosive of how things in the imagined situation really are; or it is imagined as for all its persuasiveness misleading; or whether the evidence is misleading is left unspecified. Speaking for myself, I can imagine being rationally persuaded of almost anything, provided I am allowed to imagine that the thing I am persuaded of is true, false, or of unspecified FN:46 truth-value, as I please.[46] To imagine a situation in which *p* is false, though, or one leaving *p*'s truth-value unspecified, is *not* a way of having it appear to me that *p* could have been *true*. So the only relevant case, the only one where I am in danger of conceiving an impossibility, is the one where I imagine myself believing *p* justifiably *and truly*. That understood, justification becomes a side issue. For if the belief is imagined as *true*, then whether it is imagined as justified or not, my evidence for *p*'s possibility would seem to be exactly the same. (How could the imaginability of my *knowing* that *p* be better evidence of possibility its than the imaginability of my truly believing it?)

Based on this reasoning, suppose we define *conceivability$_{itb}$* as the imaginability of veridically or truly believing that *p*. But, granted that this is different from conceivability$_{\text{ijb}}$, aren't *a posteriori* impossibilities *also* conceivable in the new

[46] Thus I can imagine some leading number theorist announcing an error in Euclid's proof from which it emerges that there is a largest prime number after all; the error takes years of training to understand, but the authorities are convinced, and I, naturally, defer to their superior knowledge. Although my *imagined* self is convinced, my *actual* self is not; I find a largest prime unimaginable, and so I suppose the imagined authorities to be mistaken.

sense? Can't I imagine truly believing that cats are robots, that Hesperus is distinct from Phosphorus, and so on?

Lurking just in the background here is a popular misunderstanding of Kripke's famous distinction between epistemic and metaphysical possibility. First it is emphasised that for Hesperus to have been other than Phosphorus is metaphysically impossible; *it could not have been* that Hesperus was not Phosphorus. Then it is explained that their nonidentity is nevertheless *epistemically* possible, since it *could have turned out* that they were not the same.

All of this is correct but the last step: the explanation of what epistemic possibility consists in. 'It could have turned out that *p*' claims, I assume, either the possibility, or the imaginability, of our coming to believe that *p* and believe it truly. On the *first* reading, as Kripke says, "*it could have turned out that p* entails that *p* could have been the case".[47] Since it could not have *been* the case that Hesperus and Phosphorus were distinct, they could not have *turned out* to be distinct. But, and this is the point, the explanation in terms of imaginability fares no better. To imagine myself truly believing that Hesperus and Phosphorus were distinct, I would have to imagine them *being* distinct; and that I cannot do, no more than I can imagine Venus's being distinct from Venus.[48]

Now it is a given that all of the usual *a posteriori* impossibilities[49] are to come out epistemically possible; this is the result for which Kripke introduced the notion. Since *not* all of these *a posteriori* impossibilities are conceivable$_{\text{itb}}$—*Hesperus* $\neq$ *Phosphorus* was our counterexample—conceivability$_{\text{itb}}$ cannot be what Kripke intends by "epistemic possibility". For much the same reason, though, conceivability$_{\text{itb}}$ is not a good reading either of "conceivable" as it occurs in the *a posteriority* objection. Unless we find *a posteriori* impossibilities "conceivable", the objection proceeds from a false premise; and to repeat, we do *not* seem to find them conceivable$_{\text{itb}}$.

Still it is hard to shake the feeling that there is *some* worthwhile sense in which we *can* imagine truly believing that Hesperus is not Phosphorus, that cats aren't animals, and so on. Since that might be the sense the *a posteriority* objection is looking for, let us consider the matter one more time. What is it to imagine yourself truly believing something? To believe truly is to believe a truth, so you imagine a situation in which you believe some true proposition. On reflection, though, it is not completely obvious how this proposition is to be identified. Is it the proposition that your *hypothetical* self entertains when it inwardly pronounces, say, 'water $\neq H_2O$', or the one that your *actual* self entertains? For these can be different.

[47] Kripke 1980, pp. 141–2.

[48] "But we *could* imagine veridically believing them to be distinct, back when we thought they *were* distinct." True but irrelevant; it remains that *Hesperus* $\neq$ *Phosphorus* is *now* epistemically possible, but *not* now conceivable$_{\text{itb}}$.

[49] Water $\neq H_2O$, gold is a compound, cats are robots, this lectem was originally made of ice, and so on.

Recall that the paper in which Putnam calls ‘water $\neq$ $H_2O$’ a conceivable impossibility contains in addition a story about how propositional content is fixed. Which proposition I believe, Putnam says, is a function not only of what goes on “in my head”—my narrow psychological state—but also of extrinsic contextual factors, including, for instance, facts about my causal interactions with the larger world. Thus the narrow psychological state, internal mental act, or what have you, constitutes only my subjective *contribution* to propositional content.[50] FN:50

How to fit beliefs themselves into the picture is a further question, and a disputed one. Some would individuate beliefs so that as long as the subjective contribution holds steady, the belief does too; variation in context affects not the belief *per se* but only the proposition believed. Others think of beliefs as having their propositional contents essentially: if I had believed a different proposition, then let my subjective condition be as similar as you like, I would have had a different belief. Rather than taking sides in this debate, suppose we concede the term “belief” to the second camp, and use “thought” to stand for the subjective contribution only. Thus my *thought* will be the internal state or act that determines, in context, which proposition I believe—what I will call the proposition *expressed* by the thought in that context. For instance, the thought which in the existing context expresses the proposition that Hesperus $\neq$ Phosphorus *would* have expressed a different proposition, a proposition with the truth conditions that Venus $\neq$ Mars, if Mars rather than Venus had been responsible for the appearances by which the referent of ‘Phosphorus’ is canonically identified.

All of which brings us back to the original question: in imagining, or seeming to imagine, myself truly believing an *a posteriori* impossibility *p*, do I imagine myself believing the proposition that my *p*-thought *actually* expresses? or believing *some other* proposition, the one that my *p*-thought *would* have expressed had the imagined situation obtained?[51] FN:51

Start with the first option: imagining myself believing the proposition that my *p*-thought *actually* expresses. Since the proposition actually expressed by my *p*-thought is the proposition that *p*, this is just conceivability$_{\text{itb}}$ again. What about the second option? Well, I can imagine believing *something* true with my *Hesperus $\neq$ Phosphorus*-thought, for as I said, I can imagine it expressing a proposition with the truth conditions that Venus $\neq$ Mars. Since I *cannot* imagine myself truly believing that Hesperus $\neq$ Phosphorus, we have uncovered a new

[50] See, for example, Dennett 1982 and White 1982.

[51] Depending on one’s theory of propositions, the same proposition *p* could be expressible, in the same world and context, by distinct thoughts *t* and $t'$ (so, the thought that the Morning Star $\neq$ the Evening Star might express the same proposition as the thought that Venus $\neq$ Venus). But then if someone thinks both *t* and $t'$ on a given occasion, the phrase “her *p*-thought on that occasion” will be ambiguous between *t* and $t'$. I will not bother about this problem except to say that it vanishes if we treat epistemic possibility as a property directly of thoughts.

kind of conceivability: *p* is *conceivable*$_{\text{ep}}$ if one can imagine, not truly believing that *p* (that very proposition!!), but believing *something* true with one's actual FN:52 *p*-thought.[52]

How does the *a posteriority* argument look in light of these distinctions, in particular its leading premise that all *a posteriori* falsehoods are conceivable? Read in terms of conceivability$_{\text{ijb}}$ or conceivability$_{\text{ep}}$, the premise is not unreasonable. For an *a posteriori* falsehood to be conceivable in *these* senses therefore says little for its possibility. Remember, though, that Hume's maxim claims evidential import only for the kind of conceivability that *portrays p* as possible. And the kinds just mentioned do not: the appearances they convey are rather that you could have been justified in believing that *p*, and that you could have believed some truth or other via the thought you actually use to believe that FN:53 *p*.[53] That leaves conceivability$_{\text{itb}}$. This *does* seem to involve the appearance of possibility, so Hume has some explaining to do if for all *a posteriori* falsehoods *p*, one can imagine truly believing that *p*. But this has not been argued, and as regards *a posteriori* impossibilities I doubt there are many who would even defend it. What we *can* do is imagine believing them justifiably, and believing related propositions truly; what we cannot do is imagine believing them, truly.

[52] The subscript "ep" is for *epistemic possibility*. Some will regard the analysis as too weak, others as too strong.

Too weak: "What I find epistemically possible ought to be constrained by my immediate evidential situation. For instance, if I know my visual field to be wholly red, then it should not be epistemically possible that it is wholly green. Yet this *is* conceivable$_{\text{ep}}$; I can imagine believing something true with the thought that my visual field is wholly green, for I can imagine its *being* wholly green." To accommodate this intuition we might try the following. Define a thought as *Cartesian* if it constitutes certain knowledge of the proposition it expresses, and it could not have expressed any other proposition; and let *c* be the conjunction of all propositions one thinks by way of Cartesian thoughts. Then *p* is conceivable$_{\text{epc}}$ if the conjunction of *p* with *c* is conceivable$_{\text{ep}}$.

Too strong: "Epistemic possibility ought to be a *weaker* notion than conceivability. Roughly it should be conceivability unconstrained by empirical beliefs. But some conceivable propositions are not conceivable$_{\text{ep}}$, for instance, the proposition that there are no thoughts." To accommodate *this* intuition, we need to arrange it so that thoughts continue to express propositions even in worlds where they do not exist. Say that the proposition a thought expresses in such a world is the one it expresses in the most natural expansion thereof to a world in which the thought *does* exist. Then *p* is conceivable$_{\text{epw}}$ if one can imagine a world that verifies the proposition that one's *p*-thought expresses therein.

Kripke offers no explicit definition of epistemic possibility, but his idea is that "under appropriate qualitatively identical evidential situations, an appropriate corresponding qualitative statement might have been [true]" (*op. cit.*, p. 142). This goes over into conceivability$_{\text{epc}}$ if by "qualitatively identical evidential situations" we understand situations satisfying the conjunction of all propositions one thinks by way of Cartesian thoughts; and by "corresponding qualitative statement" to *p* we understand a proposition *p** such that *p** is true at a world *w* iff one's *p*-thought expresses a truth there.

[53] And these things presumably *are* possible when *p* is *a posteriori* false.

## X. WHAT CONCEIVABILITY IS

Before attempting a positive account of conceivability, let me say something to lower expectations about what such an account should involve. Almost never in philosophy are we able to analyze an intentional notion outright, in genuinely independent terms: so that a novice could learn, say, what memory and perception *were* just by consulting their analyses. About all one can normally hope for is to locate the target phenomenon relative to salient alternatives, and to find the kind of internal structure in it that would explain some of its characteristic behavior. This at any rate is all I have hopes of doing for conceivability—and so much the better, in my view, if it can be done while remaining as neutral as possible on other issues. This section and the next propose an account that locates conceivability proper with respect to ~~the~~ various subscripted impostors; makes for a revealing contrast with inconceivability and undecidability; predicts that a conceived proposition will appear as possible; and does little else besides.

Here are the five main conceivability-notions that we have considered so far. Each should really be relativized to a person and an occasion, but we will be sloppy:

- p is *conceivable*$_b$ iff
  it is (not un)believable that *p*.
- *p* is *conceivable*$_{bp}$ iff
  it is (not un)believable that *possibly, p*
- *p* is *conceivable*$_{ijb}$ iff
  one can imagine justifiably believing that *p*.
- *p* is *conceivable*$_{itb}$ iff
  one can imagine believing *p* truly.
- *p* is *conceivable*$_{\mathrm{ep}}$ iff
  one can imagine believing something true with one's actual *p*-thought.

What I have been calling philosophical conceivability is none of these. Conceivability in the imaginability-of-true-belief sense comes closest, but has the following problem. I cannot imagine truly believing anything that conflicts with the hypothesis of my believing it: that I do not exist, for instance, or that no one has any beliefs. Yet many such propositions are philosophically conceivable, including the ones just mentioned.

From the way I have presented the problem you can guess its solution: I find *p* conceivable if I can imagine, not a situation *in* which I truly believe that *p*, but one *of* which I truly believe that *p*. This is the approach to be developed in what follows. And the obvious place to begin is with the nature of imagination.[54]

Imagining can be either *propositional*—imagining that there is a tiger behind the curtain—or *objectual*—imagining the tiger itself.[55] To be sure, in imagining the tiger, I imagine it as endowed with certain properties, such as sitting behind the curtain or preparing to leap; and I may also imagine *that* it has those properties. So objectual imagining has in some cases a propositional accompaniment. Still the two kinds of imagining are distinct, for only the second has alethic content—the kind that can be evaluated as true or false—and only the first has referential content—the kind that purports to depict an object.[56]

Objectual imagining, I said, may be accompanied by propositional imagining. But it is the other direction that interests me more: propositional imagining as accompanied by, and proceeding by way of, objectual imagining. To imagine that there is a tiger behind the curtain, for instance, I imagine a tiger, and I imagine it as behind the curtain. Quite possibly though I imagine the tiger as possessed of various additional properties—facing in roughly a certain direction, having roughly a certain color, and so on—and I imagine besides the tiger various other objects—the curtain, the window, the floor between them—all arranged so as to verify my imagined proposition. In short I imagine a more or less determinate *situation* which I take to be one in which my proposition holds. This is a closer approximation to what I mean by finding *p* conceivable; but "more or less determinate situation" is not quite right.

When I imagine a tiger, I imagine it as possessed of some determinate striping—what else?—but there need be no determinate striping such that I imagine my tiger as striped like *that*; the content of my imagining is satisfiable by *variously* striped tigers, but not by tigers of *no* determinate striping. Likewise for situations: even if there is much about my tiger-situation that I leave unspecified as irrelevant to the proposition at hand (e.g., the distance from the tiger's nose to the curtain), still I think of these things as fully definite in the situation itself. Thus a situation in which the tiger stands at *no* particular distance from the curtain, supposing that one can imagine this at all, is not what I have in mind.

[54] For a fuller discussion that supports on some points the approach taken here, see Walton 1990.

[55] Some philosophers use "imagine" so that imagining a thing is *imaging* it, that is, conjuring up an appropriate sensory presentation. I do *not* require a sensory-like image for imagining, and certainly not a distinct such image for distinct imaginings. (Compare Descartes on the unimaginability of chiliagons at CSM II, pp. 50, 69, 264.)

[56] "Can't the content of *objectual* imagining be truth-evaluable as well, if what one imagines is a proposition?"

This shows the importance of distinguishing the object of an imaginative act from its content. In the case described, the *object* of my imagining is a proposition. But its *content* is no more a proposition than the content of my tiger-imagining is a tiger. Rather it is something more on the order of a concept, the concept of being the proposition that a tiger behind the curtain is about to leap. Concepts being referential rather than truth-bearing, the criterion gives the right result.

By a *determinate* object, I mean one that possesses for each of its determinable properties an underlying determinate (it is not *merely* triangular, but in addition scalene, isosceles, or equilateral).[57] To imagine an object *as* determinate is to imagine it as possessing the higher-order property stated, that of possessing a determinate property for each of its determinables. There is a world of difference, then, between *imagining an object as determinate*—as possessing determinates for each of its determinables—and *determinately imagining it—specifying* in each case what the underlying determinate is. What I have been urging is that objectual imagining is determinate in the first sense but not the second. The one remaining question is whether the imagined object is itself indeterminate, as the phrase "more or less determinate situation" seems to suggest.

Suppose that it is, so that I imagine an indeterminate tiger rather than a determinate one. Then were a real, determinate, tiger to step out from behind the curtain, I ought to say that I had something more indeterminate in mind; whereas if an indeterminate tiger (!!) emerged, I ought to welcome it as just what I'd imagined. This of course get things exactly backwards. Do I imagine a determinate tiger, then? Not if this means that I am *en rapport* with one of all possible tigers, striped in one of all possible ways, etc. But to repeat a point already made, it is one thing to imagine an object as being of such-and-such a type, another for there to be an object of that type such that one imagines *it*. Understood on the first and more natural model, "I imagine a determinate tiger" describes the case perfectly.

Why should it be different, if the imagined object is a situation rather than a tiger? What we are tempted to describe as imagining a *more or less* determinate situation, is better described as imagining a *fully* determinate situation whose determinate properties are left more or less unspecified.

When I imagine a situation, I imagine a completely determinate one. Is this the same as imagining a possible world? Unfortunately not quite. Possible worlds are situations complete in *every* respect: spatially, temporally, and ontologically, for instance. But from determinacy alone these other dimensions of completeness do not follow. I may indeed imagine my tiger-situation as *part* of a complete situation, including, besides the tiger and its immediate neighbors, everything that coexists with them all laid out in some nameless pattern. But although this larger reality is in a sense *acknowledged*—I think of my tiger-situation as embedded in it—the point of calling it *larger* is that I do not imagine the whole of it in imagining the tiger-situation *per se*.

That I imagine my tiger-situation as limited is slightly awkward for our plan of explaining conceivability as the imaginability of a situation in which the

[57] Compare Locke's account of the "general idea of the triangle" as triangular but "neither oblique nor rectangular, neither equilateral, equicrural, nor scalenon" (Locke 1959, book IV, chapter 7, section 9). Lockean general ideas, if they existed, would be indeterminate in the sense intended; likewise "arbitrary objects" as discussed in Fine 1983.

conceived proposition is true. On the usual theory, propositions have truth-values not in limited situations, but in the complete situations I have identified FN:58 with possible worlds.[58] Luckily there is a way of correcting for this: As a rule, objectual imagining radically underdefines its object; so in principle it should be possible to imagine a *p*-verifying world while leaving matters visibly irrelevant to *p*'s truth-value unspecified. Granted that this is not itself to imagine a (limited) *p*-verifying situation, the two imaginings are closely related and it would seem natural for them to occur together. To look at the matter from the other direction, even if imagining my tiger-situation is not the *same* as imagining its larger world, I may well imagine the larger world *in addition*. This latter imagining is of course hopelessly unforthcoming about events outside the tiger's immediate neighborhood, but so it would be if its mission was to arrange for the truth of a proposition indifferent to those events; and so it should be, if it is to go proxy for imagining a situation in which those events have no part. I propose that the work that might have been done by the imagining of *situations* in our analysis can be done instead by the imagining of *worlds* understood mainly as containing those situations.

Now the pieces begin to fall together. Conceiving that *p* is a way of imagining that *p*; it is imagining that *p* by imagining a world of which *p* is held to be a true description. Thus *p* is conceivable for me if

FN:59 (CON) I can imagine a world that I take to verify *p*.[59]

Inconceivability is explained along similar lines:

FN:60 (INC) I cannot imagine any world that I don't take to falsify *p*.[60]

Obvious as this account may seem, it leads in interesting directions; and as it is, it fares better than any other account I know with the modal appearance test.

[58] Sense can also be made of truth-in-a-limited-situation, but it would be distracting to try to harmonize the two approaches here.

[59] It would be closer to ordinary language to distinguish 'I conceive that *p*', '*p* is conceivable for me', and 'I find *p* conceivable' as follows: (a) 'I conceive that *p*' iff I imagine a world which I take to verify *p*; (b) '*p* is conceivable for me' iff I *can* conceive that *p*; and (c) 'I find *p* conceivable' iff I find that I can conceive that *p*, presumably, by attempting to conceive it and finding that I succeed. But although my usage in this paper is *roughly* in accord with (a) and (b), to reduce clutter I have used 'I find *p* conceivable' and 'I conceive that *p*' more or less interchangeably. (Compare: 'I find it desirable/regrettable/acceptable . . . that *p*' is sometimes just a lengthier way of saying that I desire/regret●/accept . . . that *p*.)

this looks correct

● Q2

[60] *Objection*: Suppose *p* is the proposition that Socrates is a rain cloud. Then *p* is inconceivable to me; but I *can* imagine worlds that I don't take to falsify *p*, for I can imagine worlds in which Socrates doesn't exist. *Reply*: "Falsify" in (INC) is short for "fail to verify". For any world you can imagine, you take that world not to verify the proposition that Socrates is a rain cloud; hence you take it to "falsify" that proposition in the sense intended. To stress, on the intended reading (INC) is equivalent to the following: for every world I can imagine, I take that world not to verify *p*. (Brevity is not the only reason for using "falsify" rather than "fail to verify". The other reason is to discourage confusion with the much weaker condition that: for every world I can imagine, I do not take that world to verify *p*. This latter condition defines *non*conceivability.)

Tigers with round-square striping are not imaginable; neither can we imagine tigers that lick all and only tigers that do not lick themselves, or tigers with more salt in their stomachs than sodium chloride, or indeed any tigers that do not strike us as capable of existing. Assuming that this is no coincidence, two explanations suggest themselves:

(1) one cannot imagine an $X$ unless it *already* appears to one that an $X$ could exist; and
(2) to imagine an $X$ is *thereby* to enjoy the appearance that an $X$ could exist.

Which of these is more plausible? If (1) were correct, then we could never *arrive* at the view that $X$s are possible by succeeding in imagining one. Surely, though, this is the *usual* way of coming to regard $X$s as possible. For instance, it is only by learning how to imagine such things that we admit the possibility of, say, justified true beliefs that do not rise to the level of knowledge, or physical duplicates of ourselves that mean different things by their words. This shows that it cannot be a prerequisite of imagining an $X$ to be under the prior impression that $X$s can exist. Which leaves (2) as the likelier explanation: it *comes* to me that $X$s are possible in the act of imagining one.[61]

Assuming that objectual imagining works the way (2) says, it is no mystery why conceiving, in the sense of (CON), involves the appearance of possibility. By (2), when I imagine a world of such-and-such a type, it appears to me that a world of that type could really have existed. But when I take it to verify $p$, I take it that if a world like that *had* existed, then $p$ would have been the case. So, when I imagine a world which I take to verify $p$—and this is what it *is* to conceive that $p$ on the proposed account—I have it appear to me that $p$ is possible.

## XI. UNDECIDABILITY

Part of the appeal of (CON) and (INC) taken together is that they leave room for a *third* conceivability-status, such as undecidability was supposed to be. At least there is no obvious contradiction between

($\overline{\text{CON}}$) I cannot imagine a world that I take to verify $p$, and

($\overline{\text{INC}}$) I can imagine worlds that I don't take to falsify $p$;

and since these are the denials of (CON) and (INC), their conjunction defines undecidability. But although ($\overline{\text{CON}}$) and ($\overline{\text{INC}}$) are formally consistent, someone might still wonder how both could be true at the same time. For this would require that in attempting to conceive that $p$, I find myself imagining worlds

[61] Or, if this seems debatable, I hereby *stipulate* that "imagining an $X$" will denote type-(2) imagining.

such that it is *obscure* to me whether they verify *p* or falsify it. And do cases like this actually arise?

According to (CON), the task of conceiving *p* divides into two sub-tasks: imagining a possible world, and satisfying oneself that *p* is true in it. Often the world can be *stipulated* to be one in which *p* is true, as for example when Kripke stipulates that the man imagined to be President is our own Hubert Humphrey; then the verification task is trivial. But for some values of *p*, worlds in which *p* is clearly true are not clearly imaginable, or, what comes to the same thing, in clearly imaginable worlds *p*'s truth-value seems somehow uncertain. So, given his problems imagining a world in which Jacob sprouts new petals, Solomon may seek firmer ground in the hypothesis of a world where Jacob acquires petal-like appendages—whether these are petals is left obscure in deference to the possibility that Jacob is an artifact. Because he can imagine no world that he is ready to count as one in which Jacob sprouts new petals, the Jacob-proposition is not conceivable for him; but neither is it inconceivable, for he *can* imagine worlds which he is unready to describe as ones in which the proposition is false.

Another proposition I have called undecidable is not-GC, the denial of Goldbach's conjecture. Many philosophers have suggested that not-GC is rather conceivable. Michael Hooker, for instance, writes that one can

> imagine the discovery by computer of a counterexample to the conjecture, the attendant discussion of it, the subsequent revision of philosophical examples, etc.[62]

To explain where I think this goes wrong, let me describe some scenarios I clearly *can* imagine and then show how imagining *these* falls short of imagining that not-GC. For instance, I find it easy to imagine a computer printing out some unspecified even number *n*, and this being hailed on all sides as an authentic counterexample. Why wouldn't this be a case of imagining that not-GC? Because it suffices for the veridicality of *this* imagining for the following to be possible: GC has no counterexamples, but the computer produces a number *n* widely though erroneously hailed as a counterexample. Thus the truth of my imagining does *not* depend on there being a world in which not-GC, as it would if I had succeeded in imagining that not-GC.

Maybe I do better to imagine the computer producing something widely acknowledged as a *proof* that *n* is a counterexample. But again, the proof can help me to enjoy the appearance that possibly not-GC only if it is imagined to be *correct*; and since it is inconceivable to me that addition facts should vary between possible worlds, my ability to imagine the proof as correct is limited by my confidence that some number is *in fact* unavailable as the sum of two primes. Alas, I have no idea whether such a number exists, and neither (I assume) does anyone else. How then can I treat the computer's output as a correct proof? Am I to imagine it set out in convincing detail? But if the detail is only *imagined* to

[62] Hooker 1978, p. 178.

be convincing, it does nothing to increase my *actual* confidence in the proof's correctness. Am I to imagine the proof set out in *actually* convincing detail? If I could, I would call a press conference to announce my refutation of Goldbach's conjecture! So no Hooker-type thought experiment that I'm aware of shows the conceivability of not-GC. What the thought experiments *do* suggest is that not-GC is not *in*conceivable; accordingly it is undecidable.

## XII. MODAL ERROR

Ordinarily we treat perceptual appearances as *prima facie* accurate, and absent specific grounds for doubt we accept them as a basis for reasonable belief. What about conceivability appearances? Outside of philosophy, at least, they are treated in a similar fashion. Suppose that you claim to be able to imagine a world in which Oxford University exists but Cambridge does not. Perhaps we can point to some complicating factor of a kind you had not considered, e.g., one was originally a college of the other, that takes our own modal intuitions in a different direction. But if nothing of the kind occurs to us, and if attempting the thought experiment ourselves we find no difficulty in it, we are in a poor position to dispute your claim. (Imagine your reaction if we said, "still, we wonder if it is really possible", though no further complication suggested itself.)

So common sense sees appearances of both kinds as *prima facie* accurate and *prima facie* justifying. About conceivability appearances philosophers have taken a different view, but for unconvincing reasons. Can we stop worrying, then, and modalize with a clear conscience?

What makes us hesitate is not that conceiving can sometimes lead astray, but that we have so little idea how this happens. Modal error is a fact of life, and although perceptual error is too, our firmer grip on its etiology allows us to feel less the helpless victim than in the modal case. Misperception is something that we know how to guard against, detect when it occurs, and explain away as arising out of determinate cognitive lapses. That there is nothing remotely comparable for conceivability is a measure of our relative backwardness on the subject of modal error. Of course, the analogy with perception can be taken too far; a more realistic comparison might be with mathematics. Yet the system of checks and balances in mathematics is in its way most impressive of all and certainly well beyond anything encountered in the modal domain.

No wonder the advice to "trust your modal intuitions" sounds overeasy. Until our imaginative excesses are brought under something *like* the epistemological control we have in other areas, we modalize with right, perhaps, but without conviction.

Whatever their other problems, our objections at least had *models* to offer of how modal intuition goes wrong. Probably the most familiar is the one associated with the circularity objection: because you didn't appreciate *p*'s impossibility, there was nothing to prevent you from finding it conceivable. Even if this particular explanation disappoints, some *such* explanation is badly needed.

How *does* it happen that people find (what are in fact) impossible propositions conceivable? Maybe it looks like I've ruled modal error out altogether! Because what I've said is that when a proposition is unbeknownst to me impossible, it is not normally inconceivable for me but undecidable.—*Normally*, but not always. The ancient Greeks, believing that Hesperus and Phosphorus were different planets, might well have found it conceivable for the one to outlast the other. That was a mistake; Hesperus is identical to Phosphorus, so they could not have been different in any way. Or suppose that Oedipus, upset with Jocasta, finds himself imagining what life would have been like without her. Even if she had never existed, he decides, he could still have been king. Assuming with Kripke that ancestry is essential, he could not have been *anything* if she had never existed; so here is another example of modal intuition misfiring.

Sure as I might be, then, that my modal intuitions are *largely* reliable, in any particular case I have the following worry. Sometimes people have found impossibilities conceivable. Maybe I am making an analogous error when I imagine myself born on October 1, or six feet tall, or a Rosicrucian, and conclude that these things are possible for me.

## XIII. MODELS OF MODAL ERROR[63]

Is the analogy a good one, though? Remember that the ancients found it conceivable that Hesperus should outlast Phosphorus only because they took it that Hesperus and Phosphorus were distinct. What is the prior misapprehension that accounts for *my* erroneous intuition, as the ancients' denial of Hesperus's identity with Phosphorus accounts for theirs?

That the request for a backing misapprehension sounds so reasonable suggests the following model of modal error.[64] First, I find $p$ conceivable, when as a matter of fact it is impossible. Second, that $p$ is impossible emerges from the truth of some proposition $q$. Third, I do not realize this, believing instead that $q$ is false, or else that it is false that if $q$, then $p$ is impossible; and this is how I am able to conceive $p$ despite its impossibility. Explicitly, there is a proposition $q$ such that

(a) $q$;
(b) if $q$, then $\Box{\sim}p$; and

[63] This section and the next are based on Yablo 1990.

[64] Note: I do not say that *all* modal errors are captured by the models to be given here, only that many are, and especially the type most often discussed in recent modal metaphysics (see also note 67 below).

(c) that I find *p* conceivable is explained by my denial of (a) and/or my denial of (b).

('$\Box$ *s*' means: necessarily, *s*.) So, the ancients conceived it as possible for Hesperus to outlast Phosphorus because they denied the truth that Hesperus is identical to Phosphorus. If some contemporary philosophers, aware of this identity, find themselves capable of the same conception, the probable explanation is that they deny that identicals are modally indiscernible, and more particularly that Hesperus's identity with Phosphorus makes a difference in lifespan impossible. In our other example, Oedipus's false belief that Jocasta is not his mother explains how he can conceive himself being king even if she had never lived. Should he persist in his error after his ancestry is revealed, this is because he denies that *if* Jocasta is his mother, then he could not have been king without her.

Whatever you find conceivable, you are *prima facie* entitled to regard as metaphysically possible. The question is whether this *prima facie* entitlement can be defeated along the lines just indicated. Of course, if someone can *prove* that the model applies, then since (a) and (b) entail that *p* is impossible, your conclusion is refuted. But to raise legitimate *doubts* about the conclusion, reason to think that the model *may* apply ought to be enough. Thus we call proposition *q* a *defeater* if there is a reasonable chance that (a), (b), and (c).[65,66] The objector's challenge, in any particular case, is to find a defeater *q* of the conceiver's modal intuition.

Someone might object as follows. To erroneously conceive *p* as possible, why should I have to go so far as to *deny* the proposition *q* given which *p* is impossible, or to *deny* the proposition that *p* is impossible if *q* is true? Isn't it enough if I am simply *unaware* that *q*, or *unaware* that if *q* is true, then *p* is impossible? Thus consider a second, less demanding, model of modal error: there is a proposition *q* such that

(a) *q*;
(b) if *q* then $\Box\sim p$; and
(c) that I find *p* conceivable is explained by my unawareness that (a), and/or by my unawareness that (b).

Arguably this unawareness model does do a certain justice to cases which the denial model leaves untouched. At one time, for example, I suppose I found it

[65] Although it would be more in accord with existing usage to let the defeater be the *conjunction* of (a), (b), and (c). See note 67.

[66] How do I test the credibility of the conditional claim (b) that if *q*, then *p* is impossible? With any other indicative conditional, I use the Ramsey test: I pretend that I am reliably informed of the antecedent, and then I consider, under that pretense, how plausible I find the consequent. The same method works here. Suppose I want to decide whether, if salt = sodium, it is impossible for the ocean to contain more sodium than salt. Pretending that salt = sodium, I find it inconceivable that the ocean should contain these in different amounts; abandoning the pretense, I endorse the conditional.

conceivable that there should be a town whose resident barber shaved all and only the town's non-self-shavers. However, it was not because I *denied* that the scenario was implicitly contradictory that I found the town conceivable; it was because I was not *aware* of the contradiction. Or imagine that the medievals, rather than denying that dolphins were mammals, had no opinion on the matter; suppose if you like that the concept of a mammal was unknown to them. Mightn't they have conceived it as possible, erroneously mind you, for dolphins to be fish? If so, then this would be another example of a false intuition whose explanation lay not in the fact that something was denied, but in the fact that it was not believed.

As before, the objector's challenge is to identify a proposition *q* for which there is a reasonable chance that the model applies.[67] Nothing could be easier, you might think. Just let *q* be the proposition that *p*, the proposition conceived, is impossible. Then since the conceiver's intuition is still *sub judice*, there would seem to be a reasonable chance that (a) *q*, that (b) if *q*, then *p* is impossible (this is a tautology), and that (c) the conceiver's ignorance of (a) explains how she managed to conceive *p* as possible.

Yet I take it that it gives me *no* reason to mistrust my intuition that *p* is possible to be told that it might, for all I know, be due to ignorance of what might, for all I know, be the fact that *p* is *not* possible; for instance, that my ability to conceive myself with a different birthday might derive from my failure to appreciate the necessity of my actual birthday. At best the objector can argue that *if* I am necessarily born on September 30, then my failure to realize this may be relevant to my finding a later birthday conceivable. And this hardly constitutes an *objection*, no more than it is an objection to the accuracy of my impression that there are ducks around that if I am wrong, and they are decoys, then my ignorance of that fact might help to explain how I managed to take them for ducks.

Part of my point here is just that ignorance of the fact that *p* is impossible does not itself do much to explain why I would conceive it as possible. But that is not all. Even if a fuller explanation is provided, it carries little dialectical force if it depends on the prior concession that my intuition has a significant chance

[67] This is a good place to acknowledge that the models given here cannot claim to accommodate *all* defeaters. Suppose we distinguish *rebutting* defeaters, propositions *s* such that (con(*p*) & *s*) is a reason to think that *p* is impossible; *offsetting* defeaters, propositions *s* such that (con(*p*) & *s*) is not a reason to think *p* possible; and *undermining* defeaters, propositions *s* such that *s* is a reason to deny that con(*p*) is a reason to think *p* possible. And suppose we refer to conjunctions of (a), (b), and (c) as *standard* defeaters. Then standard defeaters are rebutting and offsetting (in virtue of (a) and (b)) and also undermining (in virtue of all three conjuncts). But none of our three categories is exhausted by the standard defeaters. For instance, intuition recognizes offsetting and undermining defeaters that are not rebutting. *Some* such are obtainable by generalizing the models to allow standard defeaters of con($p^*$), where $p^*$ is a fuller description than *p* of the imagined world as the conceiver understands it. But even this leaves no room for defeaters like the following: you conceived that *p* while under the influence of a mind-expanding drug; your modal intuitions are famously inaccurate; everyone but you finds *p* undecidable.

of being false. (With equal plausibility one could explain away my perceptual impression of ducks by saying that they were produced by decoy ducks, these being the usual explanation of erroneous duck-impressions.) Only if there is *independent* reason to suspect that my refusal of some relevant proposition really does put me out of touch with the facts, does that refusal call my intuition into question.

## XIV. MODAL DIALOGUE

To see how this works in practice, consider again my Cartesian intuition that I can exist in a purely mental condition. Someone might object that it is independently plausible that I am embodied, and that if so, I am embodied necessarily and so incapable of purely mental existence. About the second half of this, I have my doubts. Like most people, I take it for granted that I *am* embodied. Somehow, though, this does not seem to inhibit me from conceiving myself as disembodied. This intuition of being actually-but-not-necessarily embodied *prima facie* rationalizes my *rejection* of the conditional hypothesis stated; so I cannot regard that hypothesis as independently credible. Of course, the conditional hypothesis becomes virtually certain if we let *q* be the proposition that I am *necessarily* embodied. Now, though, it is *q* itself which wants for independent evidence.

Another candidate for the role of defeater is that I am the *same thing* as my body. But what does "same thing" mean here? If it means identical, then I doubt that the defeater is independently plausible. However *categorically* similar my body and I may be, this suggests at most that we are *coincident* (as a statue might be coincident with the hunk of clay that makes it up).[68] Evidence that we were moreover identical would be evidence that we agreed on a wide range of hypothetical, and especially *modal*, properties. Yet this can only come from conceivability considerations, which seem in fact to argue the other way! If "same thing" is understood so as to require sharing of categorical properties only, then the problem is just relocated. For now I need a reason to think that if I am categorically similar to my body, then I cannot exist without it. And to insist that categorical similarity has this consequence seems to beg the question against the otherwise intuitive view that what I am is a *person*, whose categorical properties may be those of a certain body, but with modal properties all my own.

Obviously the debate could be taken a lot further. To mention just two of the more promising possibilities, someone might try to extract a defeater from

[68] On coincidence and the categorical/hypothetical distinction, see Yablo 1987. ~~I assume for the sake of the objection that there are no temporal differences between myself and my body—for instance, that my body isn't going to outlast me.~~

Kripke's claim that my biological origins are essential to me, or from some version of the mental/physical supervenience thesis. But already we have enough to see how modal dialogue typically proceeds on the picture I have in mind:[69]

- *X* finds *p* conceivable and calls it possible;
- if *Y* chooses to challenge *X*'s intuition, she proposes a defeater *q* to explain how *X* was capable of it despite its falsity;
- if *X* is unable to accept this explanation, he takes issue either with *q* itself, or with *Y*'s claim that it casts doubt on his intuition's accuracy.

What to say—what it means—when the dialogue breaks down is the topic of the next section.

## XV. FACTUALISM ABOUT MODALITY

To defeat a modal intuition, the objector tries to motivate on independent grounds the suspicion that it derives from some prior error or oversight. Yet if conceivers disagree on fundamental enough matters—color incompatibilities, say, or the modal properties of mathematical objects—it may be difficult for either to discern on the other's part a prior lapse at all, still less one independently recognizable as such. This raises the specter of brute modal error and disagreement. Too much of *that*, someone might say, and we lose the right to speak of error and disagreement at all.

Supporting this accusation is a theory of what it is for the statements in a given region of discourse to be genuinely factual, viz. that "differences of opinion about such statements . . . will have to be traceable back to some breach of ideal rationality or material difference in the subjects' respective states of information".[70] Reason to think that there is *just no saying* how the opposition comes by its seemingly *equally* well-supported conclusions despite their falsity is "reason to think that the statements disagreed about are not objective, and so not apt to be substantially true or false".[71]

Roughly, then, the proposal is to define factual discourse by its intolerance of brute error and disagreement. There are stronger and weaker versions of this, of course, and much that could be debated in all of them, but it is hard not to feel some sympathy for the basic idea. Unless the positions one would like to call incorrect show *some* tendency to be reproachable on separate grounds, the faith that there is anything genuinely at issue can indeed become strained. The

[69] For reasons explained in note 67, the framework cannot be regarded as fully general. For instance, it doesn't cover the case where *Y* challenges *X*'s intuition on the basis that he was drugged, or that he has often been wrong before. But I believe that it covers *most* modal disputes of the kind that arise between basically competent conceivers.

[70] Wright 1986, p. 198. This is what Wright used to call the "rational command" criterion and now calls "cognitive command".

[71] Wright 1988, p. 39.

alternative is to insist on there being "facts of the matter" that only oneself and one's coreligionists are privy to—that others, through no fault of their own, get consistently wrong. And although facts like that may not be unintelligible, they do have something of a credibility problem. This is especially so when, as in the modal case, our best idea of the type of fact in question is that of an external constraint on the outcome of a certain type of investigation: in the modal case, investigation by imagination. For then our confidence that there are facts of that type in play will be limited by our confidence that an external constraint really operates; hence by our resources for explaining how, despite the constraint, we are able to arrive at opposing views.

So, our entitlement to modal factualism turns on the effectiveness of our strategies against conflicts, or seeming conflicts, of conceivability intuition.[72] (Here and below I use "conceivability intuition" broadly, as covering conceivability and inconceivability intuitions both.) What are those strategies? From the discussion above we have the following:

(1) try to show that there is no conflict of *conceivability* intuitions because what looked like *p*'s conceivability was really only its believability, or epistemic possibility, or . . .; or what looked like its inconceivability was really only its unbelievability, or epistemic impossibility, or . . .;

(2) admit that there are conceivability intuitions on either side but try to show that they are not in *conflict* because what seemed to be the conceivability (inconceivability) of one proposition was really that of some closely related other;

(3) admit that there is a conflict of conceivability intuitions but try to show that at least one of them has a defeater and is therefore open to doubt.[73]

(1) was the strategy we used with Goldbach's conjecture, when we said that it was "conceivable" only in the believability or the believability-of-possibility sense. The supposed intuition that Hesperus might not have been Phosphorus can be met with (1)—you find their nonidentity not conceivable but epistemically possible—or, what comes to the same in this case, (2)—it is not their nonidentity that you find conceivable, but only that you should have thought *something* true

[72] At least, a certain *degree* of factualism might be in order if the condition were met. In his 1988 and elsewhere, Wright sketches a system of increasingly ambitious factualisms, and offers criteria appropriate to each. Here I employ a variant of his weakest criterion. Whether modal discourse is factual in his more ambitious senses I do not discuss; Wright himself is skeptical.

[73] To apply this strategy on the conceivability side of the conflict, we use the (a) (b) (c) model as presented in the text; to apply it on the inconceivability side, we extend the (a) (b) (c) model to inconceivability intuitions in the obvious way. Suppose that historians discover that Cicero was in reality Tully's older brother (that *q*), but that unaware of this I continue to find it inconceivable that the one should have outlived the other (that *p*). My intuition is defeated because (a) *q* is true; (b) if *q* is true, then *p* is possible; and (c) I find *p* inconceivable only because I am under the misimpression that *q* is false.

FN:74 with your *Hesperus* $\neq$ *Phosphorus*-thought.[74] Another, more mundane, version of strategy (2) is to say that because of unnoticed idiolectic differences, the disputants talk past each other. Thus if we seem to disagree on the conceivability of a wet mop that holds no water, a possible explanation is that owing to differences in our concepts of wetness, the proposition I find inconceivable is not the one you find conceivable. (Sadly it is all too easy to believe that much of the current controversy over conditions of personal identity and survival—are teletransportation, brain transplant, mitotic division, etc. survivable?—owes more to our meaning slightly different things by "person" and "survive" than to any real clash of modal intuition.)

When the dissolving strategies fail, our one remaining option is to explain the conflict as arising out of some antecedent error or omission on one side or the other. To the newly crowned Oedipus, it seemed possible that he should have been king even if Jocasta had never existed; but what would you expect of someone deceived about his ancestry? The reason why some can conceive a barber who shaves all and only the non-self-shavers, while others find this inconceivable, is that the first group needs to learn more logic. And so on.

But I have been putting off the essential question: what if, after all the strategies have been tried to the best of current knowledge and ability, there remains a residue of so-far-irreducible disagreement? Well, the factualist can say, there is still such a thing as *committing* ourselves to applying them in ever more inventive ways until one finally succeeds, or, failing that, to devising new and better strategies in a similar spirit. Such a commitment could of course come to seem awfully lame, if the failures proved stubborn and the successes too minor to balance them off. But there is another scenario I like better.

How is it that substantive modal metaphysics, after years in the doldrums, has lately been making headway again? Part of the explanation might be that our methods of modal conflict management have been in a real sense improving. Already it takes an effort to recall the dispiriting conditions of, say, thirty years ago: the various half-related ideas jumbled unconsciously together under the headings of possibility and conceivability; how crude the controls were on propositional content; the anxiety about collateral information as a factor in imaginability. Especially one forgets how much easier it was then for the conversation to bog down at the first clash of modal intuition. The extent to which we have moved beyond this should not be exaggerated (more often than not we *still* bog down), but meanwhile it seems that modal dialectic has achieved an unaccustomed degree of clarity and system in a surprisingly short time. All of this has been a tremendous boost to the factualist's morale; sufficiently more of it and her commitment above might well be vindicated. "But what is the

[74] In the case of the ancients, who really *did* find it conceivable that Hesperus should have been distinct from Phosphorus, strategy (3) is used: they were capable of this conception only because they were empirically and/or philosophically misinformed.

verdict? Can modal metaphysics be brought under the discipline characteristic of a fact-finding enterprise or can't it?" I have no answer, but just a suggestion: we should try to impose that discipline in the hope that it might eventually take.

## REFERENCES

Bealer, G. (1987). "The Limits of Scientific Essentialism". *Philosophical Perspectives* 1.

Blackburn, S. (1993). "Morals and Modals", In Essays in Quasi-Realism pp. 52–74, ital
Oxford: Oxford University Press.

Bradley, F. H. (1969). *Appearance and Reality*. Oxford: Oxford University Press ~~pp. 119–141~~.
● Q3

Cleve, J. van (1983). "Conceivability and the Cartesian Argument for Dualism". *Pacific Philosophical Quarterly* 64, pp. 35–45.

Coppock, P. (1984). "Review of N. Salmon, *Reference and Essence*". *Journal of Philosophy* 81, pp. 261–70.

Cottingham, J., Stoothoff, R., and Murdoch, D., eds. (1985). *The Philosophical Writings of Descartes* (I, II). Cambridge: Cambridge University Press (= CSM).

Dennett, D. (1982). "Beyond Belief". In A. Woodfield (ed.), *Thought and Content*, Oxford: Clarendon Press. pp. 1–95.

DeRose, K. (1991). "Epistemic Possibilities". *Philosophical Review* 100, pp. 581–605.

Dreyfus, H. (ed.) (1982). *Husserl, Intentionality and Cognitive Science*. Cambridge, Mass.: Bradford.

Fine, K. (1983). "A Defense of Arbitrary Objects". *Proceedings of the Aristotelian Society*, supp. vol. 17.

Forbes, G. (1985). *The Metaphysics of Modality*. Oxford: Clarendon Press. pp. 55–77.

● Q4 Hooker, M. (1978). "Descartes's Denial of Mind–Body Identity". of Hooker, M., ed., pp.171-85 ~~pp. 171–185~~ *Descartes: Critical and Interpretive Essays* (Johns Hopkins, Baltimore, ~~1978, p/b~~)

Hume, D. (1963). *An Enquiry Concerning Human Understanding*. La Salle, Ill.: Open Court Press.

____ (1968). *Treatise of Human Nature*. Oxford: Clarendon Press.

Kneale, W. (1949). *Probability and Induction*. Oxford: Clarendon Press.

Kripke, S. (1980). *Naming and Necessity*. Cambridge, Mass.: Harvard University Press.

Locke, J. (1959). *An Essay Concerning Human Understanding*. New York: Dover.

Mason, H. T. (ed.) (1967). *The Leibniz–Arnauld Correspondence*. Manchester: Manchester University Press.

Mill, J. S. (1868). *An Examination of Sir William Hamilton's Philosophy*. Boston: W. V. Spenser.

____ (1874). *A System of Logic*. New York: Harper & Brothers.

Moore, G. E. (1966). "Certainty". In his *Philosophical Papers*, New York: Macmillan. pp.227-51

Pap, A. (1958). *Semantics and Necessary Truth*. New Haven: Yale University Press. EXTRA# ~~pp. 227–251.~~

Putnam, H. (1975). "The Meaning of 'Meaning' ". In his *Mind, Language, and Reality*, Cambridge: Cambridge University Press. pp.215-71

Putnam, H. (1990). "Is Water Necessarily $H_2O$?". In his *Realism with a Human Face*, Cambridge, Mass.: Harvard University Press, pp 54–79.

Reid, T. (1969). *Essays on the Intellectual Powers of Man*. Cambridge, Mass.: MIT Press.

Searle, J. (1983). *Intentionality*. Cambridge: Cambridge University Press.

Sellars, W. (1963). "Phenomenalism". In his *Science, Perception, and Reality*, London: Routledge & Kegan Paul.pp.60-105

Shoemaker, S. (1984). "Immortality and Dualism". In his *Identity, Cause, and Mind*, New York: Cambridge University Press, ~~pp. 60–105~~. pp.139-58

Sidelle, A. (1989). *Necessity, Essence, and Individuation*. Ithaca, NY: Cornell University Press, ~~pp. 139–58~~.

Teller, P. (1984). "A Poor Man's Guide to Supervenience". *Southern Journal of Philosophy*, supp. vol. 22. pp.137-62

Walton, K. (1990). *Mimesis as Make-Believe*. Cambridge, Mass.: Harvard University Press, ~~137–62.~~

White, S. (1982). "Partial Character and the Language of Thought". *Pacific Philosophical Quarterly* 63, pp. 347–65.

Wright, C. (1986). "Inventing Logical Necessity". In J. Butterfield, ed., *Language, Mind and Logic*, Cambridge: Cambridge University Press, pp. 187–209.

____ (1988). "Realism, Antirealism, Irrealism, Quasi-realism". *Midwest Studies in Philosophy*, xii, pp. 25–49.

Yablo, S. (1987). "Identity, Essence, and Indiscernibility". *Journal of Philosophy* 84, pp. 293–314.

____ (1990). "The Real Distinction Between Mind and Body". *Canadian Journal of Philosophy*, supp. vol. 16: 149–201; Ch. 1 above.

● Q5 ●

● Q6 ____ 1992 ~~(forthcoming)~~. "Review of Sidelle, *Necessity, Essence and Individuation*". *Philosophical Review*, 878–881.●

## Queries in Chapter 2

Q1. Author edit not clear, please confirm fullstop or comma.

Q2. Author edit not clear.

Q3. Page numbers written by author is not clear. Please check and confirm whether we have carried out as per author requirement and update.

Q4. Author edit not clear.

Q5. Check the 3rd reference marked by author here.

Q6. Author edit not clear.

# 3

# Textbook Kripkeanism and the Open Texture of Concepts

> [one imagines producing] an exhaustive list of all the circumstances in which the term is to be used so that nothing is left to doubt . . . construct[ing] a *complete definition*, i.e., a thought model which anticipates and settles once for all every possible question of usage . . . in fact, we can never eliminate the possibility of some unforeseen factor emerging . . . and thus the process of defining and refining an idea will go on without ever reaching a final stage.
>
> F. Waismann (1965)

## 1. INTRODUCTION

A lot of people appear to have drawn the same "good news–bad news" lesson from their reading of Saul Kripke on conceivability. The bad news is that conceivability evidence, particularly of a "conceptual" or "a priori" sort, is highly *fallible*. Very often one finds a statement *E* conceivable when, as a matter of fact, *E*-worlds cannot exist. So it is, for instance, with the conceivability of water in the absence of hydrogen, or of Hesperus without Phosphorus.

The good news is that (although conceivability evidence is fallible) the failures always take a certain form. A thinker who (mistakenly) conceives *E* as possible is correctly registering the possibility of *something*, and mistaking the possibility of that for the possibility of *E*. There are *illusions* of possibility, if you like, but no outright delusions or hallucinations.

The good news is important because it gives a way of living with the bad. That a statement *E* is conceivable may not itself be proof that *E* is possible; but proof is what it becomes in the absence of an $E^*$ such that it was really $E^*$ that was possible, and $E^*$ whose possibility was misread as the possibility of *E*.

I am grateful to George Bealer, Paul Boghossian, Tyler Burge, David Bzdak, David Chalmers, Mark Crimmins, Eric Funkhouser, Tamar Gendler, Sally Haslanger, Robin Jeshion, Eric Lormand, John ~~O'Leary,~~ Hawthorne, and Christopher Peacocke for comments and advice.

Now, what is the relation between $E$ and $E^*$ whereby the one's possibility is so easily misread as the possibility of the other?

The quick answer is that $E^*$ maps out the way the proposition that $E$ is presented in thought; it is, for short, a *presentation* of $E$. The usual sort of presentation takes proper names in $E$ and replaces them with descriptive and/or demonstrative phrases that, as Kripke says, *fixes their reference*; so, "water" might be replaced by "the predominant clear local drinkable stuff".

But the essential point is that $E^*$ delivers the propositional content of $E$ as a function of the circumstances that obtain where $E$ is uttered. What $E$ actually says, assuming the actual world is $w$, is the same as what $E^*$ says about $w$, i.e., what it says considered as a description of $w$.[1] Suppose for instance that $E$ is "water is plentiful". Then what $E$ actually says, pretending that the actual world is a $w$ whose watery appearances are appearances of XYZ, is what $E^*$ = "the clear drinkable stuff is plentiful" says about $w$, viz. that *its* clear drinkable stuff is plentiful, viz. that XYZ is plentiful.

Now, it comes as no surprise that the possibility of a presentation of $E$ should be confused with the possibility of $E$. A world of which $E$'s presentation is true is a world such that, *had it really obtained,* $E$ would have expressed a truth. But an understandable confusion is a confusion nevertheless. The possibility of "water is plentiful" expressing a truth is one thing—it's the possibility of there being lots of watery stuff—the possible truth of what it does express is another—it's the possibility of there being lots of $H_2O$.

Two notions of possibility, then. Our job as philosophers is (i) to clearly distinguish the two notions, and (ii) to explain how they are related. The first part is easy:

(i) an $E$ that could have expressed a true proposition is "conceptually possible", while an $E$ that does express a proposition that could have been true is "metaphysically possible".

The second part is not too difficult either. By (i), $E$ is conceptually possible iff it expresses a truth in some $w$-considered-as-actual. By definition of "presentation", the truth $E$ expresses in $w$-considered-as-actual corresponds to a true description its presentation $E^*$ gives of $w$-considered-as-counterfactual. By (i) again, for $E^*$ to be true of a counterfactual world is for $E^*$ to be metaphysically possible. Hence

(ii) $E$ is conceptually possible iff $E^*$ is metaphysically possible.

And now comes the philosophical payoff. From (i) we see why it is so often a mistake to infer a statement's metaphysical possibility from its conceivability. Conceivability (particularly of a conceptual or a priori ~~sort~~ character) tracks in the first instance *conceptual* possibility, not the metaphysical sort. It appears from (ii), though, that the inference is *not* a mistake when no obfuscating presentation can

[1] Better: the same as what $E^*$ says about $w$ on a "referential" reading.

be found, that is, when there is nothing to play the role of $E^*$ but $E$ itself. In that case, (ii) tells us that $E$ is possible in the one sense if and only if it is possible in the other.

## 2. TEXTBOOK KRIPKEANISM

The story just told can be called *Textbook Kripkeanism* about conceivability and possibility. How well it corresponds to any actual belief of Kripke's is hard to say, and something I take no stand on. What I do think is that Textbook Kripkeanism is not right. The "good news" that $E$'s conceivability ensures its possibility whenever no obfuscating presentation suggests itself is too good to be true.

About sixty years ago, the philosopher Charles Hartshorne put a neat twist on Anselm's ontological argument for God's existence.[2] Granted, he said, that existence is part of God's essence does not *itself* show that God exists; it implies only that if God *were* to exist in some world, then he would exist necessarily. God in other words is either necessary or impossible. But, God is not impossible, since we can easily conceive him. Hence God is necessary, and so actual.[3]

A response that was given even at the time is that Hartshorne is punning on "possible". All God's conceivability establishes is his *conceptual* possibility. The premise needed to establish his necessity, however, is that *he really could have existed*. Only if there is a possible world that really contains him can we say: he exists in $w$, so his essence is satisfied in $w$, so he has the property of necessary existence in $w$, so he exists in every possible world, this one included.

All of this is very familiar. The reason for mentioning it is that assuming Textbook Kripkeanism, it fails to block the argument. Let it be that God's conceivability establishes only that he is conceptually possible. Still, the gap here is not very large. A statement's conceivability suffices for its metaphysical possibility *except* in those cases where all we have cottoned on to is an $E^*$-world passing itself off as $E$.

The question is: can we find a presentation of $E =$ "there is a being whose essence includes existence" such that it is really only this *presentation* that is possible, not the proposition that it presents? The presentation would replace name-like expressions in $E$ with nonrigid descriptive phrases spelling out how we identify their referents in thought.

But, and this is putting it mildly, it is hard to think what the reference-fixing descriptions could be, or what they would replace; the statement "there is a being

[2] Hartshorne (1941). The relevant bits are reprinted in Plantinga (1965).

[3] "If 'God' stands for something conceivable, it stands for something actual" and "The necessary being, if it is not nothing, and therefore the object of no possible positive idea, is actual" (Plantinga 1965, p. 135).

whose essence includes existence" seems *already* to be about as conceptually articulate as one could want. Another way to put it is that it is hard to see what the *genuine* possibility is that we mistake for the possibility of an essentially existent being. Without a separate possibility "in the neighborhood" to point to as what was confusing us, it seems we have to conclude that it is $E =$ "there is a being whose essence includes existence" that is possible. And now it follows that a being like that truly exists.

In case anyone is not alarmed by the story so far, let me stretch it out a little. Another thing that seems clearly conceivable is that there should *fail* to be a being whose essence includes existence; it seems conceivable, in fact, that there shouldn't be anything whatsoever. Once again, it is hard to think of a presentation of "there isn't *anything*" such that it is really this presentation that is possible, and this presentation whose possibility is mistaken for the possibility of emptiness.

Now we have talked ourselves into a contradiction. Textbook Kripkeanism has the result that (Hartshorne's) God exists in some worlds but not in others. But it is a conceptual truth about this God that he exists in every world or none. The same problem arises for other "modally extreme" entities: numbers, pure sets, transcendent universals, and so on. Given Textbook Kripkeanism, they are not merely recherché, they are paradoxical. Nor can the paradox be evaded by saying that numbers and sets do not exist; it flows from the very concepts involved.[4]

## 3. CONSCIOUSNESS

If Textbook Kripkeanism could be seen at work only here, in connection with God and other modally extreme entities, it might not be worth making a fuss about. But it plays a role too in an increasingly popular objection to physicalism pressed by Frank Jackson and David Chalmers.[5]

Any physicalism worthy of the name says that the world's mental aspects are necessitated by what goes on here physically. But there is at least one sort of mental phenomenon—*consciousness*—that we can conceive going *missing* in a world that is physically just like ours. In a word, *zombie* worlds are conceivable. Doesn't this run directly against the physicalist's necessitation claim? Not according to most people. All that follows from the conceivability of zombie worlds is that they are conceptually possible; it would take their metaphysical possibility to bother the physicalist.

[4] Someone might say that in a contest between the intuition of possible existence and the intuition of possible nonexistence, the intuition of nonexistence should win out. If that's right, then the contradiction becomes a proof that God, numbers, universals, and so on do not exist.

[5] Jackson (1994); Chalmers (1996). References to Chalmers and Jackson are always to these two works. Jackson's *From Metaphysics to Ethics: A Defense of Conceptual Analysis* (Oxford: Oxford University Press, 1998) had not ~~yet~~ appeared when this article was written.

All of this is again very old news. The effect of Textbook Kripkeanism, however, is to call it into question. Space between conceptual and metaphysical possibility can open up only under fairly special conditions. And, it will be said, these conditions aren't met in the present case. Zombie worlds had better be *conceptually* impossible, then, if physicalism is to have a chance.

Now, as it happens, Jackson and Chalmers have slightly different reasons for thinking that the zombie scenario is one where the conceptual/metaphysical distinction finds no foothold. The crucial point for Jackson is that we are considering a world stipulated to be physically *just like ours*. He thinks he can get the physicalist to admit that when physical premises a posteriori necessitate nonphysical conclusions, *additional* physical premises can be found to make the necessitation a priori. Since in the zombie scenario we are allowed *complete* physical information, the additional physical premises have "already been added." So physical premises conceptually necessitate consciousness if they necessitate it at all. What makes the zombie scenario special for Chalmers is less the nature of the (physical) premises than that of the (phenomenal) conclusion. Like Kripke, he is impressed by the fact that the way the proposition that *I am in pain* is presented in thought is scarcely to be distinguished from the proposition itself. To put it in terms of presentations, $E^* =$ "I am in a state that hurts" is *necessarily* equivalent to $E =$ "I am in pain".[6] And if statements are true in the same possible worlds,[7] then there is little prospect of explaining away the apparent possibility of one as the genuine possibility of the other.

## 4. JACKSON AGAINST THE PHYSICALISTS

The Textbook Kripkeanism of Chalmers' strategy is plain to see. How Jackson fits in will take a little explaining. His essential claim, remember, is that if pain is necessitated a posteriori by physical premises, then an expanded set of physical premises necessitates pain a priori.

The argument for this begins with a puzzle. At first we are inclined to think of understanding as *knowledge of truth conditions*: for our purposes, knowledge of which worlds a sentence truly describes. If that is the correct theory, though, then understanding a necessarily true sentence $E$ should suffice for *appreciating* its necessity. And it clearly does not. I can understand "where there is $H_2O$, there is water" without having any idea of its true modal status.

[6] This is not an absolute assumption for Chalmers; see below where he tries to get by on weaker premisses.

[7] The relevant statements are not the ones in the text exactly but built on these: "things are physically like so and I am not in pain" and "things are physically like so and I am not in a state that hurts".

But the reason for my oversight is no great mystery, says Jackson. The reason is that I am under- or misinformed about what chemical substance is (in the present context) picked out by the reference-fixer of "water"; I am aware only in a potential or hypothetical sense of the truth conditions that *E* in fact possesses. That this does not prevent me from understanding *E* suggests that understanding is a matter not of knowing the conditions under which *E* is true, exactly, but

> knowing how the conditions under which it is true depend on context, on how things are outside the head. (p. 39)

A little more explicitly, it is knowing the *meaning function* $E_m$ mapping contexts in which *E* might be uttered to its truth conditions in those contexts. Since one can grasp this meaning function without knowing *E*'s actual truth conditions, simply through ignorance of which context actually obtains, the puzzle dissolves. One can't be expected to see *E*'s necessity if one doesn't know its truth conditions.

● Q1 ●Notice what this implies, however. If it is ignorance of context that enables me to miss *E*'s truth conditions, then once this ignorance is remedied, I am out of excuses. Semantic competence in other words should enable me

> to move a priori from . . . statements about the distribution of $H_2O$ *combined with the right context-giving statements*, to information about the distribution of water. (p. 39)

This takes Jackson close to his desired conclusion that whatever is metaphysically necessitated by the full physical story is conceptually necessitated by it. But a detail has been left hanging.

Why should the context-giving information be *physical* information? Couldn't the reference-fixer for "water" mention, say, the fact that it is supposed to be something *clear* and *tasteless*? Of course it could. But remember, Jackson says, we are asking after the consequences and commitments of *physicalism*. And the physicalist of all people is in no position to doubt that context is ultimately to be described in physical terms. Assuming physicalism, then, whatever is necessitated by physics is conceptually necessitated by it. This applies in particular ~~with~~ to psychology:

> the physicalist is committed to there being an a priori story to tell about how the physical way things are makes true the psychological way things are. [Note,] the story may come in two parts. It may be that one part of the story says which physical way things are, $P_1$, makes some psychological statement true, and the other part of the story, the part that tells the context, says which different physical way things are, $P_2$, makes it the case that it is $P_1$ that makes the psychological statement true. What will be a priori accessible is that $P_1$ and $P_2$ together make the psychological statement true. (p. 40)

Obviously, though, there are various psychological statements that are *not* a priori necessitated by physical ones, such as the statement that there is conscious experience. So, they are not necessitated by physical statements at all,; so physicalism is false. That completes the argument.

## 5. THE LINK WITH TEXTBOOK KRIPKEANISM

The puzzle that Jackson uses to disprove physicalism is really just the puzzle of a posteriori or nonconceptual necessity. Why isn't all necessity the *conceptual* kind? It can equally well be stated in terms of the "dual" notion of conceptual possibility, where $E$ is conceptually possible if, roughly, it is not a priori that not-$E$.[8] How can $E$ be conceptually possible without being *really* possible?

Textbook Kripkeanism has a view about this combination of features.[9] The one and only way for $E$ to be conceptually possible but not "really"—metaphysically—possible is for something *else* to be really possible, namely $E$'s presentation $E^*$. This presentation being an a priori equivalent of $E$ that specifies what $E$ says as a function of worldly context, the claim is that uttered in the right context, $E$ would have expressed a truth.

But this is very close to what Jackson tells us. According to him, the reason we don't see that not-$E$ is impossible is that the meaning function $E_{\mathrm{m}}$ telling us what proposition $E$ expresses in a given worldly context occasionally yields the result that it expresses a true proposition.[10] Thinking of the Textbook Kripkean's $E^*$ as an attempted linguistic expression of Jackson's meaning function $E_{\mathrm{m}}$, the two stories basically agree.

## 6. KNOWING WHICH

So, then: Jackson's argument is an example of Textbook Kripkeanism. The connection here is suggestive in both directions. Having seen earlier that Textbook Kripkeanism overgenerates modal "truths", e.g., it yields the contingency of theism, the suspicion is that Jackson's strategy may overgenerate as well. Having *not* seen earlier where the Textbook Kripkean goes wrong, it becomes tempting to look for signs in the Jackson argument of what might be misleading Textbook Kripkeans more generally. Our basic question, remember, is: how can an impossibility go unnoticed except under color of a suitable presentation, or now, meaning function?

[8] I am finessing something here. Jackson's puzzle is about conceptual necessity in a particular sense. $E$ is conceptually necessary iff understanding $E$ reveals it as necessary. The cognate notion of conceptual possibility is: understanding $E$ leaves it open that $E$ might be possible. The notion in the text is weaker; there, $E$ is conceptually possible iff understanding $E$ leaves it open that $E$ might be *true*. A fuller treatment would distinguish conceptual truth from conceptual necessity. But the overall argument would not be affected.

[9] I'm taking it that to be conceptually possible and to be conceptually conceivable are about the same.

[10] He would say "a possible proposition". But he shouldn't. The conceptual possibility intuition is compromised if $E_{\mathrm{m}}(w)$ is nonempty but never true in $w$ itself, as with "not all horses are actual horses".

Start with the matter of why the "contextual information" needed to boot an a posteriori necessity up into a conceptual one should be *physical* information. Jackson says that the physicalist of all people is in no position to deny that context is physical. But there has to be more to it than that. The physicality of context is one thing, the physicality of *information* about context—the information speakers need to parlay their understanding of $E$ into knowledge of its truth conditions—is another.

So let's ask again: why should physicalists think that the contextual information is physical? They are not *deniers* of nonphysical information, after all. They merely insist that it be necessitated by physical information.[11] If the necessitation were *conceptual*, then no problem; information that is *conceptually* necessitated by physical information can be considered itself physical.[12] But to insist that the necessitation is conceptual would seem to beg the question at issue.

Or maybe not. Suppose that a physical description $P$ of the context necessitates a nonphysical description $Q$. ($P$ and $Q$ might be "$H_2O$ plays such and such a role" and "$H_2O$ is water".) Then the conditional "if $P$ then $Q$" threatens to be the very sort of necessary truth that Jackson says he finds puzzling. Why isn't it *conceptually* necessary? The only possible answer is that it has necessary truth conditions in this context, nonnecessary ones the next context over.

This is reintroducing a complication we had thought to be done with. Given that $P$ and $Q$ were brought in to pin down the context of $E$ enough to settle *its* truth conditions, it seems only fair to allow that they do not bring with them *further* context-sensitivities. And now the thinker has no excuses; "if $P$ then $Q$" has got to be conceptually necessary, in which case the physicalist may as well concede that context-giving information $Q$ is indeed physical.

Notice the underlying assumption: the puzzle about nonconceptual necessities is *such* an extremely puzzling puzzle that it's *not allowed to even exist* except when Jackson's preferred strategy of solution is available. *Anyone who really and truly knows which worlds "if P then Q" is true at has got to realize that it is true at all worlds*. I want to flag that assumption because it's going to come up again. How does the argument fare from this point on?

Understanding $E =$ "there is pain" is knowing how its truth conditions vary with context. The physicalist is allowing that it takes only physical information to know which context one is in, nearly enough at least to be able to compute $E$'s truth conditions. So, someone who understands "there is pain" and possesses the relevant physical information knows which worlds are $E$-worlds. But (and let's flag this assumption too) *anyone who really knows which worlds are E-worlds thereby knows whether the E-worlds include all worlds physically just like this one.*

[11] I mean information about *this* world. Physicalism being a contingent thesis, there may be nonphysical information about other worlds that fails to be necessitated by physical information about those worlds.

[12] Alternatively, we could plug the necessitating physical information in for the contextual information, and let the contextual information be a priori deduced.

Putting the pieces together, anyone who really understands “there is pain” is in a position to parlay purely physical information about context into the knowledge that zombie worlds are impossible.

Both stages of the argument depend on hypotheses about what “else” ought to be known by someone who knows which worlds a statement truly describes. And indeed, the puzzle itself depends on such a hypothesis; knowing which worlds a *necessary* statement is true of is supposed to suffice for knowing that it is true of *every* world. Here is the general schema:

(+) knowing which worlds are $E$-worlds suffices for knowing that the $E$-worlds are (include, etc.) the $F$-worlds, assuming they in fact are.

This seems like asking a lot. For one thing, I may not have a very good idea of which worlds are $F$. Take for instance the *worlds that are physically just like this one*. Unless I know which worlds these are—and given how little I know about the physical nature of *this* world it seems an open question—knowing which worlds contain pain is clearly *not* going to tell me whether the pain-worlds include them. Or let the $F$-worlds be the class of *all* possible worlds bar none. If I am uncertain about which worlds are really possible (and I am), then there is nothing to prevent me knowing which worlds physically just like ours contain pain while still failing to know whether *all* worlds fall into this category.

But the real reason (+) doesn't work is one that applies even when we know which worlds are $F$. The real reason is that the standards for “knowing which” are *themselves* so intentional and context-driven as to prevent any easy conclusions about what the knower is now in a position to appreciate.[13] FN:13

This much seems plausible: for me to know which worlds make $E$ true, I need a way of picking out the $E$-worlds in thought, and not any old way will do. But the sort of way that suffices is not a function of the set of worlds alone. It depends (among other things) on its being the sentence $E$ that is used to designate the set as opposed to some necessarily equivalent alternative. I know which worlds $E =$ “there is pain” is true of by knowing that they are the worlds in which there is pain. (If more than that is required, I don't understand “there is pain”.) I know which worlds $F =$ “things are physically as in *our* world” describes by knowing that they are the worlds in which matters are physically as in our world—and here I might be able to reel off some specific physical requirements. Obviously though to know in *these* sorts of ways which worlds $E$ and $F$ are true of does not put me in a position to tell whether $E$ is true in every $F$-world, even if in fact it is.[14] FN:14

[13] Whether or not the space of worlds can serve as a final all-purpose matrix for commensurating meanings, the idea of using *grasp* of world-sets to explain *grasp* of meaning doesn't seem to get us much further ahead.

[14] Jackson brushes up against this issue without noticing its application to his own case:

> Suppose I hear someone say “He has a beard.” I will understand what is being said without necessarily knowing the conditions under which what is said is true, because I may not know who

shouldn't quoted material such as this be in smaller type?

## 7. CANONICAL CONCEPTION

One line of response would be to equate understanding with some sort of *unmediated*, perhaps acquaintance-like, grasp of which worlds make your sentence true; that will be postponed for a bit, until after Chalmers. Another is to insist that understanding a sentence is a matter of knowing which set of worlds it expresses in a *special canonical way*: a way that better responds to what worlds in their innermost nature are.

Some such adjustment might seem called for anyway, since otherwise the equation of understanding with knowledge of truth conditions flirts with triviality. No doubt understanding "France is a democracy" goes with knowing that the worlds it is true of are the ones where France is a democracy. But this sort of explication doesn't seem to take us very far. It would be better (one might think) if the verifying worlds could be identified not as whatever makes it the case that *E*, but, well, as the worlds they are.

Now, since the physicalist thinks that worlds are in their innermost nature physical, he will presumably insist on a *physical* specification. But then it can't be that the speaker "misses" the fact that any world physically like ours is a pain-world simply through failing to think of the pain-worlds in physical terms. Thinking of them in physical terms is a condition of understanding, and we are talking about a speaker who understands.

The claim is that, *if physicalism is true*, then to understand *E* one must be able to decide (i) on the basis of *physical* information (ii) how to make the cut between *E*- and non-*E*-world in *physical* terms. (If physicalism is true, then understanding is "physical" understanding.) This plugs the gap in Jackson's argument, and his conclusion is now reinstated. Whatever physical premises necessitate at all, an expanded set of physical premises conceptually necessitates. Merely to understand the sentences is to appreciate their truth relations.

is being spoken of . . . [Nevertheless] I am much better placed than the Russian speaker [because] I know how to move from the appropriate contextual information, the information which in this case determines who is being spoken of, to the truth-conditions of what is said. (p. 38)

In a footnote he worries that perhaps

I do know who is being spoken of: I know something unique about [that] person, namely, that he is being spoken of and is designated by a certain utterance of the pronoun "he." But this is "Cambridge" knowing who.

Is it, though? It seems to me that it may or may not be, depending on circumstances. Suppose a radio at the very same moment intones the words "he has a beer". It is hard to tell which words come from which source, and hence whether the speaker is talking about the one being described as bearded or the one being described as beered. To realize that the speaker (as opposed to the radio) is talking about the referent of $H_1$ (the utterance of "he" that goes with "has a beard") as opposed to $H_2$ (the one going with "has a beer") might be enough in this context for knowledge of who the speaker, as opposed to the radio, is talking about.

shouldn't
quoted material
be in smaller
type?

shouldn't quoted
material such as
this be in smaller
type?

Quite right, but so what? The intuition the physicalist has got to be careful not to flout is that a *normal* understanding of "things are physically like so" and of "there is pain" ~~should~~ leave open the possibility of zombie worlds. That a ∧s *physical* understanding of the same sentences should leave this possibility open is not intuitive at all. On the contrary: a physical understanding of "there is pain" is by definition an ability to tell whether worlds presented in physical terms do or do not contain pain. The only physicalist who should be bothered by the refurbished argument is the one (if he exists) who thinks ordinary understanding is *physical* understanding as defined by (i) and (ii). And that sort of physicalist deserves to be in trouble.

Everything here goes back to the assumption that the physicalist will insist on a physical specification of the verifying worlds. Why should he? Physicalism was supposed to be an ontological theory, not a theory of understanding. This distinction is trampled on when understanding is equated with canonical grasp of truth conditions. It now becomes a "consequence" of physicalism that typical speakers, to the extent that they find zombie worlds conceivable, don't really understand "there is pain"! The physicalist presumably finds this as bizarre as anyone else. Why should one's claim to understand "there is pain" depend on such an arcane and out of the way matter as the possibility of zombie worlds?[15]

## 8. CHALMERS AGAINST THE PHYSICALISTS

A word first about Chalmers' semantical framework. He and Jackson agree in associating with $E$ (as employed in a particular context) a propositional content made up of the worlds which $E$ (as used in that context) truly describes; this content is in Jackson's terms the "truth conditions" of $E$, in Chalmers' terms $E$'s "secondary intension". They agree, too, in assigning $E$ an *additional* semantical value intended to bring out how $E$'s interpretation varies with context.

The difference is that where Jackson's "additional" value is a meaning function from contexts to propositions (sets of worlds), Chalmers' "primary intension" is just another proposition.[16] A world gets into $E$'s secondary intension if $E$ is true *of* that world considered as counterfactual, and into $E$'s primary intension if $E$ is true *in* it considered as actual. For short,

$|E|_1 =$ the set of *E-verifying* worlds, the ones making $E$ true.

$|E|_2 =$ the set of *E-satisfying* worlds, or just $E$-worlds.

Both of these intensions can be seen as arrived at compositionally from the intensions of $E$'s component terms. The reason that "water $=$ H$_2$O" has a

[15] You can let it depend if you like. But don't blame the results on the physicalist; it wasn't he who told you to make understanding such a counterintuitive thing.

[16] Save for a complication about "centered" worlds, which we'll get to later.

necessary secondary intension and a contingent primary one is that "water" and "$H_2O$" agree in secondary intension only. With "water = the watery stuff", it's the other way around; the primary intension is necessary, because "water" and "the watery stuff" co-refer in all worlds-considered-as-actual, but the secondary intension is not, because a counterfactual stuff (Putnam's XYZ) describable as "the watery stuff" may not be describable as "water".

To calibrate the three accounts: *E*'s primary intension $|E|_1$ = the set of *w* belonging to $E_m(w)$ = the set of worlds in which *E* expresses a true proposition. (Some will recognize this as Stalnaker's "diagonal proposition".[17]) Its secondary intension $|E|_2 = E_m(@)$, the set of worlds falling into the proposition that *E* actually expresses. The connection with Kripke is that $|E|_1$ is the set of $E^*$-worlds, while $|E|_2$ is the set of *E*-worlds. All in all, then, we have

| *Chalmers* | *Jackson* | *Kripke* |
|---|---|---|
| *E*'s primary int. $\lvert E\rvert_1$ | the set of *w* in $E_m(w)$ | the set of $E^*$-worlds |
| *E*'s secondary int. $\lvert E\rvert_2$ | the set of *w* in the $E_m(@)$ | the set of *E*-worlds |

What is special about "there is pain" for Chalmers is that its primary and secondary intensions are the same. Unlike, say, "water is $H_2O$", the worlds in which an utterance of "there is pain" expresses a truth are the worlds at which *there is pain*. This is because our instinctive reference-fixer for "pain" (unlike "water") identifies its referent by a necessary and sufficient feature. Pain is the thing that *hurts*.

Now to the argument. If someone claims to find it conceivable that *E* although *E* is not really possible, the explanation is as follows. Conceivability intuitions track conceptual possibility, which

> comes down to the possible truth of a statement when evaluated according to the primary intensions involved . . . The primary intensions of "water" and "$H_2O$" differ, so it is [conceptually] possible . . . that water is not $H_2O$. "Metaphysical possibility" comes down to the possible truth of a statement when evaluated according to the secondary intensions involved . . . The secondary intensions of "water" and "$H_2O$" are the same, so it is metaphysically necessary that water is $H_2O$. (p. 132)

But this sort of story is not available for "pain is distinct from c-fiber firings" or "there are such-and-such physical goings-on without any pain", because

> with consciousness, the primary and secondary intensions coincide . . . The difference between the primary and secondary intensions for the concept of water reflects the

[17] Stalnaker (1987).

> fact that there could be something that looks and feels like water in some counterfactual world that in fact is not water, but merely watery stuff. But if something feels like a conscious experience, even in some counterfactual world, it *is* a conscious experience. (p. 133)[18]

## 9. "FORGET THE SEMANTICS"

Suppose though that someone disagrees (as they have done with Kripke) and says that the way the referent of "pain" is presented in thought can potentially come apart from the state itself; maybe "pain" stands for a condition of the brain importantly implicated in our suffering, a state that could in principle occur without phenomenal accompaniment.

This wouldn't necessarily bother Chalmers; his basic and underlying point, which he repeats again and again, is meant to be without prejudice to the proper semantics for phenomenal terms. The point is that we surely conceive *some* kind of world when we seem to conceive a zombie world; and that world constitutes a counterexample to physicalist supervenience *whatever* we say about the semantical issue:

> . . . nothing about Kripke's *a posteriori* necessity renders any [conceptually] possible worlds impossible. It simply tells us that some of them are misdescribed, because we are applying terms according to their primary intensions rather than the more appropriate secondary intensions . . . It follows that if there is a conceivable world that is physically identical to ours but which lacks certain positive features of our world, then no considerations about the designation of terms such as "consciousness" can do anything to rule out the metaphysical possibility of the world. We can simply forget the semantics of these terms, and note that the relevant possible world clearly lacks *something*, whether or not we call it "consciousness" . . . the mere possibility of such a world, no matter how it is described, is all the argument [against physicalism] needs to succeed. (p. 134)

This is Textbook Kripkeanism at its purest and best: even the illusion of a zombie world is a correct perception of *something*, and that something is all we need to put physicalistic supervenience to rest.

[18] If Chalmers is right about the primary and secondary intensions coinciding, this gives him a small advantage over Jackson. Jackson had to convince us that the contextual information needed to home in on the relevant secondary intensions was *physical* information. But a sentence's *primary* intension is (like Jackson's meaning-function from which it can be defined) independent of context, and so no contextual information is needed to home in on it. Agreement between the two intensions in this case means that no extra information is needed to home in on the secondary intension either. Just by understanding the sentence "there are zombies", we know what proposition it expresses. It follows that if modal intuition appears to detect a zombie world, there is no chance whatever that it is really fastening on some *other* sort of world which it is then misidentifying as one that contains zombies. Our grip on the notion of a zombie world is just too good for that.

## 10. DE RE AND DE DICTO

Now, let's grant Chalmers that the difference between conceptual and metaphysical possibility is all at the level of statements, not worlds: where worlds are concerned the two sorts of possibility are really just one. His reasoning then appears strong:

(1) it is conceptually possible for there to be zombies, so
(2) zombie worlds are conceptually possible, so
(3) zombie worlds are metaphysically possible.[19]

But although (2), on a natural reading, follows from (1), and (3) follows from a natural reading of (2), I wonder whether the two readings agree. The version of (2) entailed by (1) is

(2′) it is conceptually possible that there be zombie worlds.

(If you can imagine zombies, then you can imagine them plus their surrounding worlds.) But what you need to get (3) is

(2″) there are conceptually possible zombie worlds.

And the *de dicto* possibility of zombie worlds asserted by (2′) would seem to fall well short of the *de re* possibility asserted by (2″).

The principal charm, as I see it, of Chalmers' procedure is that he has found a way of reaping the rewards of this de re/de dicto fallacy without actually having to commit it. He maintains, remember, that

(x) conceptual possibility "comes down to the possible truth of a statement when evaluated according to the primary intensions involved" (p. 132).

This allows him to reach (2″) *directly from* (1):

(1) it is conceptually possible that there be zombies, so (by (x))
(1′) there are worlds in the primary intension of "there are zombies", so
(1″) there are worlds which if actual make "there are zombies" true, so

(since worlds like *that* would seem to be all you could want in the way of a conceptually possible zombie world)

(2″) there are conceptually possible zombie worlds.

The point is that it is (x) that saves the argument from being a straightforward modal fallacy. And if we now ask, why believe (x), the reasons turn out to be essentially Jackson's: they trace back to the assumption (+) that to know which

[19] Although "zombie world" may not be quite the right description. I'll ignore this.

worlds *E* is true in is to know a lot of other things besides. Here is how I imagine the argument going.

Chalmers tells us that we can "think of the primary and secondary intensions as the a priori and a posteriori aspects of meaning, respectively" (p. 62). What is understanding, though, if not grasping "the a priori aspect of meaning"? It follows that what a speaker understands by *E* is given by *E*'s primary intension: the worlds which, considered as actual, confer truth on *E*. If *E* is conceptually possible, that's because the speaker's understanding—her grasp of the truth-conferring set of worlds—leaves it open that *E* might be true. But, and this is where (+) comes in, it would *not* leave this open if *E* was true in no worlds whatsoever. Hence we can be assured that *E*'s primary intension is nonempty.

But now wait. To understand "there are zombies", I have to know that it is true in a world *w* iff *w* has such-and-such physical features with no consciousness. I *don't* have to know, though, whether that condition is satisfiable. It would be just as well, in fact, if I *didn't* know; any knowledge that I might have on the topic should be kept under wraps in this context. (Imagine that someone wants to test my understanding of "there are zombies" by asking which worlds it is true in; the reply "*no* worlds" would be silly *even if it were correct.*) Understanding is knowing what a world *has to be like* for "there are zombies" to be true in it, regardless of how easy or difficult it may be for worlds like that to exist.

Here is the response I expect. Just as earlier we abstracted away from controversies about primary vs. secondary intensions, let us now abstract away from the doctrine of intensions altogether. Forget about (1) in other words; we can arrive at (2) another way. All we need is the Kripkean lesson that as far as *worlds* are concerned, conceptual and metaphysical possibility are one and the same. To the extent that I see no *conceptual* obstacle to a world—to the extent that I find it conceivable—I have to admit it as possible in the only sense of the word that applies.[20] That leaves the question of course of how to *describe* this world. Chalmers is confident, though, that under any reasonable description, it constitutes a counterexample to physicalism.

But it is no doctrine of Kripke's that I first conceive worlds, and only later stop to ask what might be true of them. What would it be to find a world conceivable "in itself", as opposed to finding it conceivable that there should be worlds of some specified type? I take it that the latter phenomenon is the only real one, and that the talk of conceivable *worlds* always being possible has to be understood as code for something else: the claim that if *E* is conceivable, then *something* is possible, only perhaps not *E* itself. And that is just Textbook Kripkeanism, the view we are trying to find reason to believe.

[20] Chalmers: "every conceivable world is logically possible" (p. 66).

## 11. WHY TEXTBOOK KRIPKEANISM (ONLY) SEEMS RIGHT

At the heart of Textbook Kripkeanism lies thesis (x). What is the evidence for it? Nobody doubts that a primary-intension-like notion has shown itself to have *some* predictive value in this area. But the inference from (1) to $(1')$ presupposes that there is *no way whatever* of arranging for conceptual coherence short of including a world in the primary intension. Here is my best shot at a supporting argument.

1. $E$ is conceptually possible. (P)
2. Understanding $E$ leaves it open that $E$ might be true. (1)
3. Understanding is knowing how truth depends on worldly context. (P)[21]
4. Knowing how $E$'s truth depends on context leaves it open that $E$ might be true. (2, 3)
5. $E$ is true in some worldly context: some possible $w$ considered as actual. (4)
6. $E$ is true in $w$, considered as actual, iff $w$ is an $|E|_1$-world. (Def. of $|E|_1$)
7. So, $|E|_1$ contains at least one world. (5, 6)

This at least has the right shape to advance us from de dicto to de re possibility. The problem is that, everything above it granted, line 5 doesn't follow. All we get from 4 is that *my way of thinking* of $\{w \mid w$ makes $E$ true$\}$ leaves it open that the set might have members. And that is compatible with its being the empty set in fact.

Suppose for example that $E$ is $P$ & -$C$, where $P =$ "everything is physically like so" and $C =$ "there is consciousness". To understand $E$, it's enough to understand its conjuncts, that is, to know that $P$ is verified by the worlds that are physically like so, and that $C$ is verified by the worlds where there is consciousness. To know in *these* ways the truth conditions of $P$ and $C$ does not begin to tell me whether a world verifying the first can avoid verifying the second. Once again, understanding is knowing what a world *has to be like* to verify a statement; how easy or difficult it may be for worlds like that to exist is another matter entirely.

## 12. IMMACULATE CONCEPTION

The gap in the argument has to do with disparate ways of conceiving the same worlds. One could close it by requiring the understander to conceive the

[21] Jackson would say: how its truth *conditions* depend on context. But the difference isn't important here.

truth-conferring worlds in a *single fixed way*, or, alternatively, *in no way at all*. The first strategy has already been tried; let me not repeat it here. The second or "immaculate conception" strategy tries to relate speakers to sets of worlds *directly*, by which I mean *not* under this or that mode of presentation. Rather than knowing a *condition* that the *E*-worlds satisfy, you "know which worlds the *E*-worlds are" iff you know how to *recognize* an *E*-world when you encounter it.

Encounter it where? The encounter had better not be in imagination, because worlds are imagined under descriptions and it is the relativity to description that we are trying to get beyond. The idea has got to be that *plopped down in w* with the mission of determining *E*'s truth value there, I would conclude that *E* is indeed true. Here is Chalmers:

> What would we say if the world turned out this way? What would we say if it turned out that way? For instance, if it had turned out that the liquid in lakes was $H_2O$ and the liquid in oceans XYZ, then we probably would have said that both were water. . . . (p. 58)

The suggestion more generally is that the primary intension of my expression *E* is the mapping from worlds *w* to the extensions *I would assign to E as an actual inhabitant of w*. This will have to be a me that is idealized in various respects: computing power, mobility, ability to withstand high temperatures, and so on. But the general shape of the strategy should be clear enough.

If intensions are understood like this, then the original relativity in which I know the membership of a set of worlds under one description but not another is indeed mitigated.[22] It is replaced, though, by an *immanent* relativity in which *E*'s extension at a world varies according to my in-world representative's point of view.

An initial reason for this is that extensions tend to be presented in indexical terms. "Water" refers to the predominant clear and drinkable liquid *around here*. Hence if *w* has different such liquids in different places, there will be no simple answer to what "water" would/should be seen as referring to in *w*. This is why Chalmers says that it is not worlds *simpliciter* that go into primary intensions, but *centered* worlds fitted out with a marked space-time point or a designated individual and time.

No sooner do we recognize the need for a center, though, than we notice ways in which it needs to be enriched and expanded. Some referents are identified by their psychological effects (whatever causes *this* sensation), so room will have to be made for aspects of the speaker's psychology.[23] The center should probably also include some indication of which direction is left, and which right, and perhaps also what the speaker is attending to at any given moment, the figure/ground

[22] Why "mitigated" and not gone? (1) It's not clear that my descriptive dispositions in a given world are a priori accessible to me. (2) It's not clear that the range of worlds in which these dispositions are to be exercised is a priori accessible to me. (3) It's not clear that either sort of access can be arranged without making ourselves again vulnerable to mode-of-presentation worries.

[23] A point that Chalmers happily acknowledges.

relations in her visual field, and what may be occurring to her in memory. All of these factors can and do figure in the interpretation of the indexical phrases by which the speaker fixes the referents of her terms.

A quite different way for perspective to intrude is mentioned in a footnote attached to the passage quoted—a footnote which reinforces the impression of an investigator hypothetically parachuting down into a world with the mission of deciding what there falls into the extensions of his words. It sometimes happens that

> whether we count an object as falling under the extension of a [word] will depend on various accidental historical factors. A stimulating paper by Wilson (1982) discusses such cases, including for instance a hypothetical case in which druids might end up classifying airplanes as "birds" if they first saw a plane flying overhead, but not if they first found one crashed in the jungle. (p. 365)[24]

The center thus needs to take notice of the *order* in which various sorts of cases are presented. And this calls to mind lots of *other* factors capable of influencing the agent's referential inclinations in not overtly indexical ways: her hunches at any particular point about how representative the observed cases have been, her larger theoretical and practical projects, her beliefs about which sorts of classifications are going to serve these projects, how anxious she is to avoid multiplying entities, how *physicalistic* she is—the whole sorry mess of presumptions and prejudices that guide us in our application of old words to new cases.[25]

All right, but why should this be a problem? The reason for going hypothetically native was to secure for ourselves an unmediated grasp of primary intensions; the primary intension of a statement found conceptually possible would then *have* to contain at least one world, which world could then be used (in the case of interest) as a counterexample to physicalism. If primary intensions are made up not of worlds per se, but worlds-as-experienced-and-theorized-from-such-and-such-a-standpoint, then this rationale springs a large leak. For it could happen that whenever *w* as seen from one perspective (as fitted out with one center) makes it into the primary intension of *E, w* as seen from another perspective does not. In that case there is no determinate fact of the matter as to the emptiness or not of *E*'s primary intension. (To say that the primary intension determinately contains *w*-as-seen-from-such-and-such-a-perspective achieves nothing; our interest as modal metaphysicians is in the possibility of *w* as such, unelaborated.)

An example might be this. Suppose that my idealized self takes up residence in a world where events that I am inclined to call *pains* occur on all the same occasions as events that I am inclined to describe as *c-fiber firings*. Whether I decide that "pain" and "c-fiber firing" pick out one and the same type is hardly likely to be settled by my competence with the relevant terms; a lot will

[24] The paper by Wilson is "Predicate Meets Property" (1982).

[25] Some semi-pertinent cognitive science literature is summarized in Smith and Osherson (1995).

depend on background attitudes about ontological economy, modal intuition, the transparency of the mental, and so on. This is clear from the great identity debates of the 1950s, when it was widely assumed that mental/physical correlations would soon be found and the question was what ontological conclusions to draw.[26]

The claim is that it is utopian to expect unaided understanding to decide philosophically loaded questions, even given a full statement of pertinent facts—up to, but not including of course, facts about how those very questions are to be answered. A lot is going to depend on factors that are hard to see either as semantical or factual, with the result that a world that is counted into *E*'s primary intension on one accounting is liable to find itself counted out under another. This seriously limits the metaphysical use that can be made of our alter egos' in-world judgments. If the dualist is allowed to claim *w* as a world in which pain and c-fiber firings are *distinct*, because that is a conclusion that a well-informed inhabitant of *w* could reasonably draw, why shouldn't the identity theorist be allowed to claim *w* as a world in which they are *identical*, for the same reason?

The dualist could reply as follows. Look, you may be right about some possible worlds; there is no determinate answer to whether they in themselves, as opposed to they-as-judged-from-this-or-that-perspective, are to be described in a way that favors physicalism or in a way that doesn't. But there are *other* worlds whose anti-physicalistic import is *so clear and unmistakable* that all well-informed observers are going to agree. Take a zombie world, for instance; no one could think that pain was identical to c-fiber firings *there*, because that world doesn't *have* any pain.

But to assume that zombie worlds are indeed possible just forgets the reason we handed descriptive authority to our in-world representatives. Their role was to clear the path to a nonempty primary intension, i.e., to a zombie world. For my representative to be told outright whether *w* verifies *E* (whether others feel pain) obviously defeats the purpose, since I would be reclaiming his descriptive authority for myself. If he is *not* told outright, however, then a zombie world has no better claim to membership in $|\text{there are zombies}|_1$ than does a world like ours; after all, my representative cannot tell them apart. To the extent that the "immaculate conception" strategy buys us a world, then, physicalism is unbothered. The world might be our own, consciousness and all.

## 13. CONCEIVABILITY

One thing is clear: modal intuitions are fallible, and defeasible by reference to empirical data. If Textbook Kripkeanism isn't the way to deal with our occasional misjudgments, what is?

[26] One doesn't think of these debates as driven by differences about the meaning of "pain".

I suspect that Textbook Kripkeanism is the best we can do, if we persist in seeing modal intuition as a capacity that is at bottom conceptual in nature. Let's distinguish three progressively less implausible versions of the conceptualist thesis.

FN:27 *Extremists* say that conceptual conceivability[27] is the only kind there is. Since conceivability is a function of concepts alone, our conceiving faculty is absolutely informationally encapsulated. The role of defeaters on this view is not to educate modal intuition—like perceptual intuition in the Müller/Lyer case, it's quite unteachable—but to alert us that circumstances obtain in which it is not to be trusted. Learning that local water contains hydrogen doesn't make XYZ-water less conceivable; it just stops us from drawing the wrong conclusions from the same old mistaken intuition. It accomplishes that by slotting into a priori conditionals along the lines of "if the stereotypical features of water are grounded in property BLAH, then water is essentially BLAH" to enable results contrary to what our error-prone intuitions continue to suggest.

The objection to this is phenomenological. It is not that we are *forced* to admit that water necessarily contains hydrogen against the evidence of modal intuition. When we learn the empirical truth, our intuitions *change*, and what we used to find conceivable we find conceivable no longer.

*Moderate* conceptualists agree that empirical information has its influence; it fixes the value of the BLAH-parameter in a priori conditionals like the one mentioned above. The difference is that where the extremist sees these conditionals as external *correctives* to intuition, for the moderate they are internal to our conceiving faculty and indeed what drives it. We find *E* conceivable to the extent that we are aware of no information to suggest via a priori conditionals like "if water is made of BLAH, then it is essentially made of BLAH" that *E* is impossible. The role of defeaters on this view is not to overrule an inherently *error-prone* faculty, but to supply a *badly served* faculty—or rather the modal schemata that the faculty relies on—with a better quality of input.

This is certainly an improvement over extremism. But there is a problem about order of explanation. According to the moderate, we are forced by a priori schemata issuing from our concept of water to find hydrogenless water inconceivable. Surely, though, it's the other way around. Rather than the schema determining what we find conceivable, our faith in this (or any) schema derives from the fact that when we assume its antecedent, its consequent becomes modally intuitive. The schema is better cast as a (clumsy) post facto rationalization of a preexisting readiness to let our intuitions evolve in such-and-such ways under the impact of new information.

[27] *Roughly* definable as the non-apriority of not-*E*.

*Weak* conceptualists concede that the dispositions come first, the articulated modal schemata second. But they think that moderates are *correct* to say that modal intuition evolves under the influence of something a priori and conceptually guaranteed; their only mistake was to identify this "something" as the schemata rather than the update dispositions themselves. Weak conceptualists maintain that anyone with our concept of water is obliged to greet the news that existing water samples have such-and-such a microstructure with the same intuitional shift that we did. Of course, it is quite likely beyond our discursive powers to articulate in full detail the function from possible empirical findings to intuitional shifts that a particular concept dictates; one well-known source of perplexity is how to formulate fall-back norms, e.g., the norms telling us how to react if an aspiring natural kind concept (like that of jade) fails to pan out. It remains, however, that there is a conceptually determined truth of the matter about what modal intuitions a given evidential diet would/should evoke in relevantly endowed thinkers.

This is again an improvement. But the link that weak conceptualism postulates between concepts and evidential dispositions is implausibly tight; not enough room is left for the phenomenon of two people sharing a concept while differing in their response to evidence bearing on its application. The worry is that weak conceptualism skates dangerously close to the verificationist idea of "logical probability" relations between statements[28]—relations that all thinkers have got to respect, on pain of irrationality, when deciding how much credence to assign a hypothesis $H$ given evidence $E$.[29]

How is it that weak conceptualism comes dangerously close to logical probability? That a close association would be dangerous shouldn't need a lot of argument. Hardly anyone today thinks that there is a single objectively best epistemic response to a given body of evidence—never mind a best response settled by logic and concepts.[30] The usual view is that rational thinkers, let their concepts be as similar as you like, are liable to range widely along a number of dimensions relevant to their subsequent probabilities. There will be differences, for example, in their personal evidence thresholds; in the kinds of tradeoffs they favor between simplicity and strength; in the importance they assign to avoidance of error as against acceptance of truth; in their attachment to perfect accuracy as against verisimilitude; in how ontologically abstemious they are; and so on and so forth without obvious limit. Rational thinkers will therefore draw different conclusions from the same evidence, blamelessly but in defiance of logical probability.

[28] For more on logical probability, see Keynes (1921), and Kyburg (1970).

[29] The idea is slightly improved by demanding with Carnap that $E$ take in *all* of the evidence bearing on $H$—the "total evidence" requirement. And in fairness to Chalmers, this is all he needs.

[30] ~~Compare Putnam's characterization (somewhere) of the idea of an a priori inductive logic as "intellectual Walden Two".~~

Putnam puts Carnap-style inductive logic on his list of "fantasies of the positivist, who would replace the vast complexity of human reason with a kind of intellectual Walden II" (Putnam 1983, 234)

It remains to be explained how weak conceptualism comes "close" to logical probability. I will argue contrapositively that anyone against logical probability should reject weak conceptualism—that if there can be differences in conditional credence between (rational) subjects with relevantly similar concepts, then there can be differences in conditional conceivability between such subjects, and hence differences in their subsequent modal intuitions. The reason is simply that our modal intuitions are influenced by our beliefs. Learning that Twain = Clemens, or that water contains hydrogen, I cease to find the alternative conceivable. Hence if conceptually congruent thinkers form different beliefs in response to the same evidence, they are going to differ too in what they find conceivable.

## 14. ERROR AND DEFEAT

Our conclusion so far is that not even the weakest form of conceptualism about modal intuition has any plausibility.[31] This brings us back to our original question about how to deal with fallibility and defeasibility. What is the role of defeaters, if it is not to overrule an incurably error-prone faculty, or to correct in principle the input to a faculty that is ~~(when not abused)~~ error-proof? detect intuitive

I uncover my modal errors the same way I ~~uncover intuitional errors~~ generally: miscues more by noticing how my intuitions evolve as I become better educated, while working ~~with the people around me~~ to free myself of errors and oversights that may be prior leading me astray ~~misleading me~~. Here is a first stab at how the process works:

If *X* finds it conceivable that *E*, then she is prima facie justified in believing that *E* is possible. That justification is defeated if someone can provide her with reason to suspect the existence of a *D* such that (i) *D* is true, (ii) if *D* is true, then *E* is impossible, and (iii) that *X* finds *E* conceivable is explained by her ~~failing~~ failure to realize (i) and/or (ii).[32]

Hammurabi was able to conceive it as possible for Hesperus to exist without Phosphorus only because he didn't realize that the two were identical, and (maybe also) that identicals necessarily coexist. The medievals were able to conceive it as possible for dolphins to be cold-blooded only because they didn't realize that dolphins were mammals, and that mammals have got to be warm-blooded. And so on.

Now, it is tempting to suppose that Hammurabi and the medievals were *even at the time* aware of certain specific issues, open to independent investigation,

[31] This section borrows from Yablo (1990) and Yablo (1993).

[32] Likewise, if *Y* finds it inconceivable that *E*, then he is ●*prima facie* justified in believing that *E* is impossible. That justification is defeated if there is reason to suspect the existence of a defeater, that is, a *D* such that (i) *D* is true, (ii) if *D* is true, then *E* is possible, and (iii) that *Y* finds *E* inconceivable is explained by his failure to realize (i) and/or (ii).

● Q2

whose unfortunate resolution would have exposed their intuition as wrong. But it seems truer to the normal progress of modal inquiry that the conceiver is *not* specifically aware of her intuition's vulnerability to its eventual defeater, until the defeater comes along and does its work. Before the discovery of genes, the thought may not have been readily available that scenarios in which animal reproduction was organized along some other, non-genetic, basis were at risk of being exposed as impossible by some experiments with peas. Before it was shown how to account for locomotion, respiration, and so on in biochemical terms, the problem with a scenario in which the property of being *alive* is randomly distributed over physical duplicates must have been hard to appreciate as well. None of this is to deny that the concept of an animal, or of life, must *somehow* prepare the way for the eventual recognition that animals necessarily propagate their kind by way of genes, or that physics guarantees aliveness. But it is striking how unaware it is nevertheless possible to be of the vulnerability of one's intuition to what emerges, in the end, as its defeater. All we have to go on in cases like this is a generalized and undirected sense that defeat is quite possibly on the way, and corresponding feelings of unease about the doomed intuition—feelings that are so strong in some cases as to shift one's intuitive alliances before the defeater even arrives.

## 15. ZOMBIES

Am I the only one who feels the intuition of zombies to be vulnerable in this way? I am *braced* for the information that is going to make zombies inconceivable, even though I have no real idea what form the information is going to take.[33]

FN:33

Of course, as with the concept of life, there has to be *something* in our understanding of consciousness that "prepares the ground" for the eventual discovery that anyone just like me in physical respects must also be conscious. I guess, then, that there is room in *principle* for the project of looking for features of our concepts—of what we understand by the relevant words—that will prevent this discovery from ever being made. Such a project looks a lot less realistic, however, when we realize that grasp of meaning is not a normative crystal ball telling us what modal conclusions are to be drawn from every new empirical finding, however unforeseen or unforeseeable. One could stipulate, I suppose, that a *fully lucid understanding* of *E* would "anticipate" in some way the bearing of all possible observations on *E*'s modal status, in all possible methodological climates (etc.). But that's not the kind of understanding we have, and I imagine not the kind anybody would want.

[33] I expect to be told ~~Some may think~~ that I *should* know what forms it ~~is going to take~~ *can* take—that defeating information should slot neatly into some pre-identified schematic element in my concept of consciousness⊙ ~~in such a way as to make zombies a priori unthinkable~~. This for me would be an example of overlooking, or underestimating, the open texture of concepts.

## REFERENCES

Chalmers, D. (1996). *The Conscious Mind.* New York: Oxford University Press.

Hartshorne, C. (1941). *Man's Vision of God*. New York: Harper Row; excerpted in Plantinga (1965) pp. 122–135●. ● Q3

Jackson, F. (1994). "Armchair Metaphysics", in M. Michael and J. O'Leary-Hawthorne (eds.), *Philosophy in Mind*, Dordrecht: Kluwer pp. 24–42●. ● Q4

Keynes, J. M. (1921). *A Treatise on Probability*. London: Macmillan.

Kripke, S. (1980). *Naming and Necessity*. Cambridge, Mass.: Harvard, University Press.

Kyburg, H. (1970). "Degree-of-Entailment Interpretations of Probability", in his *Probability and Inductive Logic*, London: Macmillan, ~~●Chapter 5~~ pp. 54–67. ● Q5

Putnam, H. (1983). "Why Reason Can't Be Naturalized". In his *Philosophical Papers*, Volume 3, pp. 229–247

Plantinga, A. (ed.) (1965). *The Ontological Argument*. New York: Doubleday.

Smith, E., and Osherson, D. (eds.) (1995). *Thinking*. Cambridge, Mass.: MIT Press.

Stalnaker, R. (1987). "Semantics for Belief", *Philosophical Topics* 15, pp. 177–90.

Waismann, F. (1965). "Verifiability", in A. Flew (ed.), *Logic and Language*, New York: Doubleday, pp. 122–51.

Wilson, M. (1982). "Predicate Meets Property", *Philosophical Review* 91, pp. 549–89.

Yablo, S. (1990). "The Real Distinction between Mind and Body," *Canadian Journal of Philosophy*, supp. vol. 16, pp. 149–201; Ch. 1 above.

——— (1993). "Is Conceivability A Guide to Possibility?" *Philosophy and Phenomenological Research* 53, pp. 1–42; Ch. 2 above.

Looks OK to me!

## Queries in Chapter 3

Q1. Please check and confirm the correction marked by author in pencil here.

I don't know Oxford's policy on italicizing latinate phrases.

Q2. Please check and confirm whether this word should be in Italics or in Roman, we have captured it in Italics.

Q3. Please check and confirm whether the page numbers inserted by us is fine or not.

Fine

Q4. Please check and confirm whether the page numbers inserted by us is fine or not.

Q5. Please check and confirm the author correction here.

Corrected in situ.

# 4

# Coulda, Woulda, Shoulda

## 1. TERMINOLOGY

A main theme of Saul Kripke's *Naming and Necessity* (1980) is that metaphysical necessity is one thing; apriority, analyticity, and epistemic/semantic/conceptual necessity are another. Or rather, they are others, for although the relations among these latter notions are not fully analyzed, it does emerge that they are not the *same* notion.

'Apriority' and 'analyticity' are for Kripke nontechnical terms. They stand in the usual rough way for knowability without appeal to experience, and truth in virtue of meaning. Examples of apriority are given that it is hoped the reader will find plausible. And a schematic element is noted in the notion of knowability without experience; how far beyond our own actual cognitive powers are we allowed to idealize? Beyond that, not a whole lot is said.

Analyticity, though, does come in for further explanation. The phrase 'true in virtue of meaning' is open to different interpretations, Kripke says, depending on whether we are talking about 'meaning in the strict sense' or meaning in the looser sense given by a term's associated reference-fixing description. A sentence like 'Hesperus is visible in the evening' comes out loosely analytic but not strictly so, since the meaning proper of 'Hesperus' is exhausted by its standing for Venus.

Kripke stipulates that 'analytic' as he uses the term expresses *strict* analyticity, and he takes this to have the consequence that analytic truths in his sense are metaphysically necessary truths ('an analytic truth is one which

This paper owes a lot to discussions over the years with Ned Block, Alex Byrne, Tamar Szabó Gendler, Sally Haslanger, John Hawthorne, Frank Jackson, Joe Levine, Brian Loar, Jim Pryor, Gideon Rosen, Sydney Shoemaker, and Robert Stalnaker. A larger debt is to David Chalmers; if I still haven't got two-dimensionalism right, Lord knows the fault is not his. I thank him for years of patient explanation and good-natured debate. May heaven smile on Tamar and John for their extraordinary work on this volume, and in particular for the excellent Introduction. I am grateful, finally, to Saul Kripke for *Naming and Necessity*, three lectures so inconceivably great as to hardly seem possible.

depends on *meanings* in the strict sense and therefore is necessary' (1980: 122 n. 63)). He notes, however, that one might equally let the word express loose analyticity, and that on that definition 'some analytic truths are contingent' (ibid.).

Given the care Kripke takes in distinguishing the kind of analyticity that entails metaphysical necessity from the kind that doesn't, one might have expected him to draw a similar distinction on the side of apriority: there would be an apriority-entailing kind of analyticity and a kind that can be had by non-a priori statements. 'Hesperus is Phosphorus' is not a priori, but since its meaning is a proposition of the form $x = x$, and any proposition of that form is true, it could be considered true in virtue of meaning. I am not endorsing this particular example, just pointing out a move that could have been made.

Kripke seems, however, to take it for granted that analytic truths will be a priori knowable. In his characterization of loose analyticity he speaks, not of statements whose truth is guaranteed by reference-fixing descriptions, but ones whose '*a priori* truth is *known* via the fixing of a reference' (1980: 122 n. 63; italics added). A non-Kripkean line on the apriority of analytic statements will be elaborated below.

I said that apriority and analyticity were for Kripke (relatively) 'ordinary' notions. There are intimations in *Naming and Necessity* of a corresponding technical notion: a notion that explicates apriority/analyticity as metaphysical necessity explicates our idea of that which could not be otherwise. This technical notion—potentially a partner in full standing to metaphysical necessity—needs a name of its own. What should the name be?

'Epistemic necessity' is best avoided because, as Kripke says, to call S epistemically possible sounds like a way of saying that it is true (or possible) for all one currently knows.[1] A notion explicating apriority/analyticity should not be so sensitive to the extent of current knowledge. One doesn't know how to prove Goldbach's conjecture today, but one might tomorrow; it would then turn out to have been necessary (in the partner sense) all along.

'Semantic necessity' too is liable to mislead, since for some people, Kripke included, 'Hesperus' and 'Phosphorus' are semantically just alike, yet it is possible in the partner sense that Hesperus $\neq$ Phosphorus. As Kripke says, one is inclined to think that it could have turned out either way.

If a name is to be given, then, to the *non*metaphysical modality that features in *Naming and Necessity*, 'conceptual' is probably the least bad. It is true that Kripke doesn't use the word 'conceptual' and doesn't talk much about concepts. But his nonmetaphysical necessities do have their truth guaranteed by the way

FN:1

[1] DeRose (1991) argues that this familiar condition is not enough. If contrary information is there for the taking, and/or possessed by relevant others, then S is not epistemically possible, even if it could be true for all I myself know.

we have represented things to ourselves; and we can think of 'concept' as just evoking the relevant level of representation. Conceptual necessity will then be the technical or semi-technical notion that Kripke runs alongside, and to some extent pits against, metaphysical necessity.

## 2. CONCEPTUAL NECESSITY

An enormous amount has been done with the metaphysical/conceptual distinction. Yet, and I think this is agreed by everyone, the distinction remains not terribly well understood. One reason it is not well understood is that the conceptual side of the distinction ∧ ~~didn't~~ receive at Kripke's hands the same sort ∧did not of development as the metaphysical side.

This might have been intentional on Kripke's part. He might have thought the conceptual notion to be irremediably obscure, but important to mention lest it obscure our view of metaphysical necessity. Certainly this is the attitude that many take about the conceptual notion today. It could be argued that much of the contemporary skepticism about narrow content is at the same time skepticism about conceptual possibility. Narrow content, if it existed, would give sense to conceptual possibility: holding its narrow content fixed, S could have expressed a truth. If one rejects narrow content, one needs a different explanation, and none comes to mind. Going in the other direction, one might try to define S's narrow content as the set of worlds *w* whose obtaining conceptually necessitates that S. Lewis remarks somewhere that whoever claims not to understand something will take care not to understand anything else whereby it might be explained. If you don't understand narrow content, you will take care not to understand conceptual possibility either.

But, although many people have doubts about conceptual possibility, a number of *other* people are entirely gung ho about it. Some even treat it (and narrow content) as more, or anyway no less, fundamental than metaphysical possibility (and broad content). An example is David Chalmers. He calls S's narrow content its 'primary intension', and its broad content its 'secondary intension'. One suspects that the order here is not accidental. And even if the suspicion is wrong, the primary intension is certainly a partner in full standing.

In this paper I try not to take sides between the skeptics and the believers. My topic is how conceptual possibility should be handled *supposing it is going to be handled at all*. If I do slip occasionally into the language of the believers, that is because I am trying to explore their system from the inside, in order to see what it is capable of, and whether it can be made to deliver the advertised kinds of results. (I should say that my own leanings are to the skeptical side, though I think the issue is far from settled.)

## 3. INITIAL COMPARISONS

Kripke's theory (or picture) of metaphysical modality is familiar enough. He says that it holds necessarily that S iff S is true in all possible worlds. The word 'in' is, however, misleading. It suggests that S (or an utterance thereof) is to be seen as *inhabiting* the world(s) *w* with respect to which it is evaluated. That is certainly not Kripke's intent. His view is better captured by saying that S (that well-known denizen of *our* world), to be necessary, should be true *of* all possible worlds. Every world should be such that S gives a correct description of it. Every world should be such that the way S describes things as being is a way that it in fact is.

Conceptual possibility too is explained with worlds. To be conceptually possible is to be in some appropriate sense true with respect to—or, for short, true *at*—*w* for at least one world *w*. But what is the appropriate sense? Everyone knows the examples that are supposed to bring out how conceptual modality is different. It is conceptually possible, but metaphysically impossible, for Hesperus to be distinct from Phosphorus. This is because 'Hesperus $\neq$ Phosphorus' is true at a world that it fails to be true of. The metaphysical/conceptual contrast thus hangs on the contrast between true-of-*w* as just discussed and the notion of true-at-*w* that we must now attempt to develop.

ital ital

Here is the obvious first stab: S is true at *w* iff S as uttered in *w* is true of *w*. 'Hesperus $\neq$ Phosphorus' uttered here in the actual world means that Venus isn't Venus; uttered in *w*, it might mean that Venus isn't Mars. If, in *w*, Venus indeed *isn't* Mars, then 'Hesperus $\neq$ Phosphorus' is true at *w*. And so *w* testifies to the conceptual possibility of Hesperus not being Phosphorus.

Compare now an S that strikes us as *not* conceptually possible: for instance, 'Phosphorus $\neq$ Phosphorus'. Uttered in *w*, this means that Mars $\neq$ Mars. Since that is false of Mars, in *w* or anywhere else, *w* does not testify to the conceptual possibility of Phosphorus not being Phosphorus. Unless there are worlds where uttering 'Phosphorus $\neq$ Phosphorus' is speaking the truth, that Phosphorus $\neq$ Phosphorus is not conceptually possible.

But, and here is where the trouble starts, there *are* worlds like that. For there are worlds in which 'Phosphorus $\neq$ Phosphorus' means something other than what it actually means (say, that Phosphorus is *identical* to Phosphorus) and in which the other thing is true. So it looks like we reach the wrong result. It should not make 'Phosphorus $\neq$ Phosphorus' conceptually possible that there are worlds in which '$\neq$' expresses identity!

One remembers this sort of problem from Kripke's discussion, not of conceptual possibility, but metaphysical possibility. Let it be, he says, that *w* contains speakers (maybe our counterfactual selves) who understand S eccentrically from our point of view. That has no bearing on the issue of whether S is true of *w*:

> when we speak of a counterfactual situation, we speak of it in English, even if it is part of the description of that counterfactual situation that we were all speaking [another language] . . . We say, . . . 'suppose we had been using English in a nonstandard way'. Then we are describing a possible world or counterfactual situation in which people, including ourselves, did speak in a certain way different from the way we speak. But still, in describing that world, we use *English* with *our* meanings and *our* references. (1980: 77)

By 'tail', for example, the inhabitants of *w* might mean *wing*. If so, then assuming *w*'s horses resemble ours, they speak falsely when they say 'horses have tails'. That is irrelevant, Kripke says, to the metaphysical necessity issue. 'Horses have tails' is as true of *w* as of the actual world. This is crucial if statements are to come out with the right modal status. 'One doesn't say that "two plus two equals four" is contingent because people might have spoken a language in which "two plus two equals four" meant that seven is even' (1980: 77).

How much of this still applies on the conceptual side? Worlds where 'Hesperus ≠ Phosphorus' means that Venus ≠ Mars *can* (as we saw) bear witness to the conceptual possibility of Hesperus not being Phosphorus. So in judging conceptual contingency, we *do* want to look at *w*-speakers who, in a broad sense, mean something different by S than we mean by it here.

But there are limits; we are not interested in *w*-speakers who by 'Hesperus ≠ Phosphorus' mean that Hesperus is identical to Phosphorus, or that it's snowing in Brooklyn. It thus becomes important to know in what ways the meaning of S in the mouths of *w*-speakers can differ from the meaning of S in our mouths, for the truth of S as uttered in *w* to be relevant to the conceptual possibility of S here. Something has got to be held fixed, but what?

## 4. HOLDING FIXED

First try: S has got to mean the *very same* in *w* as it means here.

This holds too much fixed. 'Hesperus' and 'Phosphorus' as they are used here both mean Venus, and '≠' expresses nonidentity. A counterfactual utterance of 'Hesperus ≠ Phosphorus' that respected these facts would have to mean that Venus ≠ Venus; and so the utterance would not be true. But then it will not come out conceptually possible that Hesperus ≠ Phosphorus, as it should.

Second try: Corresponding expressions should mean the same, *or* have their references fixed by the same or synonymous descriptions.

This is all right as far as it goes, but there is a problem of coverage. If a reference-fixing description is one that picks out the referent no matter what, then reference-fixing descriptions are hardly ever available. One doesn't know of any description guaranteed in advance to pick out the referent of 'Homer' or 'water'. So the second proposal reduces in most cases to the first, which we've seen to be inadequate.

A third approach puts conditions not on S in particular, but on *w* as a whole: *w* bears on S's conceptual possibility if and only if it is an 'epistemic counterpart' of our world, in the sense of confronting the speaker with the same evidential situation as he confronts here. If *w* is an epistemic counterpart of actuality, then S's meaning can change only in ways that leave the evidential situation as is; that is what it takes for S's truth in *w* to bear witness to its conceptual possibility here.

A seeming advantage of the proposal is that it no longer attempts to specify the relevant aspects of meaning (the ones that are supposed to be held fixed) explicitly. The thought is that those aspects, whatever they are, are fixed *inter alia* by fixing the entire evidential situation. This is also the proposal's problem, though. Mixed in with the semantical material we want to hold fixed will be nonsemantic circumstances that should be allowed to vary. One doesn't want to hold fixed that there seems to be a lectern present, or there seeming to be a lectern present will be classified as conceptually necessary. That is clearly the wrong result. Appearances are conceptually contingent if anything is.

## 5. SUBJUNCTIVES

The kind of necessity we are calling conceptual is left by Kripke in a precarious state. Judging conceptual necessity is judging whether S *as uttered in w* is true of *w*. This collapses into triviality unless certain aspects of S's meaning are held fixed. And it is unclear which aspects are intended.

Why do the same problems not arise for metaphysical necessity? The usual answer is that with metaphysical necessity, one needn't bring in a counterfactual utterance at all. One considers whether *our* utterance, saying (or meaning) just what it actually says (means), gives a true description of *w*. But this doesn't give us much guidance in some cases.

Suppose we are trying to evaluate 'horses have tails' with respect to *w*. You maintain, reasonably enough, that what 'horses have tails' actually says is that tails are had by Northern Dancer, Secretariat, . . . (fill in here the list of all actual horses). You conclude that 'horses have tails' is true of *w* iff Northern Dancer, Secretariat, . . . (or perhaps just those of them that exist in *w*) have tails in *w*.

Someone else maintains, just as reasonably, that 'horses have tails' says that if anything is a horse, then it has a tail. She concludes that 'horses have tails' is true of *w* iff the things that are horses in *w* have tails in *w*. The two of you disagree, then, about how to evaluate 'horses have tails' at a world that contains all our horses (complete with tails) plus some *additional* horses that lack tails.

Who is right? What is really said by an utterance of 'horses have tails' and how do we tell whether it is true of a counterfactual world? These questions have no clear answers. One might, I suppose, look for answers in the theory of what is expressed, or what is said, by sentences in contexts. But it would be with a heavy heart (and not only because the notion of what is said is so slippery and vague). Almost every question in semantics can be framed as a question about what some S expresses in some context. It would be nice if we didn't have to do the full semantics of English before the truth-conditions of 'necessarily S' could be given.

If there were no way around this problem, I doubt that Kripke's approach would have found such widespread acceptance. One imagines, then, that the Kripkean has a response. Here is how I imagine it going: 'You are taking the "saying what it actually says" phraseology too seriously in some way. If any real weight were going to be laid on that way of putting it, then yes, a story would be needed about how it is determined what is said. But "saying what it actually says" is just a heuristic. It reminds us that it doesn't matter, in considering whether S is true of *w*, what the citizens of *w* mean by S. How in that case *is* true-of to be understood, you ask? One option is to treat it as primitive. But this option is problematic. It gives the skeptic about metaphysical possibility too big an opening: she can claim to find the primitive incomprehensible. It would be better if we could *explain* truth-of in terms that the skeptic, as a speaker of English, already understands. This can be done using the subjunctive conditional. To say that S is true of a world *w* is to say that *had w obtained, it would have been that S*.'[2]

Consider in this light the 'controversy' about horses and their tails. When we evaluate 'horses have tails' with respect to *w*, is it only the actual horses that matter, or do horses found only in *w* have to be taken into account as well? Suppose that although actual horses have tails, *w*'s additional horses include some that are tail-less. Is 'horses have tails' true of *w*?

The subjunctive account makes short work of this conundrum. Had *w* obtained, it would *not* have been that horses had tails; there would have been some horses with tails and some without. So 'horses have tails' is false of *w*.

Return now to the case of a *w* where 'tail' means *wing*. Does the fact that *w*-people speak falsely when they say 'horses have tails' show that 'horses have tails' is false of *w*? It doesn't, and we can now explain why in a theoretically uncontroversial way. The question is whether horses would still have had tails, if people had used 'tail' to mean wing. They clearly would have; how people talk doesn't affect the anatomy of horses. Had 'tail' meant wing, 'horses have tails' would not have been true, but horses would still have had tails.

## 6. DISPARITY

All this is to emphasize the *disparity*, in the immediate aftermath of *Naming and Necessity*, between metaphysical and conceptual necessity. The first was in

[2] See in this connection Chalmers (2000).

good shape—because it went with ‘S is true of *w*’, which could be understood as ‘it would have been that S, had it been that *w*’. The second was in bad shape—because it went with ‘S is true when uttered in *w*’, which had to be understood as ‘it would have been that S was true, had it been that *w, and had S retained certain aspects of its actual meaning*’.

Then a brainstorm was had that seemed to restore parity.[3]

Recall what we do to judge metaphysical necessity. We ask of various worlds *w* whether S (*our* S, natch) is true *of w*. The Kripkean tells us that to judge conceptual necessity, we need to ask, not whether S is true *of w*, but whether it is true (as spoken) *at w*. But maybe it wasn’t really necessary to move S over to *w*. A different option is to move *w* over to actuality: to the place where the token of S that we want evaluated in fact occurs.[4]

All right, but how do we do that? It looks at first very simple. Just as, when judging metaphysical necessity, we consider *w* as counterfactual, so, when judging conceptual necessity, we consider it as *counteractual.* We consider it as a hypothesis about what *this* world is like. Of course, we do not in general *believe* the hypothesis. But that should not deter us; we are masters at working out how matters stand on hypotheses we reject. Evaluating S with respect to counteractual *w* is asking whether S holds on the hypothesis that *w* is (contrary to what we perhaps think) this very world.

For example, it is conceptually possible that Hesperus $\neq$ Phosphorus because, if we suppose for a moment that this world is one in which Hesperus-appearances are due to Mars and Phosphorus-appearances to Venus, then clearly (on that supposition) we are *wrong* to think that Hesperus = Phosphorus. It is not that counterfactual people are wrong about *their* world. It is we who are wrong about *our* world, on a certain hypothesis about what our world is like.

This sounds like progress, but we should not celebrate too soon, because the disparity with metaphysical modality is not entirely gone.

I said that everyone would (should!) have been unhappy if they had been asked to treat ‘true of counterfactual *w*’ as a semantic primitive. We are willing

[3] At least three ideas were involved. (1) Instead of moving S over to *w*, bring *w* back to S. To do that, (2) evaluate S on the hypothesis that *w* actually obtains. To do that, (3) evaluate the indicative conditional ‘if *w* actually obtains, then S’. (1) and (2) are present to some degree in Evans (1979) and Davies and Humberstone (1980), and are explicit in Chalmers (1994). I am not aware of any discussion of (3) before Chalmers (1996, 2000). See also Segerberg (1972), White (1982), and Stalnaker (1972, 1990, 1991).

[4] A third option is to leave S and *w* where they are, and treat ‘true if’ as a trans-world primitive. This is one possible reading of Chalmers’s remark (1994) that ‘we can retain the thought from the *real* actual world and simultaneously ask its truth-value in other actual-world candidates without any loss of coherence’. He adds in a footnote that ‘Doing things this way . . . avoids a problem . . . raised by Block (1991) and Stalnaker (1991). The problem is that of what must be “held constant” between contexts . . . On my account, nothing needs to be held constant, as we always appeal to the concept from the real world in evaluating the referent at [an actual-world candidate]’ (1994:42). This is certainly one way to go. But it has its costs. If taking ‘true of’ as primitive is obscurantist, primitivism about ‘true if’ borders on mysticism (our pre-theoretical grip on the second is that much weaker).

to rest so much on true of because of the *explanation* we have been given of that notion: S is true of *w* iff, had *w* obtained, it would have been that S. It is this biconditional, with 'true of' on the left and a counterfactual on the right, that convinces us that there's a there there.

Apart, though, from some suggestive talk about what to say 'on the supposition' that *w* obtains, we have no comparable explanation of what is involved in S's being true with respect to counteractual *w*. If we use 'true if *w*' for truth with respect to a world conceived as actual, the problem is that 'true of' has been translated into English and 'true if' has not.

## 7. INDICATIVES

One proposal about this suggests itself immediately. Since 'true of' goes with a *counterfactual* conditional, 'true if' perhaps goes with the corresponding *indicative* conditional. 'S is true if *w*' says that *if w in fact obtains (evidence to the contrary notwithstanding), then S.*[5]

The proposal is intriguing because it offers to link two deep distinctions: metaphysical versus conceptual necessity, on the one hand, and subjunctive versus indicative conditionality, on the other. The reason it is only metaphysically necessary that Hesperus = Phosphorus is that there are worlds *w* such that, although Hesperus would have been Phosphorus *had w* obtained, it is not Phosphorus if *w does* obtain.

Do the two conditionals really 'predict' the two types of necessity? Before attempting to decide this, we need to remember how we got here. It was important for metaphysical necessity to keep what-is-said fixed as we evaluate S at *w*. Subjunctives are valued because they in effect do this, without dragging us into controversies about what is in fact said. It is *not* important to conceptual necessity to keep what-is-said fixed; indeed, we are willing and eager that it should change in certain respects under the impact of this or that counteractual hypothesis. (For example, we are eager for 'Hesperus = Phosphorus' to take on a content having to do with Venus and Mars.) Crucially, though, we do *not* want S's meaning to be changeable in *all* respects. (We don't want 'Hesperus = Phosphorus' to acquire a content having to do with nonidentity.) Indicatives are attractive because they seem to deliver an appropriate measure of meaning-fixation, just as subjunctives did on the metaphysical side.

Indicatives *appear* to deliver an appropriate measure of meaning-fixation. But when you look a little closer, the appearance fades. Indicatives don't in fact deliver *anything* in the way of meaning-fixation. The meaning of S as it occurs in the consequent of an indicative conditional can be changed all you want by

[5] Chalmers (2000).

putting the right kind of misinformation into the antecedent. Example: If 'tail' had meant wing, horses would still have had tails. But suppose that 'tail' *does* mean wing; it has meant wing all along, not only in others' mouths but also our own; a brain glitch (or demon) leads us systematically astray when we reflect on the meaning of that particular word. Then, it seems clear, horses do not have tails. If 'tail' as a matter of fact means wing, then to say that horses have tails is to say that they have wings. Horses do not have wings. So if 'tail' means wing, then FN:6 horses do not have tails.[6]

You may say: why should it be a problem if there are counteractual worlds at which horses lack tails? That is not the problem. The problem is that there are worlds where horses lack tails *not for anatomical reasons but on account of 'tail' ~~not~~* ‸wing *meaning‸ ~~tail~~*. If horses can lose their tails that easily, then take any S you like, it is true in some counteractual worlds and false in others. It is true in worlds where S means that X, and X is the case, and false in worlds where S means Y, and Y is not the case. This spells disaster for the indicative approach to conceptual possibility. It should not make 'Hesperus ≠ Hesperus' conceptually possible that there are worlds where people use '≠' to express identity.

## 8. NARROW CONTENT

The indicative is not the conditional we want. But it is close. We want a conditional A → C that is *like* the indicative except in one crucial respect: C is protected from a certain sort of meaning shift brought on by A.

An example of the 'good' or 'permitted' sort of meaning shift is the kind exhibited by 'Hesperus ≠ Phosphorus' on the supposition that Phosphorus-appearances are caused by Mars. An example of the 'bad' sort of meaning shift is that exhibited by 'Phosphorus ≠ Phosphorus' on the supposition that '≠' expresses identity.

It may seem that the answer is staring us in the face. The 'bad' kind of meaning shift is the kind that mucks with *S's narrow content*. Our conditional should be such that S's narrow content is the same when we condition on $w$ as when we don't. (The indicative is wrong because the narrow content of 'horses have tails' is one thing if 'tail' means wing, another thing if it doesn't.) Calling the

[6] Indicative conditionals are conditionals with antecedent and consequent in the indicative mood. Philosophers have proposed various theories of these conditionals. One, defended by Grice (1989), is that they are 'material', or truth-table, conditionals. Another, defended by Adams (1975), is that they are probability conditionals. Chalmers in recent work declares a preference for the material conditional, regardless of its relation, if any, to the indicative. (He requires the material conditional to hold a priori.) The objection in the text applies regardless. However the indicative is interpreted, A's a priori entailing C suffices for the apriority of 'if A then C'. The conditional 'if horses are wingless and "tail" means wing, then horses do not have tails' has A a priori entailing C, so the conditional is a priori.

actual narrow content NC, attention is to be restricted to worlds such that *w* obtains → S (still) means NC.

But, although helpful as an intuitive constraint, this doesn't solve our problem. This is partly because one doesn't know what the narrow content in fact is; NC has been pulled out of a hat. Second, though, to appeal to narrow content in this context gets things the wrong way around. The reason for being interested in 'S is true if *w*' was to get a better handle on conceptual necessity. But, as noted above, conceptual necessity and narrow content are two sides of the same coin. The idea is to explain narrow content using →, not → using narrow content.

## 9. TURNING OUT

Our problem now is similar to one faced earlier in connection with metaphysical necessity. It seemed that an account of true of would have to appeal to the notion of what is said. That would be unfortunate, because it would reverse the intended order of explanation. The what-is-said of an utterance (its broad content, nearly enough) is given by the worlds of which it is true. The special case in which S's broad content takes in *all* worlds is what is otherwise known as metaphysical necessity. That is why we don't want to use broad content to explain true of. Our current worry is the same, except that it concerns true if rather than true of, and narrow content rather than broad.

insert
hyphen

How did we deal with that earlier problem? By calling in the subjunctive. We said that S is true of *w* iff it would have been that S, had *w* obtained. The claim was that this construction *automatically* targets the agreement or lack thereof between *w* and S's broad content. Can a construction be found that automatically targets the agreement or lack thereof between *w* and S's *narrow* content, as the subjunctive does for broad content?

One that comes pretty close occurs in *Naming and Necessity* itself. Kripke notes that we're at first inclined to think that Hesperus and Phosphorus (although in fact identical) could have been distinct. Then we learn about metaphysical versus other types of necessity, and we lose the inclination; Hesperus and Phosphorus could not have been distinct. Even now, though, apprised of the metaphysical facts, we are still inclined to think that it *could have turned out* that Hesperus was distinct from Phosphorus.

It is this phrase 'could have turned out' that I want to focus on. Kripke is right to represent us as still inclined to think that it could have turned out that Hesperus was distinct from Phosphorus, even after we have taken on board that it could not have *been* that Hesperus was distinct from Phosphorus. The inclination persists even among practicing modal metaphysicians (who ought to know better, if there is better to know). This suggests that 'could have turned out' is special in ways we should try to understand.

It suggests it to me, anyway. Kripke apparently does not agree. He maintains that the second inclination is just as mistaken as the first. Not only could it not have been, it could not even have turned out that Hesperus was distinct from Phosphorus. This is only to be expected if ‘it could have turned out that S’ means, as Kripke hints it does mean, ‘it could have been that: S and we believed that S and with justification’. This interpretation, however, leaves it a mystery why the second inclination outlasts the first—why we persist in thinking that it could have turned out that Hesperus wasn’t Phosphorus even after giving up on the idea that Hesperus could have been other than Phosphorus.

I propose that the persisting thought is correct. Kripke to the contrary, it could indeed have turned out that Hesperus wasn’t Phosphorus. That is what *would* have turned out had it turned out that Phosphorus-appearances were appearances of Mars. It could *not*, however, have turned out that Phosphorus $\neq$ Phosphorus, even granting that ‘$\neq$’ could have turned out to express identity. That is a way for it to turn out that that ‘Phosphorus $\neq$ Phosphorus’ is true, not a way for it to turn out that Phosphorus $\neq$ Phosphorus.[7]

## 10. CONCEPTUAL POSSIBILITY

*It would have turned out that C, had it turned out that A* shares features with both the indicative conditional and the subjunctive. It resembles the indicative in making play not with counterfactual worlds, but with suppositions about *our* world. It resembles the subjunctive in that the consequent C is protected from a certain kind of semantic influence on the part of A. The way C (narrowly) represents things as being is left untouched by ‘had it turned out that A’. The role that the antecedent plays is all on the side of whether things are, on the hypothesis that A, the way that C (in actual fact, given that the hypothesis is false) narrowly represents them as being.

If ‘tail’ means wing, we said, then horses lack tails. → is supposed to be different in this respect. It should not be that *w (in which ‘tail’ means wing) obtains → horses lack tails*. That is the result we get if → is a ‘would have turned out’ conditional. For it is *not* the case that horses would have turned out to lack tails, had it turned out that ‘tail’ meant wing. It is not for linguistic reasons that

[7] Chalmers employs similar wording when he introduces primary intensions: ‘there are two quite distinct patterns of dependence of the referent of a content on the state of the world. First, there is the dependence by which reference is fixed in the actual world, depending on *how the world turns out: if it turns out* one way, a concept will pick out one thing, but *if it turns out* another way, the concept will pick out something else’ (1996: 57; italics added). I applaud the use of ‘turns out’, but I think the mood should be subjunctive—if it had turned out—rather than indicative—if it does turn out. If it turns out that ‘tail’ means wing, then horses lack tails. But that ‘tail’ means something different in *w* should be irrelevant to the question of whether *w*’s horses have tails. Otherwise conceptual necessity is trivialized. See also Jackson (1994, 1998).

horses have tails; so they are not deprived of their tails by the linguistic facts turning out differently.

One can come at → from the other direction. If Phosphorus-appearances had been due to Mars, Phosphorus would still have been Hesperus. → is supposed to be different in this respect too. We want there to be worlds *w* such that *w obtains* → *Hesperus* ≠ *Phosphorus*. That cannot happen unless the broad content of 'Hesperus ≠ Phosphorus' can be changed by conditioning it on the hypothesis that *w* obtains. Here too, 'would have turned out' delivers the goods. Had it turned out that Phosphorus-appearances were due to Mars and Hesperus-appearances (still) to Venus, it would have turned out that Hesperus ≠ Phosphorus.

What these examples suggest is that 'would have turned out' conditionals exhibit just the right combination of (i) openness to shifts in broad content, ‸and (ii) intolerance of shifts in narrow content. I therefore propose *it would have turned out that C, had it turned out that A* as the proper interpretation of A → C. And I make a hypothesis:

(M) It is metaphysically possible that S iff some world *w* is such that it would have been that S, had *w* obtained.

(C) It is conceptually possible that S iff some world *w* is such that it would have turned out that S, had *w* turned out to be actual.

More simply, S is *metaphysically* possible iff it could have *been* that S, and *conceptually* possible iff it could have *turned out* that S.

## 11. ANALYTICITY AND APRIORITY

A priori truths are truths that can be known not on the basis of empirical evidence. How well that accords with the Kripkean notion of apriority depends on one's theory of justification. There is a danger, though, of its according very badly.

One theory says that all spontaneously arising beliefs start out justified. They can lose that status only if evidence arises against them. Suppose that this view is correct, and suppose that, on pulling the curtains open, I spontaneously come to think that the sun is shining. (I don't infer that it is shining from premises about how things perceptually appear to me.) Then I know that the sun is shining, and not on the basis of empirical evidence. And yet it certainly isn't a priori, as Kripke uses the term, that the sun is shining.

Another theory has it that our most 'basic' beliefs lack empirical justification, because they are epistemically prior to anything that might be said in their support. So, the belief that nature is uniform lacks empirical backing. If we know that nature is uniform, and let's assume we do, the knowledge is not empirical. But it isn't a priori in Kripke's sense that nature is uniform.

Apriority, then, is not *any* old kind of not-empirically-based knowability, as judged by any old theory of justification. That would let far too much in. A (very familiar) objection from the other side helps us to clarify matters. If experience cannot be appealed to at all, then shouldn't it be enough to stop S from being a priori if it is through experience that we *understand* S? The answer to this is that our interest is in how S is *justified*, our understanding taken for granted.

If that is the one and only concession made, then we wind up with a roughly Kripkean notion of apriority. S is a priori iff it is *knowable just on the basis of one's understanding of S*. Or, better, it's a priori *for me* iff *I* can know it just on the basis of my understanding of S. This is why the originator of a name is apt to know more a priori than someone picking the name up in conversation. The mental state by which Leverrier understands 'Neptune' tells him that Neptune, if there is such a thing, accounts for the perturbations in the orbit of Uranus. The mental state by which others understand 'Neptune' is liable to be much less informative about Neptune's astronomical properties.

Apriority is knowability on the basis of understanding. Understanding is, one assumes, knowing the meaning. But what meaning?

Perhaps understanding is knowing meaning 'in the strict sense': the sense that ignores reference-fixing descriptions. But Kripke calls it a priori that Hesperus = Hesperus, and a posteriori that Hesperus = Phosphorus, though the strict meanings are the same. More likely, then, it is knowledge of meaning in the *loose* sense that makes for understanding. The closest thing to loose meaning in our framework is narrow content. So it does not do *too* much violence to Kripke's intentions to say that S is a priori iff one can know that it is true just on the basis of one's grasp of its narrow content.

Kripke calls S analytic iff 'it's true in virtue of meanings in the strict sense'. This definition has to be treated with some care, since the strict meaning of 'Hesperus = Phosphorus' is a singular proposition of the form $x = x$, and Kripke does not want 'Hesperus = Phosphorus' to come out analytic. (It is not a priori, and Kripke thinks that analytic truths are a priori.) Then what is his intent in speaking of 'meanings in the strict sense'? He cannot have been trying to *include* statements ('Hesperus = Phosphorus') that are true in virtue of strict meaning as opposed to loose. He must have been trying to *exclude* statements ('Hesperus is visible at night') that are true in virtue of loose meaning as opposed to strict. This is, in effect, to limit analyticity to 'Fregean' sentences: sentences to which the loose/strict distinction does not apply. S is analytic iff it is true in virtue of its Fregean meaning, that being the only meaning it has.

Now, though, one wants to know: why should it stop S from being analytic if *in addition* to its truth-guaranteeing Fregean meaning, it has a (possibly not truth-guaranteeing) Kripkean meaning? Or, to put it in narrow/broad terms, if S has a truth-guaranteeing narrow content, why isn't that enough to make it analytic, quite regardless of whether it has a broad content in addition?

True-blue Kripkeans will reply that narrow content is not (except *per accidens*, when it agrees with broad) part of *meaning*. Narrow content is *metasemantical*, not semantical.

But this, one may feel, is just terminological fussiness.[8] Even Kripke considers it a *kind* of meaning—meaning in the loose sense—and he says explicitly that some might want to define analyticity as truth in virtue of *that*. So, it does not do *too* much violence to Kripke's intentions to let analyticity be truth in virtue of narrow content. (This fits with our account of Kripkean apriority as knowability in virtue of grasp of narrow content.)

Now, finally, we can ask the question that matters: Is conceptual necessity a kind of apriority, or a kind of analyticity, or both?

I do not think there can be much doubt that it is a kind of analyticity. A conceptually necessary sentence is one true in all counteractual worlds. These worlds comprise what Chalmers calls the sentence's *primary intension*, and primary intension is his candidate for the role of narrow content. So, a conceptually necessary sentence is one whose narrow content is such that, no matter which world is actual, it comes out true. Truth guaranteed by narrow content is analytic truth.

Is conceptual necessity *also* perhaps a kind of apriority? As just discussed, the narrow content of a conceptually necessary sentence is such as to guarantee its truth. Does it follow that someone *grasping* the content is thereby in a position to *see* that S is true?

That depends on what is involved in grasping a content (let 'narrow' be understood). S's content is, roughly, a bunch of conditionals of the form: it would (or wouldn't) have turned out that S, had *w* turned out to be actual. Someone who grasps the content ~~is in a position to know the~~ knows how to evaluate the conditionals. ~~So~~ if S is conceptually necessary, then she is in a position to see, for each *w*, that had *w* turned out to be actual, it would have turned out that S. Doesn't this show that she can determine a priori that S?

No; in fact, we are still miles from that conclusion. Let it be that the speaker ~~knows~~ can tell, for each *w*, that *w* obtains → S. It is wide open so far whether ~~this~~ the resulting knowledge is a priori. Someone who grasps S's meaning is in a position ~~to come~~ to know the conditionals *somehow or other*. A priori or a posteriori is ~~an open~~ a further question.

You might think that the knowledge *has* to be a priori. If grasping S's content gives me knowledge of the conditionals, then I know the conditionals based on my grasp of S's content. Knowledge based on grasp of content is a priori knowledge.

This is ~~entirely~~ unconvincing, however. Grasping S's content 'gives me' knowledge of the conditionals only in the sense of putting me in a position ~~to come~~ to ~~know~~ tell on inspection that they are true.

[8] I myself feel it is more than that, but this is the charge made by the narrow content enthusiast whose part I am playing.

~~them~~; my advantage over non-graspers is that I have (and they lack) 'what it takes' to ~~know.~~ make this judgment. That is roughly to say that understanding S is *necessary* if one wants to know whether S if *w*, or the most important necessary condition, or the only necessary condition worth worrying ~~one has to worry~~ about. Apriority requires that understanding be *sufficient*. I have granted that understanding suffices for being in a position to *work out* whether S if *w*. If the working out involves experience, though, then the knowledge will not be a priori.

## 12. PEEKING

I said that our understanding of S might not be enough to go on, when it comes to working out whether S holds in a world *w*. The 'official story' about evaluation at counteractual worlds strongly denies this. But the possibility has a way of sneaking in uninvited. Here is Chalmers:

> [A]s an in-principle point, there are various ways to see that someone (a superbeing?) armed only with the microphysical facts and the concepts involved could infer the high-level facts. The simplest way is to note that in principle one could build a big mental simulation of the world and watch it in one's mind's eye, so to speak. (1996:76)

Say that this is right; I am able to build a mental model of *w*, and judge whether S is true in *w* by viewing the model with my mind's eye. The question is whether viewing a model of *w* and asking myself 'how it looks' S-wise is a way of coming to know S's truth-value in *w* a priori.

Here is a reason to think not. Asking yourself how something strikes you is using yourself as a measuring device. Information acquired by use of an external measuring device is a posteriori on anybody's account. Information acquired by use of an internal one seems no different. What matters is that an experiment is done, the outcome of which decides your response.

It might be argued that mental experimentation *is* different. Knowledge gained from it is acquired within the privacy of one's own mind. You determine that S without appealing at any point to information about the outside world. Shouldn't that be enough to make the knowledge a priori?

No, for you determine that you have a headache the same way. Knowledge of headaches is certainly not a priori. The modal rationalist in particular should agree, for my headache, if a priori, would be a counter-example to the proposed equation between apriority and truth in all counteractual worlds. 'I have a headache' fails in some counteractual worlds. A priori truths are supposed to hold everywhere.

Some internally acquired knowledge presumably is a priori. If you think up a counter-example to argument form F in your head, then you know a priori that F is invalid. What distinguishes this sort of case, where you do know a priori, from the case of looking at a mental model with the mind's eye?

Two things. First, when you conjure up an image of *w*, you are *simulating* the activity of really looking at it. Simulated looking is not a distinct process, but the usual process run 'off-line'. Knowledge gained by internal looking is not a priori because it is acquired through the exercise of a perceptual faculty rather than a cognitive one.

Second, some imagined reactions are a better guide to real reactions than FN:9 others.[9] Imagined shape reactions are a good guide, you say, and you are probably right. But it is hard to see how the knowledge that they are a good guide could be a priori. If the mind's eye sees one sort of property roughly as real eyes do, while its take on another sort of property tends to be off the mark, that is an empirical fact known on the basis of empirical evidence. I know not to trust my imagined reactions to arrangements ~~of~~ furniture, because they have TR often been wrong; now that I see the wardrobe in the room, ~~I realize it is far too big~~ it crowds up closer to the sofa than I had supposed. It is only because they have generally been right that I am entitled to trust my imagined judgments of shape.

The temptation to think of simulation as a source of a priori knowledge is due in part to there not being much that we are able to simulate. There might be beings who, given only the microphysical blueprint of, say, an exotic fruit, are able to imagine its color in much the way that we are able to imagine its shape. They come to know that rambutans are red, without ever laying eyes on one. I take it that no one would consider the knowledge to be a priori. These beings did not deduce the color from microphysics. Information was also needed about how that microphysics appears to human eyes. They obtained this information experimentally, by simulating an encounter with a rambutan, and using it to predict the outcome of a real encounter.

Suppose that we had been able to simulate reactions in other modalities. Suppose we could determine the taste and smell of a microphysically given item with the mind's tongue and nose. Would that make it an a priori matter how rambutans [insert chemical description here] tasted? No. How a thing tastes is an empirical question. One does not feel that it escapes being a priori only because of a contingent incompleteness in our nature. It would still have been an empirical matter how rambutans tasted, even if God had been more generous in the mind's sense-organ department.

These claims might be accepted but shrugged off as irrelevant. It doesn't matter if self-experimental knowledge is a posteriori, for any suggestion of self-experimentation was inadvertent. 'I looked at *w* and saw it to contain so-and-so's' is only a colorful description of something far more innocent: intellectually

[9] Stepping into the lake, you say, 'It's colder than I thought.' The earlier thought might have been a real judgment based on partial information (it's August, lots of people are swimming), but it might also have been a simulated judgment based on full information about the water's kinetic properties. You imagine yourself stepping into water with those properties, and it seems to feel warmer than water like that really does feel. (Most of us do something like this with temperature properties; 80 degree water is surprisingly cold.)

contemplating a world *description* and *thinking* my way to a conclusion about whether there are so-and-so's in *w*.

That is fair enough, on one condition. Self-experimentation had better not be *needed* to work out whether S holds in *w*. It had better be that one can reason from a microphysical description of *w* to a conclusion about whether or not S. *No peeking*. I assume that Chalmers would agree; for, if peeking is allowed, the inference from 'S holds in all candidates for actuality' to 'it is a priori that S' clearly does not go through. This inference is crucial to the view that Chalmers calls 'modal rationalism'.

Given how much hangs on our ability to evaluate S without peeking, one might have expected a show of vigilance on this score. If we are playing 'pin the tail on the donkey', you watch me like a hawk. You know how hard I find it to ignore information right in front of my nose. The same should apply when the game is 'decide the truth-value of S'. If it is difficult to infer S (¬S) from microphysics, I will be tempted to switch to sensory imagining. Knowing this, you will take pains that my mind's eye is completely shut, or completely covered by my mind's blindfold.

The need for vigilance is never mentioned, as far as I know, in the modal rationalist literature. Here is how the passage quoted above continues:

> Say that a man is carrying an umbrella. From the associated microphysical facts, one could straightforwardly infer facts about the distribution and chemical composition of mass in the man's vicinity, giving a high-level structural description of the area. One could determine the existence of a male fleshy biped straightforwardly enough. . . . It would be clear that he was carrying some device that was preventing drops of water, otherwise prevalent in the neighborhood, from hitting him. Doubts that this device is really an umbrella could be assuaged by noting from its physical structure that it can fold and unfold; from its history that it was hanging on a stand that morning, and was originally made in a factory with others of a similar kind. (Chalmers 1996: 76)

When I try to 'determine' these higher-level facts, I find myself relying on visual imagining at every turn. 'Keep your mind's eye scrunched tight,' I am told. I can try, but then the higher-level facts go all mysterious. The feeling intensifies when I read how 'doubts that the device is an umbrella can be assuaged'. Never mind how they are assuaged; I do not see how the umbrella idea came up in the first place.

I realize how it's supposed to go. I start with objective, geometrical information. A chain of a priori inferences leads to 'it's shaped like an umbrella'. That conclusion combines with a host of others to establish its umbrella-hood beyond any doubt. Visualization is barred, so I have no idea of how the object looks. (Eventually it may strike me that since the object is an umbrella, it probably looks like one.)

Is this possible? It helps to look at a simpler case. I am to infer a plate's shape (it's in fact round) from premises about the arrangement of its microphysical parts. The premises might take various forms, but assume for definiteness that the arrangement is specified in analytic geometry terms. I am told that the object's teeny-tiny parts occupy the points (*x, y*) such that $x^2 + y^2 < 63$. (The plate is two-dimensional, no pun intended.) If I am to reason from this to the object's shape, I must know, implicitly at least, conditionals like the following:

if R is circumscribed by the points (*x, y*) such that $x^2 + y^2 = 63$, then R is round;

if R is circumscribed by the points (*x, y*) such that $x^4 + y^4 = 63$, then R is not round.

I should know many, many conditionals of this nature, one per lower-level implementation of roundness, and, I suppose, one per implementation of nonroundness. And, most important of all, I should know the conditionals a priori, just through my grasp of the relevant English words.

But, it isn't clear that I *do* know many conditionals like these. (I am tempted to say that it's clear that I don't.) And the few that I do know, I don't seem to know a priori. It wasn't learning the meaning of 'round' that taught me the formula for circles. I worked it out empirically by graphing the formula, *looking* at the figure I had just drawn, and then *reflecting* on how I was inclined to describe the figure. (I take it that no one has their first encounter with roundness in a geometry class.)

I do not say that the above shows that you *have* to peek. There may be other ways of proceeding that haven't occurred to me. All I mean to be claiming for now is that 'one can find the umbrellas in *w* without peeking, just by virtue of one's competence with the word' is a *substantive and surprising thesis*. Theses like this need to be argued for, and no argument has been given. A priori entailment has been presented as what you would expect, unless a skeptical philosopher had got to you first.

## 13. RECOGNITIONAL PREDICATES

Now let me move on to urging in a positive way that there is only so much we can judge with the mind's eye averted. I think that one *can't* always tell, just by drawing inferences from a world description, whether the world is one where it turns out that S. If that is right, then the method that Chalmers didn't really mean to be advocating, and that figures only inadvertently in his narrative, is in some cases the only possible method. This will be argued for *observational* predicates (starting with the subtype *recognitional*), then *evaluative* predicates, then, finally, *theoretical* predicates.

What marks a predicate P as observational? The usual answer is that understanding P involves an ability to work out its extension in *perceptually* (as opposed to intellectually) presented scenarios. To determine P's extension in a world, I have to cast my gaze over that world—at candidate Ps in particular—and see how it strikes me.

Nothing has been said about the kind of appearance that marks a thing as P. Sometimes *x* is judged P because our experience of *x* has a quality Q notionally independent of P. So, *x* is tantalizing if, roughly, the experience of it makes one want to get closer and know more. Other times the experience that marks *x* as P is the experience of it as being precisely P. One judges *x* to be P because P is how it looks or feels or sounds. . . . This is what I am calling a *recognitional* predicate.

Examples are bound to be controversial, so let me just follow Kripke. Kripke says that 'the reference of "yellowness" is fixed by the description "that (manifest) property of objects which causes them, under normal circumstances, to be seen as yellow" ' (1980: 140 n. 71). We understand by yellowness whatever property it is that makes objects look yellow, or gives rise to the sensation of yellow. The predicate 'yellow' is recognitional on this view, since the yellow objects are picked out by their property of looking yellow.

Suppose Kripke is right about our understanding of 'yellow'. What are the implications for the way yellow things are identified in a candidate *w* for actuality? It's clear that *x* has to look yellow to be counted into the predicate's extension. But to whom? Perhaps it needs to look yellow to the *w*-folks, including one's counteractual self. If it is counteractual Steve's reactions that matter, then I don't need to experience *x* myself to determine if *x* is (in *w*) yellow. I can infer *x*'s color a priori from what the relevant world description says about the experiences Steve has when experiencing *x*.

But what *does* the world description say about counteractual Steve's experiences? Suppose, first, that it describes them in intrinsic phenomenological terms; banana-caused visual experiences are said to have intrinsic phenomenological property K. This doesn't yet tell me whether bananas are yellow, for I don't know that K is the phenomenology appropriate to experiences of *yellow*. I can't determine that without giving myself a K-type experience and checking its content: do I feel myself to be having an experience of yellow or of green?

Suppose, on the other hand, that counteractual Steve's experiences are described intentionally, as 'experiences of yellow', 'yellow' being the predicate whose corresponding property we are trying to identify. Then we would seem to be caught in a circle. The referent of a compound expression depends on the referents of its parts. So any intelligence we might have about what it is to be an 'experience of yellow' must come from prior information about (among other things) what it is to be 'yellow'. But then the referent of each of these two phrases depends on that of the other.[10]

Kripke must have been aware of this problem. He notes that '[s]ome philosophers have argued that such terms as "sensation of yellow", "sensation of heat", . . . and the like, could not be in the language unless they were

[10] One option is to say that yellowness and the sensation of it are identified together by means of a gigantic Ramsey-type theoretical definition. ~~This is filed under the heading 'just a pipe dream until somebody supplies details'.~~ Let me not comment on this here, except to note a possible threat to the conceivability of zombies.

identifiable in terms of external observable phenomena, such as heat, yellowness'. And he says that 'this question is independent of any view argued in the text' (1980: 140 n. 71). Kripke doesn't mind, in other words, if one can't identify sensations of yellowness until one has identified the property they are sensations of. How, if that is so, can we hope to identify yellowness by way of sensations of yellow?

Here is what I think Kripke would say. Yellowness is identified not by a *condition* on experience ('such as to give rise to sensations of yellow'), but by the experience itself. The objects I call yellow are the ones that *look* yellow. If the yellow things were identified by an experiential *condition*, then we would face the problem of working out which experiences were of the indicated type. But that is not our situation. Far from being something in need of discovery, the rom FN:11 experience of yellow is *part of the discovery process*.[11] I don't have to *identify* my yellow-experiences in order ~~to learn by their~~ exercise, any more than I have to FN:12 identify my eyes in order to learn by use of them.[12]

There is a second reason why Kripke would (should) not take 'yellow' to have its reference fixed by an experience-implicating description. What will the description say about proper viewing conditions?

This is a problem that he himself raises for a related view: the view that 'yellow' is *defined* as 'tends to produce such and such visual impressions'. Tends to produce them under what circumstances, Kripke asks? Any answer will be unsatisfactory: 'the specification of the circumstances C either circularly involves yellowness or . . . makes the alleged definition into a scientific discovery rather than a synonymy' (1980: 140 n. 71). If C-type circumstances are circumstances where we are not deceived as to yellowness, then (while it may be analytic that *x* is yellow iff it looks yellow in C-type circumstances) the definition uses 'yellow', so cannot explain its meaning. If C-type circumstances are ones where (say) the ^ type light is of such-and-such ^ ~~a composition~~, no one is suffering from jaundice, the object is not a Benham's disk rotating at such-and-such a rate, etc., then, while it may be true that *x* is yellow iff it looks yellow in C-type circumstances, it is not definitionally true, but empirically so.

[11] The issue here is much like the one raised by Putnam's 'descriptivist' interpretation of the causal theory of reference. Putnam suggests that words have their reference fixed by a causal *condition*. One finds the referent by looking for whatever stands in the right causal relation to speech. This makes for circularity problems, since one needs to know which relation causation is to work out what 'causation' denotes. From here it is a short step to radical indeterminacy of reference. The almost universal response was that reference is fixed *causally*, not *descriptively* by a condition alluding *inter alia* to causal relations. Kripke as I am reading him says something similar: reference is fixed experientially, not descriptively by a condition alluding *inter alia* to a certain sort of experience.

[12] I like what Colin McGinn says about perceptual concepts. Some think that 'When a concept is applied to a presented object that is always a further operation of the mind, superadded to the mere appearance of the object in perceptual consciousness. On my way of looking at it, concepts figure as *substitutes* for perceptual appearance— . . . they are needed for intentionality only when the object is not being perceived' (1999: 324).

If this is a good objection to the idea that ‘tends to . . . in circumstances C’ defines ‘yellow’, it would seem to be equally hard on Kripke’s own claim that ‘yellow’ has its reference fixed by that description. Either C-type circumstances are ones where we are not deceived as to yellowness, or they are ones where the light has such-and-such a composition, etc. If the first, then, while it may be a priori that *x* is yellow iff it looks yellow in C-type circumstances, the reference-fixer presupposes yellowness, and so cannot be used to identify it. If the second, then, while it may be true that *x* is yellow iff it looks yellow in C-type circumstances, it is not a priori true, as it would be if the description fixed ‘yellow’ ’s reference.

One can reply in the same way as before. What marks a thing *x* as yellow isn’t the *condition* ‘tends to produce . . . under circumstances C’. What marks *x* as yellow is that that is how it looks. Someone can of course ask, how do you know the perceptual circumstances (including the condition of the perceiver) are right? But we do not say to this person, ‘the present circumstances are of type C, and C defines rightness’. That would open us up to all the problems raised above. Our answer is, ‘Why shouldn’t they be right? What is it that leads you to suspect trouble?’ It may not be a priori that what looks yellow under conditions C is yellow, but it does seem to be a priori that what looks yellow is yellow assuming nothing funny is happening. And that is an assumption we are always entitled to, unless and until we run into specific objections.

I hope this makes clear how our grasp of a predicate can be *recognitional* rather than intellectual. I do not reason my way to the conclusion that something is yellow from premises about what looks yellow under which conditions. The belief arises spontaneously in me when I look at a thing. That *has* to be how it works, for I have in general no a priori reliable information about which viewing conditions are appropriate. The most that is a priori is that *these* conditions are appropriate, unless there is reason to think otherwise.

If P is a recognitional predicate, then I have an a priori entitlement to ‘*These* conditions are (funny business aside) such that what seems P is P’. This is an entitlement that, by its nature, does not travel well. It lapses when we move from the world that really is actual to worlds only treated as actual for semantic evaluation purposes. For in lots of those worlds, we find (what from our actual–actual perspective is) funny business.

A few special cases aside, what looks yellow, is yellow. But things could have turned out so that whipped cream looked yellow—say, because a jaundice-like staining was characteristic of healthy eyes rather than diseased ones. This would not bother the people we turned out to be (they think *our* eyes are problematic), but it does bother us as we are. Whipped cream is white, and so whoever sees it as yellow is to that extent getting it wrong.

This has two semi-surprising consequences, which for now I’ll just state without argument.

(1) Something known a priori need not hold in all counteractual worlds. It is a priori that funny business aside, what looks yellow, is yellow. But had our eyes turned out as described, objects would have turned out to look yellow that were in fact white. There is no mistake here, nor is anyone misled. Whipped cream is indeed what *they* mean by the word. It is just not what we mean by it, that is, it is not yellow.

(2) Something holding in all counteractual worlds might be knowable only a posteriori. Let F be a complete intrinsic characterization of some white chalk.[13] Could an F have turned out to be other than white? The chalk could have turned out yellow-*looking*, as already discussed. To have turned out *yellow*, however, it would have needed different (non-F-ish) intrinsic properties. So although it is a posteriori what color Fs in fact are, their color is conceptually necessary in the sense that it could not have turned out any different.

FN:13

## 14. OBSERVATIONAL PREDICATES

Everyone knows what it is for a figure to be *oval*. It is not hard to distinguish ovals from polygons, figure-eights, and so on. It is not even all that hard to distinguish ovals from otherwise ovular figures that are too skinny or too fat to count. To a first approximation, a figure is oval if it has the proportions of an egg, or a two-dimensional projection of an egg. I take it that few of us know in an intellectual way what those proportions are. What marks a figure as oval is not its satisfaction of some objective geometric condition, but the fact that when you look at it, it looks egg-shaped.[14]

FN:14

Because our grasp of *oval* is constituted in part by how its instances look, one might be tempted to group it with 'response-dependent' concepts like ~~*ticklish*~~ *irksome* or *tantalizing*. That would be a mistake. There are several respects in which *oval* is quite *un*like ~~*ticklish*~~ *irksome*, which, once pointed out, make the label 'response-*enabled*' seem much more appropriate. Another term I shall use is 'grokking concept'. (I apply these labels to concepts, but, depending on one's other commitments, they could speak more to how the concept is grasped.)

*Constitution*: Why are ~~ticklish~~ irksome things ~~ticklish~~ irksome? That might mean 'what is the evidence that they are ~~ticklish~~ irksome?' If so, the answer is that we respond to them in a certain way; they ~~tickle~~ irk us. If it means 'what qualifies them to be so regarded?', the answer has again to do with our responses. So far there is no contrast with *oval*. But suppose we now ask, 'in what does their ~~ticklishness~~ irksomeness consist?' ~~Eliciting~~ Causing

[13] More may have to be packed into F, such as prevailing natural laws.

[14] I say *looks egg-shaped* and not *looks oval* because I want 'oval' to be an example of an observational predicate that is not recognitional.

or tending to ~~elicit~~ cause a certain reaction in us is ‘what it is’ to be ~~ticklish~~ irksome. To be oval, though, is simply to have a certain shape.

*Tracking*: Our responses do not track the extension of ~~‘ticklish’~~ 'irksome'; they dictate it. It makes no sense to suggest that our tendency to be ~~tickled~~ irked by various things might not have been, or might have turned out not to be, a good guide to what is really ~~ticklish~~ irksome. It is different with ‘oval’. Our responses give us *access* to the extension of ‘oval’, but they do not dictate the extension.

*Motivation*: Why are the ~~ticklish~~ irksome things picked out experientially? There is an in-principle reason for this: we want to classify as ~~ticklish~~ irksome whatever is experienced in a certain way. Why are the oval things picked out experientially? There is no in-principle reason, but only a practical one: we have no other way of roping in the intended shapes.

*Evaluation*: Externalities are the same in *w* as here, but our responses are different. Suppose that our world had turned out to be *w*. What would have turned out to be ~~ticklish~~ irksome? That which turned out to elicit the ~~tickle~~ irked response. What would have turned out to be oval? That which *does* elicit the oval response; that which *does* look egg-shaped. For dimes to have turned out oval, they would have had to turn out a different shape.

The ‘evaluation’ contrast is the one that matters, so let me dwell on it a little. Imagine someone who thinks that ‘oval’ applies to whatever strikes the locals as egg-shaped, in any *w* you like, considered as counteractual or counterfactual. This person has misunderstood the concept. If he were right about counterfactual worlds, then

dimes would have been oval, had they (although still round) looked egg-shaped.

This is plainly untrue. If he were right about counteractual worlds, then

dimes would have turned out to be oval, had they (although still round) turned out to look egg-shaped.

This is false, too. The only way to a thing’s ovality is through its shape; you can’t change the one except by changing the other. You can’t make something oval by tinkering only with our responses.

What can we say to our confused friend to straighten him out? ‘Oval’ stands for things like *that*, the kind that we *do* see as shaped like eggs. The concept uses our responses as a tool—a tool that, like most tools, stops working if it’s banged too far out of shape. The concept presupposes that our responses are what they are, and then leans on that presupposition in marking out the class of intended shapes. This is why its turning out that we saw dimes as egg-shaped would be a way for it to turn out (not that they were oval, but) that we were taking non-ovals for ovals.

A better analogy for our concept of oval is the concept expressed by ‘*that* shape’ when we say, pointing at a sculpture, that ‘that shape is eerily familiar’—or the one expressed by ‘*this* big’ in ‘a room has to be at least *this* big [gesturing at the

FN:15 surrounding walls] to hold all my furniture'.[15] The role of '*this* big' is not to pick out whatever old size one might turn out to be perceiving: *tiny* if one turned out to have been in a tiny room suffering an optical illusion. It is, rather, that one takes oneself to be perceiving a room of a certain size, and one has no way of identifying ~~knowing~~ the size other than via its perceptual appearance.

## 15. ANALYTICITY WITHOUT APRIORITY

First there are the response-*dependent* concepts: comfortable, irksome, ~~ticklish, aggravating~~, tantalizing, painful-to-behold. Then there are the response-*enabled* concepts: oval, aquiline, jagged, crunchy, smiley-faced. Response-enabled concepts have their own distinctive pattern of evaluation at counteractual worlds. If *oval* were response-dependent, then one could determine its extension in *w* by asking what the people there saw as egg-shaped.[16] FN:16 If it is response-*enabled*, then those counteractual responses are irrelevant. Ovality is to be judged not by *as-if* actual observers, but by *actual* actual observers. A thing in *w* is oval if it is of a shape that would strike *me* as egg-shaped were I (with my sensibilities undisturbed) given a chance to look at it.

This has consequences for what comes out analytic, or conceptually necessary. Consider a world *w* about which all I'm going to tell you is that it contains Figure 1. Is 'oval' true of this figure in *w* considered as actual? The answer is clear. All we need do to determine that it is oval is look at the figure, and note that it looks *like that*—the way that ovals are supposed to look.

Once again, I have not said anything about how observers in *w* see Figure 1. Maybe there are no observers in *w*, or maybe there are, but they do not think Figure 1 has the right sort of look. It doesn't matter, for we evaluate the figure with reference ~~respect~~ to our word 'oval', understood as we understand it. Our dispositions figure crucially in that understanding, so they are part of what we (imaginatively) bring to bear on the figure in *w*.

Now let's bring in our conditional →, the conditional used to define conceptual necessity. Is it or is it not the case that *w obtains* → *Figure 1 is oval?* Would Figure 1 have turned out still to be oval, had it turned out to be shaped as shown? You bet it would. Whether an as-if actual figure is oval is completely determined by its shape. Things could have turned out so that Figure 1 did not look egg-shaped: we could have wound up with greater powers of visual discrimination, and as a result been 'bothered' by departures from an egg's precise shape that, as we are, we find it easy to ignore.[17] FN:17 But Figure 1 would not in that

[15] Peacocke (1989).

[16] So-called rigidified response dependency is for our purposes a minor variant of the unrigidified kind.

[17] Suppose for argument's sake that phenomenological similarity goes with the number of jnd's—just noticeable differences—separating one figure from another.

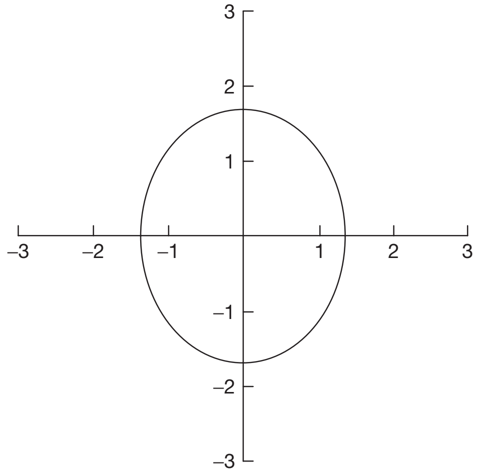

**Fig. 4.1.** Could this shape have turned out not to be oval?

case have turned out not to be oval. One wants to say, rather, that ovals would have turned out not to look the way they do look; ovals would have turned out to lack the feature by which things are recognized as ovals.[18]

Suppose we do some measurements and determine that Figure 1 is defined ~~(up to congruence)~~ by the equation $(x^2 + y^2)^2 - (x^2 - y^2) = 5$. Figures like that can be called *cassini*-shaped, or, for short, *cassinis.* (Giovanni Cassini (1625–1712) studied a class of figures of which this is one.[19]) 'Cassini-shaped' is an objective, third-personal predicate applying to all and only figures with the geometrical properties (that we all correctly take to be) exemplified by Figure 1.

Could things have turned out so that cassinis were not oval? If ovality in a world is purely a function of shape, then the answer is no. 'Cassinis are oval' is true in all worlds-taken-as-actual, which makes it (given our definition above) conceptually necessary.

But, of course, it is very far from a priori that cassinis are oval. To determine whether they are oval, you have to cast your eyes over (some of) them, and see how they look to you. There is no other way to do it. 'Cassinis are oval' is an analytic (conceptually necessary) truth that we are in no position to know a priori.

[18] I assume that the label 'oval' continues in the imagined case to be applied on the basis of egg-looking-ness. Does the fact that different things turn out to look that way make us (in that case) bad judges of ovality? Yes and no. Our counteractual responses are an excellent guide to what 'oval' would have turned out to mean. They are a bad guide, however, to what 'oval' does mean; they are a bad guide to what is in fact oval.

[19] The class of 'Cassinian ovals', although not all are really oval; indeed, not all are topologically connected.

## 16. OTHER INTENSIONS

If every world *w* is such that its cassinis are (to us) eggish-looking, then 'cassinis are oval' is analytic. Its meaning as encoded in our reactive dispositions guarantees its truth. But this is a kind of analyticity that we would not expect to make for apriority, because the route from understanding to extension and hence truth-value is inescapably observational.

To put it the other way around, one can't conclude from the fact that 'cassinis are oval' fails to be a priori that there is a counteractual world some of whose cassinis aren't oval. The premise you need for that is that 'cassinis are oval' is not *analytic*. But it *is* analytic. Given what the sentence means, it has got to be true.

Once again, the inference from (i) failure of apriority to (ii) a world that 'witnesses' the failure is crucial to modal rationalism. One might almost be forgiven for thinking that the main thing people *value* in the doctrine is its ability to deliver a counter-world. I assume, then, that modal rationalists would like, if possible, to *plug* the gap that seems to have opened up between analyticity (conceptual necessity) and apriority.

One approach harks back to the indicative account of truth in a counteractual world. For S to hold in counteractual *w* is, on that account, for it to be the case that if *w* obtains, then S. We rejected this account on the ground that it makes every sentence conceptually contingent. (If 'sibling' means parent, then sisters are *not* always siblings.) But, you may say, there is an obvious fix. It should be not merely true but a priori that if *w* obtains, then S. It is not a priori that if 'sibling' means parent, then sisters aren't always siblings. So a world where 'sibling' means parent is not on the new definition a world where the problematic sentence (some sisters are not siblings) holds.

Suppose we let S's *epistemic* intension be the set of worlds such that it's a priori that if *w* obtains, then S. And suppose that conceptual necessity is understood as necessity of the epistemic intension so defined. What happens to the argument above that conceptual necessity is a kind of analyticity but not a kind of apriority?

It might seem to fall apart. 'Cassinis are oval' may have a necessary primary intension, but its epistemic intension is contingent. (It is not generally a priori that if *w* obtains, then cassinis are oval; perhaps it is never a priori.) But then, if conceptual necessity goes with the epistemic intension, 'cassinis are oval' is not conceptually necessary. And so it no longer serves as a counter-example to the idea that whatever is conceptually necessary is a priori.

This assumes, however, that intensions built on a priori indicatives avoid the problems that were raised for intensions built on ordinary indicatives. Do they? What does seem clear is that the old examples no longer work. But this is for a correctable reason: namely, that sisters might, for all we know a priori, be one and all parents. It *is* a priori (let’s assume) that sisters (if there are any) are not numbers. And so it is a priori too that if ‘sibling’ means number, then sisters (if there are any) are not siblings.

I have said that if ‘sibling’ means number, then sisters aren’t siblings. Suppose that claim was based on empirical evidence. What would the evidence be? The only empirical fact in the neighborhood would seem to be this: ‘sibling’ does not in fact mean number. Call that the actual-meaning fact. Does it form part of my justification for believing that if ‘sibling’ means number, then sisters aren’t siblings?

If it does form part of my justification, then should I *forget* ‘sibling’ ’s meaning, or come to hold an erroneous view of it, my justification would be compromised. Say I fall under the impression that ‘sibling’ does mean number. Have I now lost my grounds for thinking that *if* it means number, then sisters aren’t siblings? Surely not. My reasons for thinking that *if ‘sibling’ means number, sisters are not siblings*, are just the same whether I believe the antecedent or not. How could forgetting what ‘sibling’ *does* mean compromise my ability to make inferences from a certain *hypothesis* about its meaning?[20]

FN:20

Where does this leave us? If my belief in the conditional is a priori, then there is a world that is not in the epistemic intension of ‘sisters have siblings’. The same argument shows that *no* statement S, however a priori in appearance, has a necessary epistemic intension. I conclude that the a priori indicative strategy is no great advance over the plain indicative strategy. Both have the same basic problem: they make all intensions contingent, and so drain the class of conceptual necessities of all its members.

It might be held that the problem is not with the aprioritizing as such, but with the type of conditional aprioritized. A second option is to call S true in *w*-considered-as-actual iff it holds a priori that (*w* obtains → S)—it holds a priori that it would have turned out that S, had *w* turned out to be actual. The intensions that result can be called *priory* intensions. If conceptual necessity is necessity of the priory intension, maybe the inference to a counter-world can be saved. Certainly it isn’t refuted by the cassini example; for although ‘cassinis are oval’ has a necessary primary intension, its priory intension is *not* necessary. (You

[20] This is intuitive on its face, but it can also be argued for in the following way. It’s agreed that I know that *if ‘sibling’ means number, then sisters aren’t siblings*. The question is whether my justification is a posteriori, because based on the actual-meaning fact. If it is, then I lack the knowledge we’ve just agreed I have. Here is why. You are not said to know that *if A, then B* unless you know something from which B can be inferred, should it be discovered that A. Your justification for the conditional should therefore be ‘robust’ with respect to A: it should be such as to *stay in place* should one come to believe that A. (See Jackson 1979.) Your justification would not be robust if the conditional were based on ¬A. Conclusion: you don’t know that *if A, then B* if your belief is based on the premise that ¬A. Since I do know that *if ‘sibling’ means number, then sisters are not siblings*, my belief is not based on the premise that ‘sibling’ does not mean number. But that is just to say that my belief is not based on the actual-meaning fact. If it is not based on that, then it is not based on any empirical evidence.

need experience to establish that, had it turned out that *w*, it would have turned out that cassinis are oval.)

^worse

The priory intension is ~~more~~ than unnecessary, however. One can *never* tell a priori whether cassinis would have turned out to be oval, had it turned out that *w*. (I ignore the case where there are no cassinis.) 'Cassinis are oval' has, therefore, *nothing* in its priory intension. The same goes for 'cassinis are not oval'. It goes in fact for most sentences whose predicates express response-enabled concepts. If one can't determine a priori whether a counteractual object is P, then that object can't be put into P's priory intension, or ¬P's either. If the priory intensions of P and its negation are empty, then so in all likelihood are the priory intensions of sentences built on P.

Concepts like *oval* are not well-represented by their priory intensions. Still, you might say, why should that matter? The point of priory intensions is to predict epistemic status: if S fails to be a priori, there should be a world that is not in its priory intension. Why should the modal rationalist want any more?

One can see why more is wanted by considering the modal rationalist's refutation of physicalism. How does that argument go, with intensions understood as priory? First premise: it is not a priori that if PHYSICS, then PAIN. Second premise: if it is not a priori that if PHYSICS, then PAIN, then there are worlds that are not in that conditional's priory intension. Third premise: worlds not in that priory intension are zombie worlds—worlds physically like ours in which no one feels pain. Conclusion: there are zombie worlds.

The argument needs priory intensions to be like primary intensions in a certain respect. If PIs are primary intensions, then worlds that are not in a sentence's PI are worlds in which S is false. Does the same hold for priory intensions? It doesn't. If PIs are priory intensions, all we can say is that there is a *w* such that it *fails to be a priori that* it would have turned out that S, had it turned out that *w*. It might still be *true* that it would have turned out that S! There might be no way for it to turn out that PHYSICS without its also turning out that PAIN.

I present this as a problem for priory intensions, but epistemic intensions are every bit as vulnerable to it. That there are worlds lying outside S's epistemic intension does not show that there are worlds in which S is false, but only that you can't always get to S a priori. (The response will come that that is enough, since any S which cannot be verified by a priori means can be falsified by a priori means. But we have examples to the contrary, such as 'this equation describes an oval'.) I don't think that anything is gained, then, by switching to an apriorized notion of truth at a world. The balloon just bulges in a different place. Yes, there is a world outside the intension, but there is no reason to think that it falsifies S, as opposed to just failing to a priori verify it. Better to stick with primary intensions as defined above. S is conceptually necessary iff it holds however things turn out.

## 17. GRASPING MEANING

Why expect an analytic (conceptually necessary) sentence to be knowable a priori? Why expect a sentence whose meaning guarantees that it is true to have the further property that we can *see* that the sentence is true just from our grasp of its meaning? There might be ways of grasping meaning that do not tell us outright whether S is true if *w*, but only how to *work out* whether S is true if *w*. If S’s meaning is grasped like that, then its not being a priori that S does not establish the existence of a falsifying world. The sentence might be (like ‘this equation
an describes ~~on~~ oval’) a posteriori but true in every world considered as actual.

The only way out is to maintain that the indicated kind of grasp is not possible. One will have to maintain that grasp of meaning always takes a certain form, a form that discloses to the grasper whether the meaning is truth-guaranteeing. If all I can do is *work out* whether $w \rightarrow$ S, then I don’t understand S. To
already understand, I have to *know* that $w \rightarrow$ S.

Say that my understanding of S is *rationalistic* if it consists in whole or part of
already my *knowing* the conditionals. The road from analyticity to apriority would be a lot smoother if all understanding was rationalistic.

On what basis, though, can other forms of understanding be ruled out? What is the problem with grasping a word’s meaning other than rationalistically? The closest thing I’ve found to an explicit discussion is Chalmers’s reply to Loar in *The Conscious Mind* (1996).

FN:21 Summarizing greatly, Loar (1990) thinks that *pain* is a recognitional concept[21] and that *C-fiber firings* is a theoretical concept, and that that is enough to make them cognitively distinct. Their distinctness notwithstanding, ‘it is reasonable to expect a recognitional concept R to “introduce” the same property as a theoretical [concept] P’. So we cannot conclude from the non-apriority of ‘C-fiber firings are pains’ that C-fiber firings aren’t pains. The failure of apriority might be because *pain* is recognitional and *C-fiber firings* isn’t. If their a priori inequivalence is explained thus, then there is nothing to stop them from co-referring. These are fine things to claim, Chalmers says, but it is not clear that they can all be reconciled.

> [Loar] gives the example of someone who is able to recognize certain cacti in the California desert without having theoretical knowledge about them. But this seems wrong: if the subject cannot know that R is P a priori, then reference to R and P is fixed in different ways and the reference-fixing intensions can come apart in certain conceivable situations. (Chalmers 1996: 373)

This might seem to be based on a misunderstanding. Observational concepts (of which recognitional concepts are a subtype) do not have their reference fixed

[21] This section is sloppy about recognitional versus observational, and also about Loar-recognitional versus recognitional in our sense.

in *any* epistemically available way; hence they do not have it fixed in a different way than holds for theoretical concepts.

What can Chalmers be thinking, then? He knows that Loar *says* that 'recognitional concepts refer "directly" . . . without the aid of reference-fixing properties' (1996: 373). He just thinks Loar is wrong about this. 'The very fact that a concept *could* refer to something else (a different set of cacti, say) shows that a substantial primary intension is involved' (1996: 373).

But, Loar can concede a substantial primary intension. The directness he is talking about is *epistemic*; one doesn't (and couldn't) *infer* that the cactus is R from its lower-level properties. A substantial primary intension is at odds only with *semantic* directness, as I now explain.

Fact: R applies to these things and not those. Why? What explains the differential treatment? If the question has an answer, as let's assume it does, it will be a truth of the form: R applies to *x* if and only if *x* is so-and-so. Consider this property of being so-and-so. It might be considered a reference-fixer for R; like a reference-fixer, it tells you how a thing has to be for R to refer to it. *Oval* too has a reference-fixer in this sense. Whether a figure is oval is not a brute fact about it, but depends on its shape.

A reference-fixer in the *theoretical* sense is a statement of the qualifications for being referred to by R, as these might be judged by a (smart enough) semantic theorist. A reference-fixer in the *ordinary* sense, though, is a statement of the qualifications for being referred to by R, as these might be explained by a (smart enough) user of the concept, trying to enumerate the factors she takes to make R applicable.

The claim about recognitional concepts is that they lack *ordinary* reference-fixers. Speakers do not apply *oval* on the basis of a condition that they know (even implicitly) that sums up the requisite features. Speakers do not know any conditions like that. They do not know any conditions that get the extension right no matter what. The condition that comes closest is *looks egg-shaped*. But, as we have seen, things could have turned out so that some *bona fide* oval had the wrong looks, and/or a non-oval had the right looks. I know an oval when I see one, and that seems to be enough.

Chalmers is right about one thing: it would be a mistake to deny recognitional concepts reference-fixers in the theoretical sense. That would be to deny that a thing's status as *oval* was a function of its lower-level properties. But if the claim is that recognitional concepts lack reference-fixers in the ordinary sense, then it would seem to be true. Speakers don't (and often can't) determine extensions a priori by asking what has the R-making properties.

How does all this bear on the issue that Loar and Chalmers are primarily interested in: the issue of physicalism? Chalmers, you will recall, argues as follows. It is not a priori that if PHYSICS, then PAIN; so the primary intension cannot contain every world; so there are worlds physically like this one in which pain is lacking, so physicalism is false.

The problem is (once again) with the inference from *not a priori* to *less than full primary intension*. With certain concepts the link between apriority and primary necessity breaks down. And the way it breaks down gives the physicalist an opening. She can say this: *Pain* (like *oval*) is a grokking, or observational, concept. That being so, whether an objectively described state is a case of pain cannot be determined just by rational reflection. One has to 'sample' the state by experiencing it from the right sort of first-personal perspective.

Two consequences should be noted. First, suppose there were a world *w* physically like ours but without pain. That world would do nothing to explain the non-apriority of 'if PHYSICS, then PAIN'; or rather, it would do nothing that couldn't be done just as well by a world *with* pain. For *w* to help, our intuition of non-apriority would have to be owing to our awareness of *w*. But the relevant fact about *w* (that it lacks pain) is not available to us as students of its microphysical description. Just as you can't tell whether *w* has ovals except by sampling its shapes, so you can't tell whether it has pain except by sampling its brain states.

Second, not only is a world like *w* of no particular help in explaining the failure of apriority, it isn't needed. Suppose that *v* is a world just like ours in every physical respect. The question of whether there is (say) pain in *v* is the question of whether there is anything there that *hurts* if sampled in the right sort of first-personal way. Whether a state hurts when sampled by someone in the state is not the kind of thing that can be decided from the armchair. If we are trying to explain why PHYSICS doesn't a priori entail PAIN, a world whose zombieness can't be a priori ruled out works just as well as a true zombie world would.

## 18. EVALUATIVE PREDICATES

Our grasp of a concept is *rationalistic* if it consists in whole or in part of a certain kind of knowledge: knowledge of conditionals of the form *w obtains* → $x, y, z$, . . . *are the Cs*. Suppose that your conditionals put $x, y, z$, . . . into a concept's extension in *w*, while mine count $x, y, z$, . . . out. Then, by Leibniz's Law, your concept and mine are not the same. A single concept cannot have conflicting extensions in the same world.

Now, in some cases, it seems quite right that disagreements about what goes into the extension should make for differences in the identity of the concept. If you and I can't agree about whether to call a certain almost-round figure oval, and this is not because of misinformation, error, or oversight on either side, then probably we have different concepts; probably we mean slightly different things by the word. There is no question of trying to work out who is really correct, because our beliefs are not really in conflict.

Similarly, if we can't agree about whether recently widowed 98-year-old males are bachelors, and not because either of us is misinformed or confused, then probably we mean slightly different things by 'bachelor'. There is no question of trying to work out who is really right, because we aren't really disagreeing.

A phrase sometimes used for concepts of this kind is *intolerant of brute disagreement*; if we have the same concept, we should not 'brutely disagree' about what falls under it. Are all concepts like that? Imagine that we disagree about whether it was wrong of Smith to tell a lie in hopes of saving his child embarrassment. The disagreement can't be traced back to differences in factual information, or miscalculation or oversight on either side. Does this show that we mean different things by 'wrong'?

The usual view is that it doesn't. People who disagree about the extension of 'wrong' (and where the disagreement does not trace back to . . .) do not necessarily mean different things by the word. Likewise for disputes about what is beautiful or fitting or reasonable. You will get people angry if you brand these disputes 'merely verbal', just because you can't see any good way to bring the two parties into line. Some concepts, then, are *tolerant of brute disagreement.*

A lot of philosophers would claim something even stronger. So far is the meaning of 'right' from dictating a particular view of its extension that it positively *rejects* the notion that such dictation is possible. If I try to represent your side of a moral controversy as based in a misunderstanding of 'right', then I am the one who misunderstands. Questions of rightness are supposed to be *contestable* in the (rather minimal) sense that someone who brutely disagrees with you can't be charged on that basis alone with meaning something different by 'right'. Some concepts, then, seem to be *intolerant of intolerance of brute disagreement.*

How do we grasp the meaning of 'right'? If our grasp is rationalistic, then (assuming we mean the same by 'right') all of us know the same conditionals *w obtains* → $x, y, z,$ . . . *are right and other things aren't.* Someone operating with different conditionals attaches a different meaning to the word. In that case, though, the concept is intolerant of brute disagreement. And our concept of rightness is, on the contrary, intolerant of such intolerance.

That is one argument for the conclusion that we do not grasp evaluative concepts in a purely rationalistic way. Here is another. Recall a well-known puzzle about right and wrong. On the one hand, you can't derive an ought from an is. 'If N then M', where N is descriptive and M is evaluative, cannot be known a priori. On the other hand, it does seem to be a priori that the evaluative facts are fixed by the descriptive ones. There is a tension here; we have trouble seeing how the two claims are supposed to hang together. But we do not get an outright contradiction unless it is supposed that our grasp of evaluative concepts is rationalistic.

Assume with the rationalist that if it is not a priori that S, then there's a counteractual *w* such that ¬S. Then, from the fact that N does not a priori entail M, we can infer the existence of a *u* such that *u obtains* → $(N \;\&\; \neg M)$. Since it's also not a priori that if N, then ¬M, there should be a world *v* such that *v* → (N & M). But if N is descriptively complete (as we are free to suppose), then these two worlds taken together constitute a counter-example to the thesis that there can be no moral differences without underlying descriptive differences.

It could be objected that all *u* and *v directly* show is that things could have turned out so that N & M, and they could have turned out so that N & $\neg$M. To get to $\Diamond$ (N & M) and $\Diamond$(N & $\neg$M), one needs to assume that M does not change in broad content between *u* and *v*. But that is a fair assumption, for the facts relevant to reference determination are descriptive facts, and these are the same in both worlds. Hence we can argue as follows:

(1) It is not a priori that if N, then M, or that if N, then $\neg$M.
(2) If it is not a priori that S, then there's a *w* such that $w \rightarrow \neg S$.
(3) There are *u* and *v* such that $u \rightarrow (N \,\&\, \neg M)$ and $v \rightarrow (N \,\&\, M)$.
(4) M does not change in broad content between *u* and *v*.
(5) $\Diamond$ (N & M) and $\Diamond$(N & $\neg$M).

But (5) is an a priori falsehood. Somewhere or other a big meta-ethical mistake has been made.

I claim that the puzzle has nothing essentially to do with ethics. Consider the conditionals, 'if *x* is cassini-shaped, then it is oval', and 'if *x* is cassini-shaped, then it is not oval'. Neither is knowable a priori. Shouldn't there then be a pair of worlds *u, v*, exactly the same in geometrical respects but such that $u \rightarrow$ *cassinis are oval*, while $v \rightarrow$ *cassinis are not oval*? These worlds threaten to show that there can be differences in respect of ovality without underlying geometrical differences.

Where the ovality argument goes wrong is easy to see. The problem is (2). You can't get a world where cassinis are not oval out of the fact that it's not a priori that they are oval. If our grasp of ovality were purely rationalistic, then the failure of apriority *would* call for a counter-world. But it isn't, so it doesn't.

The morality puzzle can be pinned on the same mistake. You can't get a world where N and $\neg$M out of the fact that it's not a priori that if N, then M. It would be different if our grasp of rightness were rationalistic; then we would have a genuine paradox on our hands. I conclude that it isn't rationalistic. A similar argument can be given for other evaluative concepts. None, I claim, are grasped rationalistically. None are grasped in what modal rationalists consider to be the one way in which a concept can be grasped.

## 19. THEORETICAL PREDICATES

Consider, finally, theoretical predicates: acid, energy, force, mass, species, cause, mereological sum, essential nature. What can be said about our understanding of these? Do we understand 'energy' by knowing a lot of conditionals of the form 'had it turned out that *w*, such-and-such would have turned out to be the energy'?

Here are two arguments to the contrary, both harking back to the discussion of evaluative predicates. Suppose that we *do* (*qua* understanders of 'energy', etc.) know all these conditionals—that our concept of energy not only fixes, for each possible scenario, but discloses to us, for each of these scenarios, where the energy is to be found. How is it, then, that you and I continue to disagree about where the energy is to be found? (You say there is energy stored up in the curvature of space, while I deny it.) After all, there is a conditional known to both of us (as understanders of 'energy') that decides the matter. The explanation must lie in one of two places. It must be that

(i) someone is misconstruing the lower-level facts, and so picking the wrong conditional,

or:

(ii) someone is misconstruing their own mental states, specifically, the belief with that conditional as its content.

Whichever of these applies, our disagreement has the character of a misunderstanding. One or the other of us is laboring under a misimpression, and will (or should) change his or her tune when the mistake is pointed out. Of course, there is always the possibility that we associate *different* conditionals with 'energy'. In that case, though, we are not disagreeing at all; we mean different things by the word, so are talking past each other. None of the three scenarios allows for substantive disputes. Someone has made a mistake of type (i) or (ii), or else we are arguing over words.

This is almost as hard to accept here as it was in the evaluative case. Some disagreements *are* merely verbal, and some are based in correctable false impressions. The usual view, though, is that there's third category: honest-to-God conflicts about what it is reasonable to believe, between people in command of the same lower-level facts. The effect of the rationalistic theory of grasp is to eliminate this third category. ^a

The extension of 'energy' in a world is a function of what the correct scientific theory is. To find that theory, one must appeal at some point to considerations of naturalness, simplicity, nonarbitrariness, and the like—in a word, to considerations of *reasonableness*. (The positivists were the last to seriously question this.) Reasonableness is an evaluative concept and, as such, response-enabled. You can't hand responsibility over to 'rules of reasonableness'; there are no such rules, or at any rate not enough of them. You have to let yourself be led to some extent by your gut.

There are places where Chalmers sounds this theme himself. Figuring a concept's extension, he says, is not just grinding out conclusions. Judgment and discretion may be called for:

> the decision about what a concept refers to in the actual world [may] involve [] a large amount of reflection about what is the most reasonable thing to say; as, for example, with

> questions about the reference of ‘mass’ when the actual world turned out to be one in which general relativity is true, or perhaps with questions about what qualifies as ‘belief’ in the actual world. Consideration of just what the primary intension picks out in various actual-world candidates may involve a corresponding amount of reflection. But this is not to say that the matter is not a priori: we have the ability to engage in this reasoning independently of how the world turns out. (1996: 58)

I suppose that we do have this ability. We can ask ourselves what is the most reasonable thing to say on various hypotheses about how the world turns out. It is not clear, though, how that argues for the matter’s being a priori. We can also ask ourselves where the ovals are on various hypotheses about how the world turns out. Our conclusions in the second case aren’t a priori, so why should they be a priori in the first?

If the oval example shows anything, it’s that the move from ‘we can tell independently of how things turn out’ to ‘we can tell a priori’ is a *non sequitur*. For ‘we can tell independently’ may just mean that we can stage simulated confrontations with nature on various hypotheses about the form nature takes. It may not be obvious that searchers after the most reasonable hypothesis are doing this. But it seems to me that they are. Judgments of reasonableness and plausibility are arrived at by exercising a type of sensibility.

To be sure, the sensibility involved is not a perceptual one. And there seems less cause for worry about simulated plausibility judgments being a bad guide to real such judgments.[22] But the fact that sensibility is required should still give pause. It means that if you and I disagree about a sentence’s truth-value in *w*, there may be no more we can say to each other than ‘I find your position unreasonable’. The claim that everything but consciousness is a priori entailed by physics thus comes down to this: if two people disagree about a sentence’s truth-value in *w*, each will find his or her own position to be the more reasonable one, unless the sentence is about consciousness, in which case each side concedes the rational defensibility of the other. Even if this were true, it is hard to see an argument for metaphysical dualism in it. And it is not true; the zombie hypothesis is *much less reasonable* than the hypothesis that what people seem to be feeling, they are feeling.

## 20. LOGICAL EMPIRICISM AND MODAL RATIONALISM

There were two dogmas of empiricism. One was the analytic/synthetic distinction. The other was ‘semantic reductionism’—the idea that each statement is linked

[22] The moral case is arguably intermediate in these respects. Sensitivity to the moral aspects of things has often been likened to good vision or a keen sense of smell. And our horror at an observed case of, say, euthanasia or abortion may catch us by surprise, given our approving reaction to the imagined case. (Why else would right-to-lifers work so hard at getting us to *look* at what is being done?)

by fixed correspondence rules to a determinate range of confirming observations. Quine held that the two dogmas are 'at bottom the same'. For the correspondence rules are in a sense analytic. They give the sentence its meaning, so cannot fail as long as that meaning holds fixed. The dogmas are at least notionally different, though, and my focus will be on the second: the conception of correspondence rules as analytic, and therefore a priori. Although I will follow Quine in speaking mostly of analyticity, it is the apriority that is my real concern.

How is a modal rationalist like a logical empiricist? They seem initially very different. The empiricist has analytic correspondence rules connecting theory to experience. Modal rationalists aren't proposing anything like *that*. Yes, people have to be able to tell a priori whether S is true in a presented world. Gone, though, is any thought of that world being presented *in experiential terms*. There is no case~~, then,~~ for a charge of *phenomenalistic* reductionism.

If one looks, though, at Carnap's writings on protocol sentences, it turns out that his sort of reductionism did not have to be terribly ~~experiential~~ phenomenalistic either. Under the influence of Neurath, Carnap thinks that it is somewhat of an open question which sentences ought to be counted as protocols. Sometimes a protocol sentence is said to be any sentence 'belonging to the physicalistic system-language' which we are prepared to accept without further tests.[23] Often it is said to be a matter of *convention* which sentences will count as protocols. The important point for us is that Carnap thinks there are a priori rules connecting theoretical statements with protocols, whatever protocols turn out to be.

Another seeming difference emerges from Quine's complaint that Carnap overlooks the 'holistic nature of confirmation'. The complaint might be understood like this: One never knows whether S is really correct until all the observational evidence is in. Hence any rules portraying S as verifiable on the basis of limited courses of experience—courses of experience small enough to be enjoyable by particular observers—would be untrue to the way in which confirmation actually works.

This complaint the rationalist can rightly claim to have answered. He never represents *partial* information as enough to ensure that S; the rules he contemplates take as input *complete* information:

> [Quine says that] purported conceptual truths are always subject to revision in the face of sufficient empirical evidence. For instance, if evidence forces us to revise various background statements in a theory it is possible that a statement that once appeared to be conceptually true might turn out to be false.
>
> This is so for many purported conceptual truths, but it does not apply to the supervenience conditionals that we are considering, which have the form 'If the low-level facts turn out like this, then the high-level facts will be like that'. The facts specified in the antecedent of this conditional effectively include all relevant empirical factors. . . . The

[23] Ayer (1959: 237).

> very comprehensiveness of the antecedent ensures that empirical evidence is irrelevant to the conditional's truth-value. (Chalmers 1996: 55)

This is a good answer as far as it goes. But there are aspects of Quine's critique that it does not address. Quine says that

> the dogma of reductionism survives in the supposition that each statement, taken in isolation from its fellows, can admit of confirmation or infirmation at all. My countersuggestion, issuing essentially from Carnap's doctrine of the physical world in the Aufbau, is that our statements about the external world face the tribunal of sense experience not individually but only as a corporate body. (1961: 41 in rpt.)

The problem here is not that S's confirmational status is underdetermined until all the empirical evidence is in. The problem is that S's confirmational status is not fully determined even by the full corpus E of empirical evidence. The degree to which E confirms S, Quine thinks, is tied up with the extent to which E or aspects of E are deducible from S. But nothing of an observational nature is deducible from S except with the help of a background theory T. Hence the degree of support that E lends to S depends on which background theory we use.

This complaint would be easily evadable if there were an analytically guaranteed fact of the matter about which theory E selects for. One could simply ask whether E supports S relative to the E-preferred theory, whatever it might be.

One has to assume, then, that this is what Quine is really concerned to deny. He denies that there are analytic connections between total corpuses E of empirical evidence and theories T of nature. Without these, there can be no analytic connections between E and particular statements S. A number of things suggest that analytic confirmation relations are indeed the target:

> ● Q1 I am impressed, apart from prefabricated examples of black and white balls in an urn, with how baffling the problem has always been of arriving at any explicit theory of the empirical confirmation of a synthetic statement. (Quine ●1961: 49 ~~in rpt.~~)

This could be taken to mean just that the sought-after theory of confirmation would have to be very complicated. But Quine has something different in mind. He is aware, after all, of Carnap's attempts to work out a logic of confirmation ^that ~~which~~ would tell us what to believe on the basis of given evidence. He is aware, too, that the attempt failed even for the simplest sort of examples. Carnap came up with a whole array of confirmation functions, none of them looking a priori better than the rest.

Where does this leave us? One problem with analytic confirmation relations concerns total evidence. This the rationalist has addressed. But there's a second problem: 'total science, mathematical and natural and human, is underdetermined by experience' (Quine 19~~5~~1: 45 ~~in rpt.~~). The version of underdetermination Quine needs is really a rather mild one. He needn't deny that there is an objectively best theory relative to a given body of evidence. He needn't even ^6 deny that there's a single most rational theory to adopt. All he need claim is that the choice between theories compatible with the evidence cannot be based just on our grasp of meaning. It 'turns on our vaguely pragmatic inclination to adjust one strand of the fabric of science rather than another. Conservatism figures in such choices, and so does the quest for simplicity' (Quine 1951: 49 in rpt.).

This can be reconciled with the analytic view of confirmation relations only by supposing that my grasp of the language tells me how conservative I should be, and how important simplicity is, and how these sorts of desiderata trade off against one another. If two scientists judged the trade-offs differently, at most one could be considered to be speaking correctly—that is, in accordance with the meanings of her words. That, however, is not how the science game is played.

The interesting thing is that Carnap *agrees* that it's not how the science game is played. His goal, as he usually describes it, is not to uncover the true nature of meaning, but to give us tools for making our discursive practice more rational and efficient. He thinks that disputants should pick a common framework and then resolve their disagreements by reference to its assertion rules:

> it is preferable to formulate the principle of empiricism not in the form of an assertion . . . but rather in the form of a proposal or requirement. As empiricists we require the language of science to be restricted in a certain way. (Carnap 1936–7: sect. 27)

Based on passages like this, one recent commentator has summarized the view as follows:

> Criticisms of the meaning/belief distinction rest on the lack of a principled criterion for [semanticality]—no empirical method can be found for making it. However, for Carnap, such a distinction is to be reached by agreement in a conflict situation. Maximize agreement on framework issues and situate disagreement on either empirically answerable problems or on questions of a pragmatic nature about the framework. (O'Grady 1999: 1026)

One can argue about whether this would really be helpful. All I am saying right now is that not even Carnap believes that it is how we really operate: that our actual practice lends itself to a distinction between semantic factors in assertion and doxastic ones.

Is there anyone who does believe that this is how we operate? The modal rationalist does, or at least, such a view is not far from the surface. We are told that grasp of S's meaning, or at least the kind of grasp you need to count as understanding S, is knowing which worlds *w* are such that had this turned out to be *w*, it would have turned out that S. This applies not just to observation-level statements, but to theoretical statements as well. It is part and parcel of knowing T's meaning to know what the world would have had to be like for it to be the case that T. And that is not obviously different from Carnap's idea of analytic confirmation rules.

I say 'not obviously different', because there may be room for maneuver on the issue of what is involved in 'knowing which worlds are S-worlds'. I have been assuming that worlds are given in 'lower-level' terms, whatever exactly that might mean. What if worlds are described more fully than that, perhaps as fully as possible? There would be no need to infer that theory T applied; it would be given that it applied in the world's initial presentation. This seems tantamount to saying that one knows the S-worlds as, well, the
as S-worlds, or the worlds such that ~~if~~ they turned out actual, it would ~~turn~~ out had have turned
that S.

But, if a 'homophonic' grasp of the set of verifying worlds were all one needed, ~~then~~ there would be no reason to expect a sentence to be knowable a priori just because its primary intension contained all worlds.

This is clear from Chalmers's discussion of physicalism. Consider again the conditional 'if PHYSICS, then PAIN'. It is claimed that the only way for this to be non-a priori is for there to be worlds not in its primary intension: there have got to be zombie worlds. If our grasp of primary intensions was homophonic, the failure of apriority would present no puzzle, hence no puzzle to which zombie worlds might be offered as a solution. The reason I don't know a priori that if PHYSICS, then PAIN is that I can't tell a priori whether the primary intension of 'if PHYSICS, then PAIN' contains all worlds. I can't tell that because I can't tell a priori whether the PHYSICS worlds are a subset of the PAIN worlds. If they are a subset, there is no puzzle as to why the understander doesn't realize it, because it is assumed from the outset that PHYSICS worlds are, for all she knows a priori, worlds without PAIN.

How, then, are worlds presented to the meaning-grasper? She must be able to pick out the S-worlds on the basis of their ground-level properties. 'If the low-level facts turn out like this, then the high-level facts will be like that' (Chalmers 1996: 55). These conditionals are ~~thought~~ to be analytic; indeed, they held
are true in virtue of the aspect of meaning to which we have a priori access. This is why I say that modal rationalists are committed to something *like* the analytic confirmation relations advocated by Carnap and rejected by Quine. The rationalist who wants to escape Quine's criticisms has got to (a) show that the criticisms don't work even against logical empiricism; (b) show that the cases are relevantly different.

To accomplish (a) would be to find a mistake in Quine's reasoning. Maybe, for example, it's just untrue that theory is underdetermined by evidence. To accomplish (b) would be to show that what the modal rationalist says is different enough from what the logical empiricist says that the Quinean critique doesn't generalize. Maybe, for example, the lower-level facts on the basis of which we can tell a priori whether S are quite unlike the 'empirical' facts on the basis of which we can't tell a priori whether S. I won't pursue the matter any further here, but I suspect that the prospects for doing either of these things are not terribly good.

## 21. DIGRESSION: IMAGINATIVE RESISTANCE

Hume, in 'The Standard of Taste', points out something surprising about our reactions to imagined circumstances. Reading a story according to which S, I try to imagine myself in a situation where S really holds. The surprising thing is that we can do this quite easily if S is contrary to *descriptive* fact, but have a great deal of trouble if S is contrary to *evaluative* fact. Reading that Franco drank from the Fountain of Youth and was made young again, you don't blink twice. But reading that it was good that little Billy was starved to death since he had, after all, forgotten to feed the dog, you want to say, 'it was *not* good, I won't go along'.

FN:24 Call that *imaginative resistance*.[24] Why does it happen? A number of explanations have been tried. Do we resist because what we're asked to imagine is conceptually false? No, because (i) counter-moral hypotheses are *not* conceptually false (remember essential contestability), and (ii) lots of conceptually false scenarios are *not* resisted (as readers of Calvino and Borges will attest).

Do we resist because what we're asked to imagine is morally repugnant? No, because we balk at aesthetic misinformation as well. 'All eyes were on the twin Chevy 4 × 4's as they pushed purposefully through the mud. Expectations were high; last year's blood bath death match of doom had been exhilarating and profound, and this year's promised to be even better. The crowd went quiet as special musical guests ZZ Top began to lay down their sonorous rhythms. The scene was marred only by the awkwardly setting sun.' Reading this, one thinks, 'If the author wants to stage a monster truck rally at sunset, that's up to her. But the sunset's aesthetic properties are not up to her; nor are we willing to take her word for it that last year's blood bath death match of doom was a thing of beauty.'[25] FN:25

Do we resist because the scenario is repugnant along *some evaluative dimension or other*? No, because it is not only evaluative suggestions that are resisted. You open a children's book and read as follows: 'They flopped down beneath the great maple. One more item to find, and yet the game seemed lost. Hang on, Sally said. It's staring us in the face. This is a *maple* tree we're under. She grabbed a five-fingered leaf. Here was the oval they needed! They ran off to claim their prize.' Reading this one thinks, 'If the author wants it to be a maple leaf, that's her prerogative. But the leaf's physical properties having been settled, whether it is oval is not up to her. She can, perhaps, arrange for it not to have the expected mapley shape. But if it does have the expected shape, then there is not a whole lot she can do to get us to imagine it as oval.'

Imaginative resistance arises not only with evaluative predicates, but also with (certain) descriptive ones: 'oval', 'aquiline', 'jagged', 'smooth', 'lilting'. What

[24] ~~On imaginative resistance~~, see Gendler (2000), Moran (1989), and Walton (1994). caps

The author [25] ~~She~~ knows this, moreover. Why make a suggestion you know will not be accepted? There might be any number of reasons, but most likely she is just pulling our leg.

do these predicates have in common? P makes for imaginative resistance if, and because, the concept it expresses is of the type I have called 'grokking', or response-enabled.

Why should resistance and grokkingness be connected in this way? It's a feature of grokking concepts that their extension in a situation depends on how the situation does or would strike us.[26] 'Does or would strike us' *as we are*: how we are represented as reacting, or invited to react, has nothing to do with it. Resistance is the natural consequence. If we insist on judging the extension ourselves, it stands to reason that any seeming intelligence coming from elsewhere is automatically suspect. This applies in particular to being 'told' about the extension by an as-if knowledgeable narrator.

## 22. (CONCEPTUALLY) CONTINGENT A PRIORI

I have called a lot of claims a priori. But not much has been done to explicate the notion; the focus has been more on conceptual necessity. I doubt that it is possible to explain apriority in all its guises with the materials at hand. But I'll try in the next few sections to clarify a particular type of apriority as far as I can. (Nothing argued so far depends on what is coming next.)

'Water contains hydrogen' is touted in *Naming and Necessity* as an example of an a posteriori metaphysical necessity. 'Cassinis are oval' has been touted here as an example of an a posteriori *conceptual* necessity. A posteriori conceptual necessities are the counterpart in our system of the a posteriori metaphysical necessities that Kripke emphasized.

One might wonder whether we have anything to correspond to Kripke's *other* famous category: the category of a priori but (metaphysically) contingent truths like 'Neptune is the planet if any responsible for . . .'.

I suggested above that 'unless we are greatly misled about the circumstances of perception, a figure is oval iff it looks egg-shaped' was a priori, or close enough for present purposes. But of course things *could* have turned out so that we were unable to see eggs in oval figures. Things could have turned out so that we never saw anything as egg-shaped.

Had things turned out so that nothing looked egg-shaped, would the world have turned out to be oval-free? The answer seems clear. How we see things is irrelevant to how they are shaped. It would have turned out that there were ovals which, however, did not look the way ovals are supposed to look.

I make no prediction about what we would have *said*. It may be that we would have said 'there are no ovals'. That is irrelevant unless the meaning that 'oval' would have turned out to have in that circumstance is the meaning it has

[26] I assume that fictional situations are presented as counteractual, not counterfactual. One is to think of them as really happening.

actually. And it seems clear that the meanings are different. If people say 'there are no ovals' in a world geometrically just like ours, they do not mean the same thing by 'oval' as we do.

'Unless . . ., a figure is oval iff it looks egg-shaped' is an a priori but conceptually contingent truth. It could have turned out that we were not prone to see ovals as egg-shaped, perhaps because we were not prone to see anything as egg-shaped. And, approaching it from the other end, it could have turned out that almost-circular figures looked to us egg-shaped, despite not being oval.

This seems at first puzzling: how can it be a priori that 'oval iff looks egg-shaped' when it could have turned out otherwise? One has to remember that the scenario where it turns out otherwise is *also* a scenario where it turns out that 'oval' doesn't mean what we all know it does mean. A scenario in which 'oval' changes meaning can no more stop 'oval iff looks egg-shaped' from being a priori than one in which '=' means nonidentity can stop 'Phosphorus = Phosphorus' from being a priori.

## 23. APRIORITY VERSUS CONCEPTUAL NECESSITY

I said that 'oval' could have turned out not to mean what we all know it does mean. What we all know it does mean is *oval*. So I could equally have said that it could have turned out that 'oval' did not mean oval. I do not shrink from this way of putting it, or even the claim that it could turn out (though it won't) that 'oval' *doesn't* mean oval.

I admit, however, that these claims sound funny. If we accept that 'oval' could have turned out not to mean oval, then it seems like we should regard as not completely insane someone (Crazy Eddie) who says that 'oval' *doesn't* mean oval. He could turn out to be right! Intuitively, though, there is no chance whatever of Crazy Eddie's turning out to be right.

What does it take for Crazy Eddie to be vindicated? It is not enough that, letting S be the sentence he uttered, it could have turned out that S. The scenario in which it turns out that S could be a scenario in which S has changed meaning. You are not vindicated unless what you said turns out to be right; it's not enough that what you turn out to have said turns out to be right. Otherwise Warrenites would be vindicated if 'Oswald acted alone' turned out to mean that Oswald had help, and he did. There is no danger of Crazy Eddie turning out to be right, because, letting M be the (actual) meaning of his words, had it turned out that M, it would have turned out that M was not what he said!

I assume that 'it could turn out that . . .' is an intensional context—that is, a context treating synonyms alike. Since 'sister' is synonymous with 'female sibling', and it could turn out (though it won't!) that 'sister' does not mean female sibling, it could turn out that 'sister' does not mean sister. The reason why it sounds funny to say it is that the statement strongly *suggests* something absurd: namely, that someone who conjectures that ‘sister’ doesn’t mean sister could turn out to be right.

Another (not incompatible) way to explain the funniness is this. There is a use of ‘it could turn out that S’ on which it means that it is not a priori that ¬S. In that (alternative) sense of the phrase, it really *couldn’t* have turned out that ‘sister’ didn’t mean sister. For we know a priori that ‘sister’ means sister. If it doesn’t sound as bad to say that ‘sister’ could turn out not to mean female sibling, that might be because we don’t know a priori that it does mean female sibling.

Compared to conceptual necessity, apriority is an elusive notion. One reason has already been noted. If it is a priori that ‘sister’ means sister, but not that it means female sibling, then ‘it is a priori that . . .’ is not an intensional context; it cares about the difference between synonyms. (‘It could ~~have~~ turned out that . . .’ (in the alternative epistemic sense) is therefore not intensional either.)

Stranger even than the failure of intensionality is the following. The class of a priori truths is often claimed to be closed under (obvious) logical consequence. This can’t be right, if a well-known account of apriority is even roughly correct. It is a priori that S, according to the well-known account, if one can know that S is true just on the basis of one’s grasp of S’s meaning. Suppose I know that A and that A ⊃ B just through my grasp of the two sentences’ meanings, and then I infer B. If this is my reason for believing B, then I do not know it a priori. For my belief is based in part on my grasp of A’s meaning, and A is a different sentence from B. ital

The failure of logical closure helps us resolve a puzzle. There are many things I know a priori. For instance, I know a priori that sisters are sisters, and that Hesperus = Hesperus. If ‘S’ is a sentence I understand, then I would seem to FN:27 know a priori that ‘S’ is true iff S.[27] (More on this claim below.)

But I rarely, if ever, know a priori that a sentence ‘S’ is true; for truth-value depends on meaning, and my knowledge of meaning is a posteriori. I have to learn what a sentence means, even a sentence of my own idiolect. And my views on the topic are rationally defeasible under the impact of further evidence.

The question is, why can’t I combine my a priori knowledge that sisters are siblings with my a priori knowledge that if they are siblings, then ‘sisters are siblings’ is true, to arrive at a priori knowledge that ‘sisters are siblings’ is true?

The problem is not that I can’t *modus ponens* my way to the conclusion that ‘S’ is true, starting from premises known a priori. The problem is that, having done so, it is not just in virtue of understanding ‘“S” is true’ that I know that ‘S’ is true. The understanding I have of ‘S’ plays a role too, and that is something over and above my understanding of ‘“S” is true’. (I can understand the latter while momentarily forgetting what ‘S’ means, or while entertaining a skeptical hypothesis to the effect that it means something other than I had thought.) Since

EXTRA #

EXTRA #

[27] Notice the quotation marks. Use/mention distinctions that had been left to context are here marked explicitly.

I cannot claim to know that 'S' is true just in virtue of my understanding of that very sentence, I cannot claim to know a priori that 'S' is true.

## 24. APRIORITY

~~What can we say about apriority to~~ How to explain these puzzling features? Since apriority is a matter of what my grasp of a sentence's meaning tells me, our account has got to bring in grasp explicitly. What aspect of grasp could function to tell me that the sentence is true? A state that tells me something is a state whereby I possess information. So our account should be in terms of the information I possess whereby I grasp meaning. Call this my *grasp-constituting information* about 'S'. The proposal is that

(AP) it is a priori (for me) that S iff for some G
(a) that 'S' is G is part of my grasp-constituting information, and
~~and~~
(b) being G conceptually necessitates being true.

Let's revisit some earlier questions with (AP) in hand.

*How can it be a priori that 'sister' means sister yet not a priori that it means female sibling?* That 'sister' means sister is part (all?) of the information whereby I grasp 'sister'. I do of course realize 'on the side' that to be a sister is none other than to be a female sibling. But that is a collateral belief which does not figure in my grasp. Suppose the belief changed in response to some *outré* counter-example; that would be a change in what I thought sisters were, but not a change in what I meant by 'sister'.

*Why are the a priori truths not closed under logical consequence?* Having deduced B from A and A ⊃ B, I am in possession of information given which B has to be true. But there is no reason to expect the information to be grasp-constituting with respect to B; on the contrary, the information by which I grasp A is likely to be involved. To know B a priori, I need to know it on the basis of the information whereby I grasp B.

*How can an a priori truth fail to be conceptually necessary?* The information G that conceptually necessitates that 'S' is true might not be conceptually necessary information. If 'S' has a conceptually contingent property that conceptually necessitates that 'S' is true, all I can conclude about 'S' truth-wise is that it *is* true given how matters actually stand. Conceptual necessity requires more than this: 'S' must be true on *any* hypothesis about how matters stand, including the false ones.

Example: I am newly arrived in the royal court. A helpful attendant explains that 'the king' is to be understood so that 'the king is the guy giving orders, wearing the crown, and so forth' comes out true. I come as a result to know a priori that the king is the guy giving orders, and the rest. Now, as a matter of fact, it is Richard who is doing all these things; as a matter of fact, it is Richard who is the king. But things *could* have turned out so that it was an impostor Richerd who was giving orders, and so on. Would the king then have turned out to be Richerd?

I have certainly been given no reason to think so. I was told that 'the king' stood for the order-giver by someone who supposed (correctly) that the order-giver was Richard. They leaned on that supposition in defining 'the king' as the order-giver. Leaning on a supposition that they knew could turn out to be false, they were careful *not* to say that the king would still have been the order-giver however things had turned out. And indeed, he wouldn't: things could have FN:28 turned out so that the king was Richard, while the order-giver was Richerd.[28] It is conceptually contingent that the king = the order-giver. Still, I know it a priori.

*Why are some conceptually necessary truths not a priori?* Sometimes the information that a speaker possesses about 'S' whereby she grasps its meaning is information that exhibits 'S' as true. Other times, it isn't. I am not sure what a typical understanding of 'cassinis are oval' involves, but one is not expected to realize that it is true. You should perhaps know that things looking egg-shaped are to be counted oval. But that doesn't enable you to work out that cassinis are oval until you've laid eyes on one.

*If 'sisters are siblings' can turn out not to be true, yet sisters cannot turn out not to be siblings, then in some counteractual world sisters are siblings and 'sisters are siblings' is untrue. Why isn't this a world in which the T-biconditional fails?* It *is* a world where the T-biconditional fails. It could have turned out that 'sisters are siblings'
^was ^~~is~~ untrue although sisters ^~~are~~ siblings. This seems odd until we remember that ^were
^ could it ^~~can~~ happen only if 'sisters are siblings' turn~~s~~^ out not to mean what it does ^ed
mean. A world where it turn~~s~~^ out not to mean what it does mean is a world ^ed
where my grasp-making information fails. A world where that information fails is irrelevant to the issue of whether that information entails the truth of the biconditional—and so to the issue of whether it holds a priori that 'sisters are siblings' is true iff sisters are siblings.

*Why does the feeling persist that if it is not a priori that S, there is a counteractual world in which ¬S*? There is an argument to this effect that almost works. If there

[28] A related sort of presupposition is discussed by Putnam: 'Suppose I point to a glass of water and say "this liquid is called water". . . . My "ostensive definition" of water has the following empirical presupposition: that the body of liquid I am pointing to bears a certain sameness relation to . . . most of the stuff I and other speakers in my linguistic community have on other occasions called "water." If this presupposition is false because, say, I am without knowing it pointing to a glass of gin and not a glass of water, then I do not intend my ostensive definition to be accepted. Thus the ostensive definition conveys what might be called a defeasible necessary and sufficient condition. . . . If it is not satisfied, then one of a series of, so to speak, "fallback" conditions becomes activated' (1975: 225). I would add only that the series tends to be a finite one. Some defeats you recover from; an (itself defeasible) backup condition kicks in. Eventually, though, the backups are exhausted, and the definition just fizzles.

are no counteractual worlds in which ¬S, then every counteractual world is an S-world. A fact like that surely figures in the information whereby we understand 'S'.[29] The fact entails that 'S' is true, and so grasp-making information entails that 'S' is true, and so S is a priori. Contraposing, if S is not a priori, then it does not hold in all counteractual worlds. But, the sentence beginning 'a fact like that surely figures' assumes our grasp is rationalistic. The feeling persists because we forget that there are other ways to understand.

FN:29

## REFERENCES

● Q2 ●Adams, E. W. (1975). *The Logic of Conditionals*. Dordrecht: D. Reidel.

Almog, Joseph, Perry, John, and Wettstein, Howard (eds.) (1989). *Themes from Kaplan*. New York: Oxford University Press.

Ayer, A. J. (ed.), (1959). *Logical Positivism*. New York: Free Press.

Carnap, Rudolf (1936–7). 'Testability and Meaning'. *Philosophy of Science*, 3:419–71; 4: 1–40.

Chalmers, David (1994). 'The Components of Content'. Philosophy/Neuroscience/Psychology Technical Report 94–04, Washington University; <http://www.u.arizona.edu/~chalmers/papers/content.html>.

____ (1996). *The Conscious Mind: In Search of a Fundamental Theory*. New York: Oxford University Press.

____ (2000). 'The Tyranny of the Subjunctive'; <http://www.u.arizona.edu/~chalmers/papers/tyranny.txt>.

Davies, Martin, and Humberstone, Lloyd (1980). 'Two Notions of Necessity'. *Philosophical Studies*, 38: 1–30.

● Q3 DeRose, Keith (1991). 'Epistemic Possibilities'. *Philosophical Review*●, 100: 581–605.

Evans, Gareth (1979). 'Reference and Contingency'. *Monist*, 62: 161–89.

Gendler, Tamar Szabó (2000). 'The Puzzle of Imaginative Resistance'. *Journal of Philosophy*, 97(2): 55–81.

Grice, H. P. (1989). 'Indicative Conditionals'. In *Studies in the Way of Words*, Cambridge, Mass.: Harvard University Press pp. 58–85.

Jackson, Frank (1979). 'On Assertion and Indicative Conditionals'. *Philosophical Review*, 88: 565–89; repr. in Jackson (1991), 111–35.

____ (1991). *Conditionals*. Oxford: Oxford University Press.

____ (1994). 'Armchair Metaphysics'. In Michaelis Michael and John O'Leary-Hawthorne (eds.), *Philosophy in Mind*, Dordrecht: Kluwer, 23–42.

____ (1998). *From Metaphysics to Ethics: A Defense of Conceptual Analysis*. Oxford: Clarendon Press.

Kripke, Saul (1980). *Naming and Necessity*. Cambridge, Mass.: Harvard University Press.

Loar, Brian (1990). 'Phenomenal States'. *Philosophical Perspectives*, 4: 81–108.

McGinn, Colin (1999). *Knowledge and Reality*. Oxford: Clarendon Press.

[29] It also assumes that ~~This is a bit of an exaggeration, since~~ knowing of each $w$ that $w \rightarrow S$ suffices for ~~is not yet~~ knowing that every $w$ is such that $w \rightarrow S$ ~~for all $w$~~.

Moran, Richard (1989). ‘Seeing and Believing: Metaphor, Image, and Force’. *Critical Inquiry*, 16: 87–112.

O’Grady, Paul (1999). ‘Carnap and Two Dogmas of Empiricism’. *Philosophy and Phenomenological Research*, 49: 1015–27.

Peacocke, Christopher (1989). ‘Perceptual Content’. In Almog et al. (1989), 297–330.

Putnam, Hilary (1975). ‘The Meaning of “Meaning” ’. In *Mind, Language, and Reality* (Cambridge: Cambridge University Press), 215–71.

Quine, Willard van Orman (1951). ‘Two Dogmas of Empiricism’. *Philosophical Review*, 60: 20–43; repr. in Quine (1961), 20–46.

____ (1961). *From a Logical Point of View*, 2nd edn. New York: Harper & Row.

Segerberg, Krister (1972). ‘Two Dimensional Modal Logic’. *Journal of Philosophical Logic*, 2: 77–96.

Stalnaker, Robert (1972), ‘Assertion’. In Peter Cole (ed.), *Syntax and Semantics*, ix, New York: Academic Press, 315–32.

____ (1990). ‘Narrow Content’. In C. A. Anderson and J. Owens (eds.), *Propositional Attitudes*, Stanford, Calif.: Center for the Study of Language and Information pp. 131–45.

____ (1991). ‘How to Do Semantics for the Language of Thought’. In Barry Loewer and Georges Rey (eds.), *Meaning in Mind: Fodor and his Critics*, Oxford: Blackwell pp. 229–38.

Walton, Kendall (1994). ‘Morals in Fiction and Fictional Morality’. *Proceedings of the Aristotelian Society*, supp. vol. 68: 27–50.

White, Stephen (1982). ‘Partial Character and the Language of Thought’. *Pacific Philosophical Quarterly*, 63: 347–65.

Should be Quine 1961: 45

## Queries in Chapter 4

Q1. Please check this year.

Q2. Please check and confirm the page numbers written by author in the margin has been updated correctly.

Page numbers have
been updated correctly.

Q3. Author correction is not clear.

This was the editor's
correction not mine

# 5

# No Fool's Cold: Notes on Illusions of Possibility

A lot of philosophers are *pessimistic* about conceivability evidence. They think it does not prove, or even go very far towards justifying, interesting modal conclusions. A number of other philosophers are *optimistic*; they think it does justify, and perhaps even establish beyond a reasonable doubt, that lots of interesting things are possible. Nothing very surprising there. What is slightly surprising is that both groups ~~can~~ claim to find support for their attitude in the work of Saul Kripke.

Pessimists say: Kripke shows that conceivability evidence is highly and systematically *fallible*. Very often $E$ seems possible, when as a matter of fact, $E$-worlds cannot be. So it is, for instance, with the seeming possibility of water in the absence of hydrogen, or of Hesperus distinct from Phosphorus, or of this table turning out to be made of ice. Let the pessimistic thesis be

(P) oftentimes $E$ seems possible when it is not, so conceivability evidence is not to be trusted.

Optimists reply: yes, Kripke finds conceivability evidence to be fallible, but that is only half of the story. The rest of the story is that *the failures always take a certain form*. A thinker who (mistakenly) conceives $E$ as possible is correctly registering the possibility of *something*, and mistaking the possibility of *that* for the possibility of $E$. There are *il*lusions of possibility, if you like, but no *de*lusions or hallucinations. Let the optimistic thesis be

(O) carefully handled, conceivability evidence can be trusted, for if impossible $E$ seems possible, then something else $F$ is possible, such that we mistake the possibility of $F$ for that of $E$.

This paper was presented at the UNC Greensboro conference on imagination and possibility, with comments by Keith Simmons. Thanks to Keith for exposing various gaps in the argument, not all of which I have been able to deal with here. Thanks to Kit Fine, Tamar Szabó Gendler, Janine Jones, and Saul Kripke for discussion at the conference, and to David Chalmers and Tyler Doggett for extremely helpful written comments provided more recently.

The optimistic thesis (O) represents conceivability evidence as in a sense *in*fallible. If (O) is correct, then that *E* seems possible, while it may not establish that *E* is possible, does succeed in establishing the disjunctive conclusion that either *E* is possible or *F* is. And indeed in certain cases we can get all the way to the first disjunct, because *F* is tantamount to *E* or entails *E*. This, the optimist continues, is the situation we encounter in the last few pages of *Naming and Necessity*, where Kripke argues against the identity theory of mind. It seems possible that pain is not c-fiber firings, and the *F* that supposedly snookers us into thinking *E* possible is tantamount to that original *E*. (I will be questioning that argument in due course.)

It seems likely that both groups are overinterpreting Kripke. Certainly Kripke is not a pessimist, because he closes the book with a positive argument of the sort that pessimists are bound to find fault with. And although this is not as clear, he seems to stop short of outright optimism too. He says (in "Identity and Necessity") that "the only model *I can think of* for what the illusion might be . . . does not work in this case" (1977: 101; emphasis added). Others are welcome to argue in favor of some other model that does not require a genuinely possible *F*. Kripke is skeptical, to be sure: "it would have to be a deeper and subtler argument than I can fathom and subtler than ever appeared in any materialist literature that I have read" (1977: 101). But although Kripke has his doubts about the availability of an alternative model, he does not entirely rule it out. (One is reminded of Carnap's position in "Empiricism, Semantics, and Ontology": I can't make sense of the question of realism my way; maybe others can find a different way, but it won't be easy.)

So the door is open, technically anyway, to "a deeper and subtler argument" aimed at establishing that *some* seeming possibilities do not reflect *any* sort of genuine possibility. Whether this deeper and subtler argument can be given has not been terribly much explored.

One idea sometimes encountered is that there are differences in how pains and c-fiber firings are entertained in thought that *all by themselves* explain why each would seem possible without the other. Thomas Nagel's version of this idea is that c-fiber firings are imagined perceptually—"we put ourselves in a conscious state resembling the state we would be in if we perceived it"—while pain is imagined sympathetically—"we put ourselves in a conscious state resembling the thing itself" (1974: note 11). He maintains that:

> the relation between them will appear contingent, even if it is necessary, because of the independence of the disparate types of imagination. (1974: note 11)

Chris Hill says in a similar vein that the relation appears contingent because our concept of c-fiber firings is theoretical while our concept of pain is phenomenological. Between concepts like that "there are no substantive a priori ties," and the absence of such ties allows us to "use the concepts to conceive coherently of situations . . . in which there are particulars that fall under one of the concepts but do not fall under the other" (1997: 75).

This sort of approach is in one way too broad and in another too narrow. It is too broad in that it threatens to undermine conceivability arguments that most of us find attractive. It certainly *seems* to me that my dog Ruby could have been in severe pain right now; that's what you normally get for harassing a porcupine. But then so it would, according to Nagel, what with Ruby being imagined perceptually and the pain sympathetically.

^modal

I agree that the ^ appearance here should not be taken seriously, if it arises in the way Nagel says. That we do take it seriously suggests that the explanation may not be quite so simple. And indeed there are independent reasons to think matters are not so simple. If appearances of contingency resulted just from "disparate types of imagination", then one would expect more to seem possible than in fact does. After all, it is not just the dog that is imagined perceptually but everyday objects in general. Consider the rock that Ruby is perched on. All the Nagelian conditions are in place, yet it does *not* seem that the rock could have been in pain right now. It takes more to tempt us into an illusion of possibility than Nagel supposes.

What about Hill's version of the idea? It seems to me, as I consider this cup of vinegar, that a cup of $H_2O$ could look just the same. But then so it would, on Hill's view, for *looking the same* is a phenomenological concept, while our concept of $H_2O$ is theoretical. Once again, though, this cannot be all there is to it, for there are cases where Hill's conditions are met and the appearance of contingency is lacking. A cup of molten lead does *not* present itself as capable of looking like this.[1] FN:1

How is the Nagel-type approach too narrow? By focusing so intently on subjective versus objective, it just reinforces the impression that Kripke is trying to create: namely, that any response to his argument is going to require some kind of special pleading on behalf of the mental. I cannot rule it out, of course, that the proper response *does* require special pleading. But it would be better if we could identify a general constraint on modal illusions that is independently motivated and that just happens to deliver the desired results when applied to the intuitions supporting mental/physical dualism.

I want to explore some of these issues by looking at the role of *actuality* in modal judgments. Actuality comes in under two separate headings. On the one hand it can figure in the *content* of a modal judgment. The thing that seems possible—the condition that seems like it could have obtained—can have the

[1] Tyler Doggett and Daniel Stoljar point out that the Nagel worry also pulls the rug out from under standard objections to behaviorism and functionalism. Given any behavioral property B, we can imagine being in pain without exhibiting B, and vice versa. Perhaps the appearance of contingency here is due just to the fact that pain is imagined sympathetically and B perceptually.

notion of actuality in it. This is in fact quite common. One says, for instance, FN:2 "this lemonade is cold but it could have been colder".[2] Colder than what? Colder than it actually is, of course. If C is the "how cold was it?" parameter, then our judgment is roughly this

seems $\Diamond$ (C exceeds $C_@$).

Or perhaps we are doing a puzzle where five irregularly shaped pieces of plastic have to be rearranged into a square. We look the pieces over, and it strikes us that the thing can be done. What seems possible, however, is not that the pieces can be made to form a square *after being melted down and recast as rectangles*; it's that they can be made to form a square with their actual shapes and sizes held fixed. If the shape and size of piece X is S(X), then our judgment is

seems $\Diamond$ (the Xs form a square & $\forall X\ (S(X) = S_@(X))$).

A remark attributed to Richard Taylor gives us a third example. "Why are people so sure they could have acted otherwise?" he asks. "After all, nobody ever has." One reason we think this is that it very much *seems* as though we could have acted otherwise:

seems $\Diamond$ (my action was of a type T incompatible with the type $T_@$ of the action I really did perform)

To have a schema for judgments of this kind, what seems possible is that a certain parameter P should have taken a value ~~so~~-and-so related to the value it actually takes: thus

seems $\Diamond$ (. . . & P is ~~so~~ and so related to $P_@$ & . . .). thus

That is the first way actuality can come in. It leads pretty directly to a second way. Whether or not it seems possible for some parameter to assume a value thus ~~so~~-and-so related to its actual value is not independent of what we know, or think we know, about what the actual value in fact is, or indeed of other information we possess about actuality. It would not have seemed possible for the pieces to be rigidly rearranged into a pentagon if we had believed each piece to be square, or round. It would not have seemed possible for the lemonade to be colder if it was believed to be at zero degrees already. It might not have seemed possible for us to act otherwise were we convinced that Frankfurt's nefarious neurologist (made omnipotent if necessary) stood ready to reprogram our brains if we tried.

There is a temptation, perhaps, to treat this as just more content. But the temptation should be resisted, because it imports more into the content than belongs there. Our judgment is not

seems $\Diamond$ (this lemonade is colder than $N^\circ$ C).

[2] Could have been colder as a liquid, I mean. Assume for the sake of the example that so-called frozen lemonade is not really lemonade.

After all, we may have little positive idea what temperature the lemonade is in degrees centigrade. What seems possible is that the lemonade should be colder than it is, and why it seems possible has to do with the lemonade's felt temperature.[3]

If our sense of the temperature doesn't figure in content, though, what role *does* it play? It plays what might be called a *presuppositional* role. The judgment is conditioned on our temperature experience's not being too misleading. One thinks, "unless I am very much misled about how cold this liquid is, it could have been colder". Besides appearing in the *content* of a model judgment, then, actuality can figure in the *background* to the judgment, that is, the beliefs or presuppositions that allow the seemingly possible thing to seem possible.

Back now to the main issue. The optimist says that whenever there is the illusion that *E* is possible, there is a related hypothesis *F* that really is possible. For instance, it seems that Hesperus could have been distinct from Phosphorus because there really could have been two planets there, one responsible for Hesperus-appearances and the other for the appearances we enjoy of Phosphorus. I have said a little about *E*, the content of the (perhaps mistaken) intuition, but nothing about *F*, the hypothesis that is supposed to really be possible.

Kripke does not even pretend to give us a general strategy for recovering *F*—what I will call *the underlying possibility*—from *E*. What he does do is, first, sketch lots of highly convincing examples; second, suggest that at least some of the time, it is good enough to replace names in *E* with corresponding reference-fixing descriptions; and third, characterize *F* as the "appropriate corresponding qualitative contingent statement". He explicitly refrains, though, from giving a "general paradigm" for the construction of the proposition whose possibility fools us into thinking *E* possible.

A number of other writers have been bolder. Some say that there is the illusion that *E* is possible because the sentence "*E*" could (with its "meaning" in some sense of that world held fixed) have expressed a true proposition, albeit not the proposition it expresses in fact. So,

(a) it could have happened that "*E*" expressed a true proposition.

I myself once conjectured that *E* seems possible because we could have thought something true with the thought (the internal mental act) whose content in this world is *E*. So,

(b) it could have happened that thinking the *E* way was thinking truly.

The best-known suggestion along these lines is that *E* seems possible because there are worlds such that if (contrary to what we perhaps suppose), they are actual, then *E*. So a third hypothesis is that

[3] Specifically, with its feeling warmer than lemonade on the verge of freezing feels.

(c) things could have been a way such that, if they actually *are* that way, then *E*.

All these proposals are variations on the theme of *E* seeming possible because what it says is correct, if a certain not-impossible world is actual. Nothing important is lost if we ignore ~~any differences~~ the differences between them and speak simply of the *if-actually* account of illusions of possibility.

The if-actually account works extremely well in some cases. The reason it seems possible that the table should turn out to be made of ice is that there are worlds with the property that if they are actual, then it *is* made of ice. The reason it seems possible that Hesperus should have been other than Phosphorus is that there are worlds with the property that if they are actual, it really is other than Phosphorus. It turns out, though, that the account cannot deal correctly with actuality-based modal contents. I will build up to this slowly.

Ivory-billed woodpeckers had been thought extinct; recently, though, a man named David Kullivan reported spotting a pair of them. I happen to believe this report, but not everyone does. Knowing that his word would be doubted, Kullivan was tempted (let us say for purposes of the example) to shoot one of the woodpeckers and bring its body back as proof. According to me, believing as I do that ivory-billed woodpeckers exist, had Kullivan shot one, there would have been fewer ivory-billed woodpeckers than there are. To me, then

seems $\Diamond$ (there are fewer ivory-billed woodpeckers than actually).

Now suppose that I am wrong and there are no ivory-billed woodpeckers. Then I am under an illusion of possibility; a smaller number seems possible, but there cannot be fewer than none. What explains my illusion? The story would have to be that this seems possible because there is a world such that if it is actual, then there *are* fewer ivory-billed woodpeckers than there actually are. And that makes no sense.

Of course, there is no peculiarly *modal* illusion here; where I go wrong is in believing in ivory-billed woodpeckers in the first place. But consider a second example. It seems possible that Hesperus could have turned out to be distinct from Phosphorus. It seems, for instance, that Phosphorus could have turned out to be Mars rather than Venus. Another thing that seems possible is for Phosphorus to have turned out to be *Xorg*, a solar planet over and above the planets that exist in fact. It seems possible, then, that there should have been more planets than actually: all the actual ones, including Hesperus, and then in addition Phosphorus = Xorg.

seems $\Diamond$ (there are extra planets; Hesperus is Venus but Phosphorus is new).

The story would have to be that this seems possible because if we are wrong and the morning-visible planet is "new", then there really *are* more planets than actually. And that clearly cannot be right. Again, it strikes us that gold could have turned out to have a different chemical makeup. The illusion that gold could have failed to be the 79th element *can* be explained, notice. But I may not know that gold is any kind of element; my thought is just that it did not *have* to turn out with that chemical makeup, whatever its makeup in fact is. This illusion cannot be explained on the if-actually model, for we would need a world such that gold has a different makeup than it actually does on the supposition that ~~this~~ that world is actual.

So the if-actually account cannot explain certain illusions of possibility, those in which the hypothesis that seems possible involves a contrast or comparison with actuality.[4] Why should we bother about this? The reason for bothering is that it tells us something about how people are thinking of the modal illusion problem. The if-actually account is exceedingly popular. (I stress that Kripke does not endorse it.) Why, if there is a class of illusions it does not address? *It must be that this class of illusions has not been much on people's minds.* People have been assuming, implicitly anyway, that the contents of error-prone modal judgments are *actuality-neutral* in the sense, roughly, that facts about which world is actual are irrelevant to what the judged hypothesis says. Perhaps to be safer I should just say that there has been a tendency to downplay or underestimate the actuality-based aspects of these contents, and to play up or overestimate their actuality-neutral aspects.

One sort of problem this bias in favor of neutrality leads to has already been seen. But the problem that interests me is not that certain actuality-based illusions will prove difficult to explain, but that certain such illusions will be "explained" too easily. This is how it would happen:

(1) What seems possible is a hypothesis $E$ that is actuality-based.
(2) An actuality-neutral (or more neutral) hypothesis $E'$ is covertly substituted.
(3) One explains the illusion that $\Diamond E'$ as a subtle misreading of $\Diamond F'$.
(4) It would take a very much *grosser* misreading of $\Diamond F'$ to fall under the illusion that $\Diamond E$.
(5) One thinks the $E$ illusion has been explained when really it has not.

I will give examples in a minute. But first let me link the worry up with what I take to be an important feature of Kripke's procedure.

Kripke does not just want to show how someone *could* fall under the misimpression that, say, Hesperus could have failed to be Phosphorus, by

[4] One natural idea about actuality-involving illusions (suggested independently by Robert Stalnaker and David Chalmers) is this: they are to be explained by saying there is a world $w$ such that if $w$ is actual, then the actuality-involving proposition is *possible*. It seems possible for there to have been fewer ivory-billed woodpeckers because this really is possible on the hypothesis that Kullivan's story is true. But the intuition that Hesperus could have been an additional planet is not based in any factual misinformation of the sort we might try to correct by treating $w$ as actual. The feeling is not that assuming Phosphorus is other than Hesperus, it could have been Xorg. The feeling is that Phosphorus, although (it turns out) identical to Hesperus, could have been distinct from it in a way that bumped up the number of planets.

misinterpreting what was in fact a different possibility. That would be easy, since a sufficiently confused person could presumably misinterpret anything as anything. He wants to show that we plausibly *do* fall under the modal misimpression by misinterpreting a different possibility. It is not just that an intuition of *E*'s possibility *could* be but that our intuition of its possibility plausibly *is*, based on the mistaking of one possibility for another.

An example of someone who seems to underestimate the aspiration here is Michael Della Rocca in "Essentialism and Essentialists" (~~*Journal of Philosophy*~~ 1996). Say that Lumpl is the lump of clay composing the statue Goliath. It seems possible that Lumpl could have failed to be Goliath, or any other statue; it seems possible, indeed, that Lumpl could have existed *in the complete absence of statues*.

shouldn't this be Della Rocca 1996?

(a) seems ◊ (Lumpl exists without any statues).

Della Rocca maintains that this intuition is (or might be for all Kripke has to say about it) explained by the possibility that *a lump of clay handled by artisan A at time T* should have lacked all these properties.

(b) really ◊ (a lump handled by *A* at *T* exists without any statues).

I suppose that (b) *might* perhaps explain the illusion of someone for whom the reference of "Lumpl" was fixed by "the lump of clay handled by *A* at *T*". But "Lumpl" in *our* mouths has its reference fixed by "the lump composing the statue Goliath". (That is how I introduced the term above, and that is the usual way of introducing it.) So, the genuine possibility needed to explain away *our* intuition is

(c) really ◊ (a lump composing the statue Goliath exists without any statues).

But there is no such possibility as (c); it cannot happen that a lump both composes a certain statue and fails to coexist with any statues. The scenario that (c) calls possible, and whose possibility would be needed to explain away the intuition that Lumpl could exist without statues, makes no sense.

I seriously doubt, then, whether our *actual* intuition of Lumpl without statues can be defeated as easily as Della Rocca suggests.[5] The only real possibility in the neighborhood is the one recorded in (b). And there is no way on earth that we are

FN:5

[5] Della Rocca brushes up against this problem in a footnote. "One might, perhaps, see some other property as the property in terms of which Lumpl is identified. Even if some other property is the identifying property, the argument that I am about to give would not be affected because I shall show that *any* property that might plausibly be seen as the property in terms of which Lumpl is identified would be a property that allows a Kripkean reconstrual of our intuition of contingency in this case to go forward" (1996: 197). I do not see that he ever shows this. What he does say is that "Lumpl seems to be identified in terms of the designation, 'lump formed by, etc.', or some similar designator. Any such designator would allow the reconstrual to go through" (1996: 197–8). This is false, unless "similar" means "designator H such that there could be an H without Goliath existing". The designator "clay composing Goliath" is a~~n~~ ~~obvious~~ counter-example.

ministerpreting *that* as the possibility of Lumpl without any statues. The proof that (b) does not explain (a) is just that stare at (b) as long as you like, one cannot imagine being so confused as to have been fooled by it into supposing that (a). One is not at all tempted to say: oh, I see, once you point out the difference, it's because *this* really is possible that I supposed *that* to be possible.[6]

The kind of principle I am relying on here is familiar from psychoanalysis. Here is what in my brief (well, . . .) experience psychoanalysts tell you. "You are under the impression that nobody loves you. I submit that this is an illusion. A cruder sort of doctor might say, here is how the illusion arises, take my word for it. But I would never dream of asking you to take my word for it. No, the test of my explanation is whether you can be brought to accept the explanation, and to accept that your judgment is to that extent unsupported." The analogy is good enough that I will speak of the

*Psychoanalytic Standard* Assuming the conceiver is not too self-deceived or resistant, $\Diamond F$ explains *E*'s seeming possibility only if he/she does or would accept it as an explanation, and accept that his/her intuition testifies at best to *F*'s possibility, not *E*'s.

This is a high standard, but what makes Kripke's approach so convincing is that this is the standard he tries to meet, and mostly *does* meet. Philosophers have been telling us for centuries that this or that common impression is false; and we have for centuries been shrugging them off. What makes Kripke special is that he gets you to *agree* that you are making the mistake he describes.

I said that Kripke "mostly" meets the psychoanalytic standard. This is because I think that with at least some of the illusions he discusses, the standard is *not* met, and is perhaps unmeetable. Let me start with an example where a psychoanalytically acceptable explanation *can* be given. I will then argue that a crucial feature of the example goes missing in Kripke's treatment of certain other examples.

Kripke says, ". . . though we can imagine making a table out of another block of wood or even from ice, *identical in appearance to this*, and though we could have put it in this very position in the room, it seems to me that this is not to imagine this table as made of wood or ice, but rather it is to imagine another table, resembling this one in all external details, made of another block of wood, or even of ice" (1980: 114; emphasis added).

Imagine someone, call them Schmipke, expressing puzzlement about Kripke's procedure: "Hasn't Kripke gone to a lot of unnecessary trouble here? Why

[6] Della Rocca (2002 and personal correspondence) agrees that the (b) possibility is not *judged* explanatory. He thinks, however, that any attempt to justify this judgment winds up begging the question at issue: which modal intuitions are windows on possibility and which are illusions of possibility?

does he impose this condition of *identical in appearance* with the actual table? 'Identical in appearance' suggests that the other-worldly table looks just like the real one *to us*: if both of them were sitting here side by side, we could not tell them apart. This is suggested as well by the language he uses in 'Identity and Necessity': "I could find out that an *ingenious trick* has been played on me and that, in fact, this lectern is made of ice" (1977: 88). The ice has to be 'cleverly hardened' in the shape of a table, and presumably painted too. Otherwise it would not be a spitting image of our actual table, as Kripke clearly intends. Is any of this really necessary? Why does Kripke ask *w* to satisfy the *actuality-based* condition that its table looks or would look just the same to us? What is wrong with the *neutral* condition of, not identical *in* appearance, but simply: identical appearances?"[7]

FN:7

This seems a fair question, so let us try it. Until further notice, all we require from *w* is that there ~~is~~ be an icy table there, and that the people looking at it (perhaps counterfactual versions of ourselves) have the same experiences qualitatively speaking as we do looking at our table. It is of course compatible with this that the tables look to *us* very different. But then our reason for thinking of the icy table in *w* as "in disguise", cleverly tricked up to look like wood, no longer applies. Now that we have dropped the identical-in-appearance requirement, the icy table can be made any number of ways. Let it be, say, a table-shaped, table-sized, but otherwise perfectly ordinary frosty white block of ice. Of course, it needs to be added that the observers in *w* are spectrum-inverted with respect to observers here, so that the qualitative appearances they enjoy in front of a frosty white object are just like the ones we enjoy when looking at an otherwise similar brown object. But if both of those changes are made at once, then the experience of observers there looking at their table is just like the experience we enjoy looking at ours.[8]

FN:8

Note that there is *some* slight support for Schmipke's position in the text. Kripke says that what the icy table intuition comes to is that "I (or some conscious being) could have been qualitatively in the same epistemic situation that in fact obtains, etc." *He does not say the conscious being has to resemble me in any important respect*. The counterfactual being's brain might be wired so that it is in the same qualitative state standing in front of an icy table as I am standing in front of a wooden one. So, contrary to what we said above, it could be that Kripke is imposing only the neutral condition of *icy table, appearances XYZ*.

The question is, does the revised explanation meet the psychoanalytic standard? Does it explain our illusion that *this table could have turned out to be made of ice*, to point out that had our brains been different, a regular icy table would have caused in us the same qualitative state that a wooden table does cause in us? I

[7] Or, if that is not neutral enough, let the condition be not that observers in *w* enjoy qualitatively identical appearances, but that they enjoy qualitative appearances PQR. I will ignore this complication.

[8] Schmipke concedes the possibility of spectrum inversion.

tend to think it does not. Because what seems possible is that this table *with relevant perceptible properties held fixed* could have turned out to be ice. No one is going to be tempted into thinking *that* possible by reflection on the possibility that we see a regular icy table as brown, because in that scenario the perceptible properties *change*. The color of the table goes from brown to white.[9]

FN:9

It may help to consider an analogy. Say that I am under the impression that that animal there [pointing] is a zebra, when really it is a horse. Dretske's explanation is this: "The horse is painted to look just like a zebra. When two things look just the same, the one is easily mistaken for the other. It makes sense then that you would take this horse for a zebra." That corresponds to the Kripkean explanation of the "could have turned out to be ice" illusion. Because the table's appearance is indistinguishable from that of disguised ice, one naturally concludes that it could *be*, or have been, disguised ice.

Imagine now a second, Schmipkean explanation of my zebra illusion. "The horse is not painted at all. And you're enjoying ordinary horsy phenomenology. But there is this guy counter-Steve, a counterfactual variant of yourself, who has zebraish phenomenology when looking at a horse, and horsy phenomenology when looking at a zebra. Because your phenomenology is indistinguishable from that of counter-Steve looking at a zebra, it makes sense that you would take this horse for a zebra." That corresponds to the Schmipkean explanation of the "could have turned out to be ice" illusion. Because my actual table phenomenology is indistinguishable from my alter ego's ice phenomenology, I am led to suppose that this table could be, or have been, a regular old hunk of ice.

Is it just me, or does the first pair of explanations work better than the second? "I am liable to confuse A with B because they look the same to me" sounds quite plausible. If things look the same, then one is indeed liable to confuse them. "I am liable to confuse A with B because the same looks result if it is me looking at A or counter-Steve looking at B." *There is no chance at all* that I am confusing myself with counter-Steve, even if his phenomenology is just the same. Counter-Steve is by definition a person who sees things differently than I do. (One might as well worry that our planet has all along been Twin-Earth, making water not $H_2O$ but XYZ.)

So we have the following principle: to explain why *this*, understood to present like so, seems like it could turn out to be Q, one needs a possible scenario in which something *superficially indistinguishable* from it does turn out to be Q. The counterfactual thing has to look the same, not to the counterfactual folks, but to us. I will call that a *facsimile* of the actual thing. And I will refer to the principle as the facsimile or fool's gold principle.

[9] A property is perceptible iff when an object perceptually appears to have it and does not, we have misperceived. Not all properties figuring in the content of a perceptual state are perceptible in this sense. Our experience may represent the table as wooden, but it is not as if our eyes are playing tricks on us if it is well-disguised ice.

Kripke gives two models for the explaining-away of the intuition that A could be Q. First is the reference-fixer model:

(RF) it seems possible for A to be Q because it really is possible that the so-and-so is Q, where "the so-and-so" is a descriptive condition fixing "A" 's reference.

Then there is the epistemic counterpart model:

(EC) it seems possible for A to be Q because it really is possible for A* to be Q, where A* is a facsimile of A.

The epistemic counterpart model might seem the more accommodating of the two, because it does not require anything in the way of reference-fixing descriptions. But there is a respect in which the reference-fixing model is more accommodating and indeed *too* accommodating.

The epistemic counterpart model requires an A* indiscernible in relevant respects from A, what we have called a facsimile of A. Can this requirement be enforced by asking A* to satisfy some carefully constructed reference-fixing description D? It is not at all obvious that a suitable D can be found. One ~~obvious~~ possibility is "the thing that puts me into qualitative state 279". The picture this gives is:

$$\begin{tikzcd} \text{me-in-@)}\ QS_{279} \arrow[r] & A \\ \text{me-in-}w)\ QS_{279} \arrow[r] & A^* \end{tikzcd}$$

Can the arrows please be right-to-left instead of left-to-right?

Here we have dissimilar observers in distinct worlds confronting two (perhaps readily distinguishable) objects and reacting the same way. (EC) by contrast envisages a single observer confronting two objects to which she responds identically:

$$\begin{tikzcd} & A \\ \text{me-in-@)}\ QS_{279} \arrow[ur] \arrow[dr] & \\ & A^* \end{tikzcd}$$

Perhaps we can arrange for the second picture by letting D be the "the thing that puts me *as I actually am* into qualitative state 279". But this forgets that "the thing that actually puts me in state 279" stands in counter-Steve's mouth for A*. We are left again with the first picture.

One could try to force the second picture by letting D be "the thing that *in* $\alpha$ puts me into state 279", where $\alpha$ is a stable designator of actuality; it picks out our world @ no matter in which actual or counterfactual context it is uttered. But the point of a reference-fixing description is that it is supposed to be a piece of language that directs us to the referent across a range of counterfactual situations. And the term "whatever in $\alpha$ puts me into state 279" is not even *understandable* in counterfactual situations. Had things been different, we would not have been thinking, "too bad things are so different here, it would be ~~how much~~ better to live in a non-counterfactual world like $\alpha$".

Two pictures have been sketched of how to explain away modal illusions. Which of the two is meant to apply in the case of the icy table? Passages like "I (*or some conscious observer*) could have been in qualitatively the same epistemic situation" (1980: 142; emphasis added) suggest the first picture. But there are also passages like this:

> . . . it seems to me that this is not to imagine this table as made of wood or ice, but rather it is to imagine another table, *resembling this one in all external details*, made of another block of wood, or even of ice. (1980: 114; emphasis added)

"Resembling in all external details" means, I take it, that we would not notice if the one table were instantaneously substituted for the other. And that is the second picture. The reason this matters is, once again, that the first picture fails to explain the illusion. It defies credulity that my feeling that *this table* could have been made of ice is based on the fact that *my brain* could have been such that suitably carved ice elicited in me the present sort of appearances.

But let us not dwell too long on the icy table example, since Kripke uses it mainly for illustration. His real interest is in the kind of modal illusion that arises in science. Here is some heat; is it some type of molecular energy?[10] One has to conduct further tests, and, like any tests, they could come out either way. So there is the appearance that heat could turn out to be a certain type of molecular energy, and the appearance that it could turn out to be something else. The second appearance is an illusion. How does Kripke propose to account for it?

> the property by which we identify [heat] originally, that of producing such and such a sensation in us, is not a necessary property but a contingent one. This very phenomenon could have existed, but due to differences in our neural structures and so on, have failed to be felt as heat. (1980: 133)

It might be, for instance, that due to differences in our neural structures *high* mean molecular energy—henceforth HME—felt cold, and *low* mean molecular energy—henceforth LME—felt hot. Does this explain in a psychoanalytically satisfying way our feeling that it could have been LME that was heat rather than HME? Does pointing to possible differences in our neural structures explain why this cold seems like it could have turned out to be HME?

Here is the worry. With the table, remember, what seemed possible is not only that ice could have paraded itself in front of *someone or other* who saw it as I see wood, but that there could have been ice that *I with my existing sensory faculties* would have seen as wood. To explain *that* seeming we needed a facsimile of the table—a spitting image of it—that was in fact ice. Likewise what seems

[10] Like Kripke, I will run heat together with temperature.

possible in the case of LME is not just that it could have paraded itself in front of *someone or other* who felt it as hot, but that *I* with my existing neural structures could have found it to be hot. To explain that seeming, we need a counterfactual facsimile of heat that turns out on closer inspection to be LME. There should in other words be the possibility of LME-type fool's heat. Similarly, to explain the seeming possibility of cold turning out to be HME, we would need the possibility of fool's cold that was found by scientists to be HME.

Is there fool's heat of this type, or fool's cold? I do not see how there could be. It may be possible to slip a cleverly disguised icy table in for this wooden one with no change in visual appearance. *But it is not possible to slip cleverly disguised LME in for HME and have it feel just the same.* Having substituted low ME for high, there is no way to preserve the appearances but to postulate observers who react differently than ourselves to the same external phenomena. But then what we are getting is not really fool's heat but something more like *dunce's* heat. You would have to be pretty confused to see in the possibility of rewiring on *your* side the explanation of why a switcheroo seems possible *on the side of the phenomenon you are sensing*. Whether fool's heat is absolutely impossible I don't know. But what does seem clearly impossible is for *LME to be fool's heat*, because it by hypothesis feels the opposite of hot; it feels cold.

Kripke is right, or anyway I am not disagreeing, when he says that "the property of producing such and such a sensation in us . . . is not a necessary property", because we could have been wired differently. LME could, it seems, have produced what we call sensations of cold. That is not what I am worried about. What worries me is that the property of interest is not that but *producing such-and-such a sensation in us as we are*. And this property is, I suspect, necessary. There would seem to be three factors in how an external phenomenon is disposed to feel: its condition, our condition, and the conditions of observation. If all these factors are held fixed, as the notion of fool's heat would seem to require, then it is hard to see how the sensory outcome can change.

Someone might say: that LME can't be fool's heat doesn't show that there can't be fool's heat at all. Surely there is *something* in some faraway world that although not HME feels or would feel hot to us as we are. Suppose that is so,[11] and call the something ABC ("alien basis caliente"). ABC is all you need to

[11] Kripke actually discusses something like this in *Naming and Necessity*. "Some people have been inclined to argue that although certainly we cannot say that sound waves 'would have been heat' if they had been felt by the sensation which we feel when we feel heat, the situation is different with respect to a possible phenomenon, not present in the actual world, and distinct from molecular motion. Perhaps, it is suggested, there might be another form of heat other than 'our heat', which was not molecular motion; though no actual phenomenon other than molecular motion, such as sound, would qualify. Although I am disinclined to accept these views, they would make relatively little difference to the substance of the present lectures. Someone who is inclined to hold these views can simply replace the term . . . 'heat' with . . . 'our heat'. . . ." (p. 130 n. 68)

explain the illusion that heat could have been other than HME in the approved Kripkean fashion, that is, in terms of a genuine underlying possibility.

But, granted that one *can* explain, or try to explain, the illusion in this way, would the explanation be correct? I am not sure that it would, for the following reason. Our feeling that heat could have turned out to be something else is *indifferent* to whether the something else is alien ABC or actual LME. It would be very surprising if the feeling had two radically different explanations depending on the precise form of the something else. The LME form of the illusion *cannot* be explained by pointing to a possible facsimile of heat that really is LME. (Whether LME can be fool's heat is a factual question, and the answer is that it can be at best dunce's heat.) Therefore the ABC form of the illusion ought not to be explained with a possible facsimile either.

I have been arguing that *strong* epistemic counterparts, or facsimiles, are needed to explain illusions of possibility. However, there are some illusions to which epistemic counterparts, strong or weak, might seem altogether irrelevant. It seems possible not only that heat could have failed to be HME, but also that HME could have failed to be heat. Kripke treats the latter illusion as reflecting the genuine possibility that HME might not have felt hot. Given that epistemic counterparts do not figure here at all, the insistence that any epistemic counterparts should be strong may seem to leave Kripke's explanation untouched.

Once again, I appeal to the principle that similar intuitions should receive similar explanations. Our intuition that HME could have turned out to be *something* other than heat differs only in specificity from the intuition that it could have turned out to be *cold*. Weak epistemic counterparts of cold are of no use in explaining the latter illusion; it does not matter what "those people" (the residents of *w*) think. But if other-worldly observers are irrelevant here, then they are irrelevant to the unspecific intuition as well.

The upshot is that if S is a sensed phenomenon like heat, and P is a physical phenomenon like LME, then other-worldly observers are no use in explaining *either* why S seems like it could have been other than P, or why P seems like it could have been other than S. Since, as we have seen, actual observers cannot explain these apparent contingencies either, it seems that there is no psychoanalytically satisfying explanation in Kripke for the appearance that S is only contingently related to P.

But, someone might say, this just shows we have been going about it the wrong way around. Rather than looking for a strong epistemic counterpart of *heat* that is LME, we should be looking for a strong epistemic counterpart of *me* to whom LME feels hot.

I do not deny that such a person is possible; the question is what he can do for us. It seems not an accident that the intuitions explained by facsimiles of the table are intuitions about what is possible for *the table*. Likewise, the intuitions explained by gold-facsimiles are intuitions about *gold*, for example, that it could have turned out to be iron pyrites. One would expect, then, that the intuitions explainable by reference to me-facsimiles are in the first instance intuitions about me. Am I the sort of person who has heat sensations in response to HME, or the sort of person to whom LME feels hot? There is the feeling (suppose for argument's sake that it is an illusion) that I could have been the second sort of person. How does this feeling arise? Well, a possible strong epistemic counterpart of mine *does* have heat sensations in response to LME.

But it is one thing to explain apparent de re possibilities for ourselves, another to explain apparent de re possibilities for heat. When we ask, "did heat have to be HME or could it have been LME?", and answer that it could have turned out either way, we are caught between two seeming possibilities *for heat*. The proof of this is that the seeming possibility of heat being LME does not depend in the least on there being Steve-like beings around to whom LME feels hot. (Perhaps heat's being LME creates conditions inhospitable to life.) The intuition that heat could have been LME *although there was no one around to realize it* cannot be explained by pointing to a possible me-facsimile reacting differently to LME, simply because it is stipulated in the intuition that no observers are present.

Here is the position so far. It is not hard to disguise a genuinely icy table so that it looks wooden. So if Kripke wants to explain the seeming possibility of this table A being made of ice, he has at his disposal a facsimile A* of the table that really is made of ice. Sometimes, though, the appearance is closer to the reality, and facsimiles of A are no more capable of possessing the seemingly possible property Q than A is itself. How the second sort of illusion arises is an interesting question, but a question for another paper.[12] The claim for now is just that we cannot explain the second sort of illusion by pointing to a world where an A-facsimile really is Q, because such a world is not possible.

Kripke says, "perhaps we can imagine that, by some miracle, sound waves somehow enabled some creature to see. I mean, they gave him visual impressions just as we have, maybe exactly the same color sense. We can also imagine the same creature to be completely *insensitive* to light (photons). Who knows what subtle undreamt of possibilities there may be?" (1980: 130). He asks, "Would we say that in such a possible world, it was sound which was light, that these wave motions in the air were light?" He says no, "given our concept of light, we should describe the situation differently" (1980: 130).

I agree. The indicated world does not testify to the genuine possibility of light being pressure waves in the air. But now let us ask a slightly different question. Does it explain the *seeming* possibility of light having turned out to be waves in the air? Again the answer is no. For that you would need sound to be a facsimile of light. And it is not, for the obvious reason that airwaves do not look the least bit like light. But then what *does* explain the seeming possibility of light turning

[12] I suspect that the explanation is often as simple as this: there is a facsimile of A that might *for all we know a priori* be Q.

out to be compression waves in the air? I am not going to comment on that. What we do know is that the explanation is *not* in terms of a genuinely possible strong epistemic counterpart.

One further example, this time not taken from Kripke. Suppose that Q is a broadly geometrical property our concept of which is recognitional. Q might be the property of being jagged, or loopy, or jumbled. It might be the property of "leftiness", which we recognize by asking if the figure in question appears to be facing left (in the manner of 'J' and '3'), or right (in the manner of 'C' and '5'). I will focus for no particular reason on the property of being *oval*. Everyone knows how to recognize ovals, but nobody knows the formula (there is no formula to know). The one and only way to tell whether something is oval is to lay eyes on it and see how it looks. A thing is judged oval iff it looks more or less the shape of an egg.

Now suppose I tell you that *cassinis* are the plane figures, whatever they may be, defined by the equation $(x^2 + y^2)^2 - (x^2 - y^2) = 5$. Is being a cassini a way of being oval? I take it that until you do the experiment, this is an empirically open question. Cassinis could turn out to be oval or they could turn out not to be. You need to draw the figure and see how it strikes you.[13]

This seems not too different, intuitively, from the way LME needs to be sampled to determine whether or not it is heat. Presumably the Kripkean will want to give the same sort of explanation. Just as there are worlds where HME feels hot and worlds where it feels cold, there are worlds where cassinis look egg-shaped and worlds where they look to be shaped like bunny ears or figure-8s.

But this is all a mistake, since for cassinis to look other than egg-shaped to us as we are is impossible. There may perhaps be counterfactual observers who due to their greater visual acuity are bothered by departures from the exact profile of an egg that we ourselves hardly notice. To them, cassinis do not look egg-shaped. But those observers can no more explain the seeming possibility of cassinis' turning out not to be oval than spectrum-inverted observers can explain the seeming possibility of the table's being made of ice. This is because what seems possible (until we do the experiment) is that cassinis look other than egg-shaped to us as we are, with our existing sensory endowment.[14]

[13] Cassinis as I have defined them are oval. (They belong to the class of "cassinian ovals"—oddly, most cassinian ovals are not egg-shaped at all.)

[14] It is not as easy as one might think to throw the facsimile requirement over as too onerous. If the appearance that A could be Q is sufficiently explained by noting that *dunce's A* can be Q, then more ought to seem possible than in fact does. It should seem, not only that this brown table could have turned out to be icy, but that it could have turned out to be icy-*looking*, that is, white—for there is (we are assuming) a world where white tables cause the same sort of experience as this brown table causes in me. Similarly the Eiffel Tower should seem like it could have turned out to be three feet in height. For again, a reduced Tower should present to similarly scaled-down observers the same narrow appearances as I enjoy of the real Tower here.

What is the bearing of all this on Kripke's arguments against the mind–body identity theory? Kripke holds that any supposed identities between mental states and physical ones "cannot be interpreted as analogous to that of the scientific identification of the usual sort, as exemplified by the identity of heat and molecular motion" (1980: 150). This is because the model that explains away contrary appearances in the scientific case is powerless against the appearance that pain can come apart from c-fiber firings. Which is more plausible, that the model should suddenly meet its match in illusions about pain and c-fiber firings, or that the model fails to explain away anti-materialist intuitions because those intuitions are correct?

This argument rests on a false assumption: namely, that dualist intuitions, if mistaken, would be the sole holdouts against the epistemic counterpart model of illusions of possibility. The model breaks down already in scientific cases like the illusion that this heat could exist without HME (and vice versa).[15] One need not know how exactly the scientific illusion arises to suspect that a similar mechanism might be behind the corresponding illusion about pain.

I do not say the cases are analogous in every respect. The disanalogy stressed by Kripke is this: Identity theorists about heat can concede the existence of a world *v* where HME gives rise to sensations of cold. Materialists cannot, however, concede the existence of a world *w* where c-fiber firings are not felt as pain, because not to be felt as pain is not to *be* pain.

But this puts the materialist at a disadvantage only if we assume that *v* is what it takes to explain why this cold seems like it could have been HME, and *w* is what it takes to explain why this non-pain—this pleasure, say—seems like it could have been c-fiber firings. And my claim has been that intuitions like this cannot be explained by *v* and *w* at all—*unless* their HME and c-fiber firings are such as to feel the relevant ways to us as we are.[16]

The materialist may seem still at a disadvantage, for the following reason. How other-worldly HME feels, we know. It feels hot. But whether other-worldly c-fiber firings are bound to present as pain is not clear. Certainly if they *are* pain, then insofar as it is essential to pain to feel a certain way, that is how c-fiber firings are bound to feel. But what if we suppose with the dualist that c-fiber firings are not *identical* to mental states but *cause* them? The c-fiber firings in *w* might affect minds (ours included) differently than the c-fiber firings here.

I think we should grant Kripke that a world like *w, if it existed*, would explain the dualist intuition, at the same time as it verified that intuition. But that is just to say that the intuition would be well explained by *w* if it were correct, which

[15] One doesn't notice this because Kripke lowers the bar, dropping the facsimile requirement at precisely the point that it threatens to make a counterpart-style explanation unavailable.

[16] Of course there may be other reasons to think *v* exists, e.g., the well-attested phenomenon of the same stimulus causing different perceptual reactions in different perceivers. There are not to my knowledge any well-attested phenomena to suggest the possibility of a world like *w*.

does nothing to show that it is correct. The premise Kripke needs is that we *still* find ourselves with reason to postulate *w* even if we suppose for reductio that it is the identity theory that is correct; this is what supposedly makes materialism a self-undermining position.

But the stronger premise, we have seen, is false. This suggests to me that Kripke's argument is not in the end successful.

Does this make me a pessimist about conceivability evidence? Not at all. It does put me at odds with

(O) carefully handled, conceivability evidence can be trusted, for if impossible *E* seems possible, then something else *F* is possible, such that we mistake the possibility of *F* for that of *E*.

But although this was called the optimistic thesis above, a better term might have been super-optimistic or Pollyannaish—because for a type of evidence to *never* mislead about its proper object (the real possibility confusedly glimpsed, in this case) is exceedingly unusual and perhaps unprecedented.[17] The thesis we want, I think, is that

(O′) carefully handled, conceivability evidence can be trusted, for when impossible *E* seems possible, that will generally be because of distorting factors that we can discover and control for.

Kripke's first great contribution to conceivability studies was to have seen the need for a technology of modal error detection in the first place. His second great contribution was to have made a start at developing this technology. There is no need to foist on him a third "contribution" of identifying the one and only way modal illusions can arise.

[17] Berkeley suggests a similarly Pollyannaish thesis about perception in *Three Dialogues between Hylas and Philonous*.

> *Hylas*: What say you to this? Since, according to you, men judge of the reality of things by their senses, how can a man be mistaken in thinking the moon a plain lucid surface, about a foot in diameter; or a square tower, seen at a distance, round; or an oar, with one end in the water.
>
> *Philonous*: He is not mistaken with regard to the ideas he actually perceives; but in the inferences he makes from his present perceptions. Thus in the case of the oar, what he immediately perceives by sight is certainly crooked; and so far he is in the right. But if he thence conclude, that upon taking the oar out of the water he shall perceive the same crookedness . . . he is mistaken . . . his mistake lies not in what he perceives immediately and at present, (it being a manifest contradiction to suppose he should err in respect of that) but in the wrong judgment he makes concerning the ideas he apprehends to be connected with those immediately perceived. (3rd Dialogue)

Where the Kripkean super-optimist treats seeming failures of imagination as failures of interpretation, the Berkeleyan one shifts the blame rather from experience to inference. The insistence that there are severe, a priori discoverable, limits on our liability to ~~make~~ mistakes about a subject matter often goes hand in hand with idealism about that subject matter. This seems to me a further reason not to associate Kripke with the super-optimistic thesis (O).

## REFERENCES

Berkeley, George (1979). *Three Dialogues between Hylas and Philonous*. Indianapolis: Hackett Publishing Company.

Della Rocca, Michael (1996). "Essentialists and Essentialism". *Journal of Philosophy* 93: 186–202.

——— (2002). "Essentialism versus Essentialism". In Tamar Szabó Gendler and John Hawthorne, eds., *Conceivability and Possibility*, Oxford: Oxford University Press, 223–52.

Hill, Chris (1997). "Imaginability, Conceivability, and the Mind–Body Problem". *Philosophical Studies* 87: 61–85.

Kripke, Saul (1977). "Identity and Necessity". In Stephen P. Schwartz, ed., *Naming, Necessity, and Natural Kinds* (Ithaca: Cornell University Press).

——— (1980). *Naming and Necessity*. Cambridge, Mass.: Harvard University Press.
● Q1 •pp. 66–107.

Nagel, Thomas (1974). "What Is It like to Be a Bat?". *Philosophical Review* 83: 435–50.

The pagination is for "Identity and Necessity"

## Queries in Chapter 5

Q1. Author edit not clear.

# 6

# Beyond Rigidification: The Importance of Being *Really* Actual

Rereading *Naming and Necessity* in the light of later, two-dimensional, developments, it can seem that Kripke was not playing fair in his critique of Frege's sense theory.

The sense theory for our purposes says that with each namelike expression *n* is associated a collection of properties. The namelike expression is linked to the properties in three ways:

| | |
|---|---|
| *modally* | being *n* goes **necessarily** with having the properties |
| *epistemically* | being *n* goes **apriori** with having the properties |
| *conceptually* | being *n* goes in **understanding** with having the properties |

Each of the links leads us to expect a phenomenon that turns out not to obtain:

| | |
|---|---|
| *modal* | Water is possible without hydrogen. |
| *epistemic* | Cats are as an apriori matter small furry animals. |
| *conceptual* | Nothing counts as Peano unless it discovered Peano's Axioms. |

Because of these false predictions, the sense theory is rejected, and a new theory, or picture, is put in its place.

This paper has not previously been published. It was given at the 3rd *Barcelona Workshop on Issues in the Theory of Reference* (2003)~~, and is rendered here more or less as presented there; hence the perhaps overly conversational style. I owe~~ thanks to Mark Johnston, Scott Soames, Hagit Benjabi, Joan Weiner, and Brian Weatherson for ~~their~~ comments and criticism. caps

But the ink is hardly dry on this critique when Kripke turns around and points out other phenomena, *also* predicted by the sense theory, that *are* in his view genuine ~~there~~:

| | |
|---|---|
| *modal* | Watery stuff is possible w/out hydrogen. |
| *epistemic* | A meter is as an a priori matter the length of stick S. |
| *conceptual* | Nothing counts as heat unless it feels a certain way. |

Not only are these predictions *correct* by Kripke's lights, his own account of them, in terms of reference-fixing descriptions, bears a certain resemblance to the rejected explanation in terms of sense.

That some of the phenomena Fregeans point to do obtain on Kripke's theory, and are explained in broadly analogous ways, could make a person suspicious. (Not me! I am channeling a perspective that I do not share.) Perhaps Kripke's radical-seeming conclusions are a function less of his evidence than the order of presentation. A more logical approach, one might think, would be to first use the sense theory's *true* predictions to motivate the theory, then bring in its false ones as a guide to the theory's proper development. Senses should be chosen with an eye to the importance of not falling into these particular traps.

This way lies the two-dimensionalist reimagining of Kripke, first convincingly elaborated in Davies and Humberstone's "Two Notions of Necessity" (1980). Names *are* constitutively linked to property clusters, on this view, only not the ones we'd supposed.[1] Sometimes the false predictions reflect just a bad choice of *associated properties*. 'Cat' means something more like 'whatever shares deep explanatory features with *these* things'. 'Peano' means something more like 'whoever the people I learned the word from were talking about'. This is what's going on with the epistemic and conceptual problems.

Other times, however, the false prediction shows that we have misjudged *the character of the association*. Being water goes with being the transparent, potable stuff not across all counterfactual worlds, but all worlds "considered as candidates for actuality"—all *counteractual* worlds. This is addressed by switching to 'the *actual* so-and-so'. The description is rigid in one dimension, since the actual so-and-so *would* have been one and the same whatever world *had* obtained. (That takes care of the modal mis-prediction.) But along another dimension, it refers to whatever turns out *actually* to have the properties, supposing for argument's sake that the given world is actual.

I have called this a reimagining of Kripke. Not everyone sees it that way. The 2D line has taken on such an air of inevitability of late that it can seem,

[1] From here on I use 'name' (very!) broadly to cover any expressions that do not in Kripke's view have Fregean senses.

at times, that what separates Kripke's position from later developments is just that Kripke is more confusing. I find it all the more interesting, then, that this is not Davies and Humberstone's attitude at all. They see the 2D view as distinct from Kripke's; indeed, they point to an issue about language use that resolved one way supports the 2D view and resolved another way supports Kripke. It is because they are not sure how to resolve this issue that they describe themselves as "not confident that the suggested [2D] view is correct" (p. 20).[2] This is important, since if two-dimensionalism is correct no matter *how* we talk, then the view is lacking in substantive content. Davies and Humberstone are, to my knowledge, the last two-dimensionalists to associate the view with a potentially falsifiable claim about English. One ought to be grateful, for they are giving us here a rare opportunity to see what its substantive content might be.

What is the issue that Davies and Humberstone are not sure how to resolve? They start by looking at the cases where the 2D account works best: descriptive names *à la* Gareth Evans. Part of what makes 'Julius' a descriptive name, on their reading of Evans, is that

> One can understand sentences containing 'Julius' without knowing *of* any object that it is being said to be thus and so. (p. 7)

It will be hard for the name to retain this feature, they think, if it comes into everyday use.

> Imagine that every speaker of the language . . . had a visual confrontation with Tom and was told 'This man is Julius.' . . . Given the knowledge which each speaker would now have (knowledge by *acquaintance*) of Julius it would be natural for the semantical function of 'Julius' to change. (p. 20)

This is (I suppose) because knowing Julius just as the zip-inventor no longer suffices for understanding, once his identity becomes known. "Run, Julius has a gun!" someone says. An out-of-the-loop Evans student grabs your arm: "I don't get it—we're to run from whoever turns out to have invented the zip?" His failure to realize that we are to run from a certain particular person, not just from the zip-inventor whoever it might be, shows that he doesn't fully appreciate what "Julius has a gun" means. This is why it might seem that acquaintance with the referent destroys 'Julius''s career as a descriptive name. But what's bad for the goose is bad for the gander:

> . . . consider the fact that practically every speaker of our language *has* had a visual confrontation with (a sample of) the chemical kind $H_2O$ accompanied by the words 'This stuff (this chemical kind) is water (is called 'water')'. Is it not unlikely that 'water' remains, in our language, a merely descriptive name of $H_2O$? (p. 20)

[2] Later: "it is no part of our position that the suggested view is the ultimately correct view of the way 'red' functions in English" (p. 22).

If water is known to us as this familiar stuff of our acquaintance, then when someone says "water is refreshing", we know that this familiar stuff of our acqaintance is being called refreshing. Suppose a Martian chemist walks in who knows water only by description. She drinks a glass of water and says, "Mmm, that's refreshing. I wonder if water is refreshing?" That might ~~again~~ appear to mark her as not fully understanding the word.

The claim is that insofar as water is known to us as this familiar stuff of our acquaintance, it will be hard for 'water' to be a descriptive name. Contraposing, if water is to be a descriptive name, water had better *not* be known as this familiar stuff of our acquaintance. It had better be that

> physical ostension of a sample of $H_2O$ accompanied by the words 'this stuff' . . . is similar not to physical ostension of a man accompanied by the words 'this man' but rather to physical ostension of a screen accompanied by the words 'the man behind this screen'. (p. 20)

Of course, the analogy here—pointing to water is like pointing to ~~whatever~~ whoever is behind the screen—is strained at best. It's unclear why identifying water ostensively as 'this stuff here' should be compared to ostending Julius indirectly as 'whoever is behind the screen'. I take it that this is part of the reason why D & H are "not sure the suggested view is correct".

Suppose we agree with D & H that 'water' is not a descriptive name if the referent is known as *that familiar stuff of our acquaintance*, as opposed to *whatever lies behind water-appearances*. This still doesn't tell us why D & H are worried that 'water' doesn't have a 2D semantics. That 'water' doesn't satisfy Evans's *definition* of a descriptive name, given our acquaintance with its referent, is no doubt interesting. But the question is why 'water' would stop *behaving* like a descriptive name—the way 2D semantics *says* a descriptive name should behave—when we become acquainted with its referent.

Here is a story that seems of the right general type, drawing on work of Jim Pryor and John Campbell.[3] Judgments can be made by way of other judgments. I might judge that the President is holding a dog by judging that Bush is holding a dog, in the belief that Bush is the President. What is special about 'Julius' is that *the one and only way to judge that Julius is F is by judging that the zip-inventor is F, in the belief that Julius is the zip-inventor*. The minute we learn how to go in the other direction, judging that the zip-inventor is F by judging that Julius here is F, in the belief the zip-inventor is Julius, 'Julius' ceases to be a descriptive name. That of course, is what happens when we meet the guy; we see (and judge) that Julius is (say) drunk and judge thereby that the zip-inventor is drunk.

Now, why should descriptive namehood as just explained have the result that '*n*' refers on all counteractual hypotheses to whatever is actually so-and-so, and why should loss of descriptive namehood interfere with that result? When it

[3] Pryor (1999); Campbell (1999).

comes to deciding whether Julius is so-and-so, I am doing something that is of its nature done when and because one decides the zip-inventor is so-and-so. This applies in particular to deciding whether Julius is drunk in such-and-such a counteractual world. If you ask me how I decide whether Julius is on a given counteractual hypothesis drunk, there is only one possible answer: I decide whether the zip-inventor is drunk on that hypothesis.

Suppose on the other hand that I have other ways of deciding that Julius is so-and-so. Suddenly there is the possibility of, as we might put it, original intelligence about Julius—I realize Julius is so-and-so not by first realizing something else—and also "variably derivative" intelligence—intelligence based on other descriptions Julius is thought to satisfy. Now I cannot rest my decision purely and simply on the issue of whether the zip-inventor is on the given hypothesis drunk. On the one hand, here is a guy with the same parents as Julius (= Tom), who looks and acts just like Julius and leads a very similar life, and he is drunk. On the other hand, here is the guy who invented the zip on this hypothesis and he is not drunk at all. Is Julius drunk on the given hypothesis or not? It ~~as~~ is not as though I have rank-ordered Julius's traits so as to know which way to jump when and if the traits come apart. Unless our grasp of a name takes a very particular and unnatural form—a form that precludes independent lines of sight on the referent—judgments about Julius's counteractual state are bound to be problematic.

It makes sense, then, that the two-dimensionality of a term should stand or fall with the uniformly *derivative* character of judgments about the term's referent; they are always reached via the same descriptive route. How is it to be determined whether judgments about water, heat, and so on are uniformly derivative in that way?

Certainly ~~One could try to argue directly that~~ speaking of water *feels* very different from speaking about an I-know-not-what hidden behind some veil of appearances. But the argument for a two-dimensional interpretation was never that it *feels* right. The argument was that it *explains* a lot, and on a more economical basis. This brings us back around to the modal, epistemic, and conceptual phenomena with which we began. The rigidifier says: look, I can explain these phenomena just as well as Kripke and with a lot less fuss and bother.

● Q2

## WHOSE EXPLANATIONS ARE BETTER?

● Q1 • Our job as friends of D&H's skeptical side is to ask whether this explanatory advantage claim is true. It isn't, we'll be arguing. We'll be arguing, in fact, that the rigidifier's explanation is oftentimes worse than Kripke's, and worse in ways plausibly blameable on the treatment of water-judgments as of their nature made by way of descriptive judgments. There will be three kinds of worseness involved. Depending on the case, we'll be saying to the 2D explainer either that (1) you

EXTRA #

are using a cannon to kill a mouse; (2) you are hitting a lot besides the mouse; or (3) you have missed the mouse.

Obviously it's the type-(3) criticisms that are the most interesting, so let me say briefly how the mouse is liable to escape. According to me, the rigidifier's interpretation of "actually" makes certain sorts of concept inexpressible. One cannot in the 2D framework express concepts whose extension is tied to what is *really* actually the case, as opposed to what might be *hypothesized* to be actually the case. The rigidifier may thus wind up explaining the wrong datum—one in which our actual concept of thus and such has been replaced by some 2D surrogate.

FN:4 So, to mention an example that will come up later, we have the concept of a "spitting image" or "look-alike" of something now under observation.[4] This is used in turn to explain other concepts. By fool's gold we mean a certain kind of look-alike of real gold.

Why is the concept of fool's gold inexpressible in a 2D setting? Because to be fool's gold is to *look to us as we are* like gold, not to us as we might be hypothesized to be. To see this, consider ~~y~~our attitudes towards fool's gold ~~so described~~. What am I worried about, ~~for instance,~~ when I worry that this supposedly gold ring might perhaps be made of fool's gold? Is the worrisome hypothesis that the ring might be a substance like iron pyrites that I *as I am* cannot tell apart from real gold? Or is it that the ring might be a substance like charcoal that looks to me *as I am hypothesized to be* like gold, because (on the worrying hypothesis) gold looks to me a dull black? The first answer is clearly the right one. I am worried I am in a world where the ring "looks golden" despite not being gold. I am not worried I am in a world where the ring looks ~~however~~ gold looks to me *in that world* whatever way without being gold. Lacking a concept of real actuality—of how the ring looks to me, not as I am hypothesized to be, but as I am—the two-dimensionalist cannot express what I am worried about.

At least, the two-dimensionalist cannot express it directly, by invoking the (truly) actual world as such; for 2D actuality is by nature shiftable, by moving to a new hypothesis about which world is actual. Another option would be to try to pick our world out by description. One could determine *empirically* that this world has certain ~~relevan~~t features, and write those features explicitly into the definition of 'fool's gold'. For instance, fool's gold is whatever looks like gold to observers with X-type brains (a particular empirically determined sort of brain). But this, as Kripke says, "makes the definition into a scientific discovery". rom

[4] Compare the concept *audible* to the concept *plausible*. Both arguably have constitutive links to a certain kind of response. *Audible* seems amenable to 2D treatment. Whatever observers can directly hear, on a given hypothesis, is on that hypothesis audible. However, it is not the case that whatever observers find plausible, on a given hypothesis, really is plausible on that hypothesis. Imagining that we all find Scientology—its content and the evidence for it unchanged—plausible is not imagining it to really be plausible. (There are connections here with the problem of imaginative resistance.)

It is an empirical, not definitional, truth that fool's gold has certain effects on people with X-type brains. If one insists that the above is a definition, then it is a definition of some other term, one that sounds like "fool's gold" but expresses a different concept. If the two-dimensionalist took this route, he could fairly be charged with changing the subject and explaining the wrong phenomenon.

single quotes

Here is another example of "explaining the wrong phenomenon", this one to do with illusions of possibility. It seems possible that gold could have had a different atomic number. The illusion can be explained in 2D terms, *if* by "different" one means "different from 79". (Supposing a certain world *w* to be actual, gold has an atomic number of 80.) But suppose we don't *know* gold's atomic number, and the illusion is rather that gold's atomic number is contingent; gold could have had a different atomic number than it does have. This seeming possibility the two-dimensionalist cannot so easily explain—for it makes no sense to say that if *w* is actual, then gold has a different atomic number than it has actually. (Compare the *de dicto* reading of "I thought your yacht was longer than it actually is.") The closest he can come is to explain the illusion that it could have had a different atomic number than 79. But that is a different illusion. So there is a second example of explaining the wrong phenomenon, what above I called missing the mouse.

A third, perhaps more controversial, example, relates to Davies and Humberstone's suggestion that the 2D meaning of '*x* is red' might be *x has that physical property which actually standardly ~~causes~~ produces red* sense data in perceivers* (p. 22). They say, "we find at fn 71 [of *Naming and Necessity*] . . . a very clear anticipation of the present suggestion for secondary quality words" (p. 28). What Kripke says in footnote 71 is that

> the reference of 'yellowness' is fixed by the description 'that (manifest) property of objects, which causes them, under normal circumstances, to be seen as yellow (i.e. to be sensed by certain visual impressions . . . )'.

But what is meant by "normal" circumstances here? It seems to me that Kripke does *not* mean "actually normal" circumstances where "actually" is used in the shifty 2D way. He does not mean whatever circumstances we might *imagine* to be normal in the course of imagining the reference-fixing description deployed in alternative settings.[5] One piece of evidence is that Kripke never offers a criterion C for normal circumstance-hood; he never tells us, as it were, what screen to look behind. Another is that he questions whether such a criterion is possible:

FN:5

> If one tries to revise the definition of 'yellow' to be, 'tends to produce such and such visual impressions under circumstances C' . . . one will find that the specification of the circumstances C either circularly involves yellowness or plainly makes the alleged definition into a scientific discovery rather than a synonymy. (1980: 140)

[5] He is not, for instance, thinking that white paper falls into the extension of 'yellow' in worlds where it is normal to view objects under yellow light.

This recalls our point above about fool's gold. Bringing neurological findings into the definition of 'fool's gold' makes it no longer a definition but a scientific discovery about the referent; construed as a definition it defines not our term 'fool's gold' but a hitherto unknown homonym. Likewise, bringing empirical findings about normality into the definition of 'yellowness' gives us either a scientific discovery about yellowness or a definition not of 'yellowness' but of a new term spelled the same way.

Then what *does* Kripke mean by 'normal' in that footnote? I suspect the Kripkean notion has a demonstrative element; we presume that *these* conditions—the ones that mostly obtain around here—are normal. Normal conditions are like water. They're those familiar conditions of our acquaintance; we recognize them when we see them. The rigidifier, lacking a concept of real-actuality, cannot follow Kripke in this. He will have to specify normal conditions *descriptively* and without reliance on the concept of yellow. He certainly doesn't know how to do this from the armchair, so he will have to do an *empirical* study of viewing conditions, including the conditions that obtain inside our heads. But this, as Kripke says, "makes the alleged definition into a scientific discovery". If you want it to be a definition, it's a definition of shmolor concepts, not color concepts. The modal, epistemic, and conceptual phenomena as they arise for *our* concepts will be left unexplained.

## EXPLAINING THE *CONCEPTUAL* DATUM

Both sides agree that it can sometimes be important to the understanding of a name '*n*' to realize that the referent should have certain properties. But they offer different explanations of this, according to their different views of meaning. (I assume that understanding is in some sense knowing the meaning.) The rigidifier maintains that '*n*' 's meaning is the same as that of 'the actual G', which comes to the fact that '*n*' stands no matter which world is actual for whatever is actually G there. It would seem then that

(2DU) Understanding '*n*' is knowing that no matter which world is actual, *x* is *n* iff *x* alone is actually G.

Given this, the rest is a slam dunk. Knowing that '*n*' stands for the unique G *no matter what* is certainly sufficient for knowing that a thing should be G if it wants to be *n*.

Since for Kripke the meaning is just the referent, understanding for him comes (so I assume) to knowing what the name stands for. This might sound like saying that to understand is to know of the appropriate *x* that '*n*' stands for *x*. But I suspect that is not Kripke's view. A couple of passages suggest what more might be involved.

[If a Martian]

> ~~[I]f someone else~~ detects heat by some sort of instrument, but is unable to feel it, we might want to say, if we like, that the concept of heat is not the same even though the referent is the same. (p. 131)
>
> [A] blind man who uses the term 'light', even though he uses it as a rigid designator for the same phenomenon as we, seems to us to have lost a great deal, perhaps enough for us to declare that he has a different concept. (p. 139)

The Martian has a defective or eccentric understanding of 'heat', but why? It is not, I think, that the Martian fails to know of any *x* that 'heat' stands for *x*. For we can suppose she senses heat some other way; she can *see* it, let's say, with her telescopic vision. Just as we know of the phenomenon *x* that we feel that 'heat' stands for *x*, she knows of the phenomenon *x* that she is looking at that 'heat' stands for *x*. But Kripke would still, I think, say that "her concept of heat is not the same even though the referent is the same" (p. 131). The Martian's problem is not that she fails to know of the correct *x* that it is the referent of 'heat'.

By one's *idea* of heat, let us mean whatever it is in one's head that enables one to form thoughts about heat so described: thoughts of the sort one would express by saying "heat is so-and-so". The Martian certainly has an idea of heat, for she has thoughts to the effect that "heat looks like a bunch of rapidly vibrating particles". So what is she missing?

Proposal, meant to be in a Kripkean spirit: What sets the Martian apart is that her heat-idea is *abnormal*. All of *our* heat-ideas have certain properties in common that the Martian's idea lacks. To know what 'heat' stands for is to know that it stands for heat, where heat is conceived not by any old ~~idea of heat~~ but a *normal* ~~*idea*~~.[6] I will call this *knowing in the normal way* that 'heat' stands for heat. Putting this together with the claim about understanding, we get

heat-idea FN:6 heat-idea

(KRU) Understanding '*n*' is knowing in the normal way that '*n*' stands for *n*.

So, for instance, I might acquire the word 'Mt Everest' by being told it stands for the world's highest mountain, located somewhere in Asia, or 'the Sun' by being told that it stands for *that*, the shiniest object in the sky. Something like this is, let's assume, the normal idea of Mt Everest, or of the Sun. ~~In my case, and I would~~ assume in ~~yours~~, understanding 'Mt Everest' ('the Sun') is knowing that it stands for Mt Everest (the Sun) as thus normally conceived.

caps

How does this compare to what the rigidifier requires for understanding? Both sides agree, let's say, that I am expected to know that 'Mt Everest' stands for Mt Everest, conceived as the highest mountain. The difference is that (KRU) is content if I know this is true *as matters stand*. (2DU) says I should know it unconditionally, that is, *no matter which world is actual*. Do I?

— Q3

Given that my understanding of 'Mt Everest' comes entirely from ~~the~~ teacher's explanation, I know Mt Everest is the tallest mountain *no matter which world*

my

[6] Crimmins (1989).

*is actual* only if my teacher has told me that it is. But she ha~~d~~ s told me only that being Mt Everest *does* go with being the tallest mountain. Was the stronger claim perhaps implicit? It would seem not. She would be shocked and horrified own students to hear me telling my ~~brother~~, “~~oh, by the way,~~ if Kanchenjunga should turn so out to be tallest, then ‘Mt Everest’ stands for Kanchenjunga”. Her message is this: “presuming I am not greatly mistaken about which mountain is tallest, ‘Mt Everest’ stands for the tallest mountain.” Similar remarks apply to ‘the Sun’. “/” No one’s understanding of “the Sun” tells them it is Sirius B if we are massively deluded and Sirius B is the star responsible for the appearances by which we identify the Sun.

FN:7 This shows, I think, that the 2-D picture of understanding is in one way[7] much more demanding than the Kripke picture, and on the face of it more demanding colon ● Q4 than the truth. The next question is; is any of the additional expertise imputed by the 2D picture actually *needed* to explain the phenomenon of associated properties?

A reason to doubt it is this. The phenomenon to be explained has to do with *necessary* conditions on the referent: to be yellowness, a property should look a certain way; to be 100°C, a temperature should be the boiling point of water at sea level; to be the Sun, a thing should be the shiniest object in the sky. Someone who doesn’t expect the referent to have these properties doesn’t understand the term as we do. But of course, it is one thing to think that *if x is the referent, it needs to have certain properties*, another to think that *if x has those properties, it needs to be the referent*. conditional puts The first ~~concerns~~ necessary conditions on the referent, the second offers sufficient conditions for being the referent. When the two-dimensionalist insists that no matter which world is actual, yellowness is whatever feels a certain way, she is talking (at least) about *sufficient* conditions. She is thus like the imaginary counterpart of my teacher who says, *I don’t care which mountain turns out to be tallest, that’s the one we call ‘Mt Everest’.*

Not only is this extra instruction irrelevant to the explanation of our intuition that to be heat, say, a thing should feel like *this*, it “explains” an intuition that rom we *don’t* have: viz. that if, evidence to the contrary notwithstanding, it is not fire and soup that feel the relevant way but snow and Jell-o, then it is the condition of these latter that we have in mind by ‘heat’. When I identify heat as what feels like *this*, standing before a fire, I mean what *does* feel like this, given what I know about how various things feel. (If I learn the word from a teacher, her message is not “heat is whatever presents like so, and now I cast my fate to the

[7] *Less* extravagant than KRU, and than the truth, in another way. 2DU doesn’t require you to know what ‘*n*’ stands for. After all, it’s your ignorance of this that’s supposed to explain how you can understand ‘water’ without realizing it stands for $H_2O$. But this “ignorance of the referent” is a tendentious redescription of ignorance of some of the referent’s essential properties. Why should you need to know the essence of water to know that it’s what the word ‘water’ stands for? If that were the requirement, then I don’t know what my own name stands for. (Thanks here to Brian Weatherson.)

winds", it's "heat is what presents like so, presuming, as why shouldn't I, that I am not totally misremembering or otherwise mistaking my actual perceptual reactions".)

So far we've had an example of using a cannon to kill a mouse—using a necessary and sufficient connection when the explanation draws only on the necessity—and an example of hitting some neighboring mice—"explaining" a cast-our-fate-to-the-winds intuition we don't actually have. Next an example of missing the mouse.

Suppose Kripke is right that the Martians have a different concept of heat if they don't feel it as we do. How does the rigidifier propose to explain this? Well, the Martian does not know that no matter which world is actual, 'heat' is what causes heat-sensations. A problem which I won't be discussing is, why *can't* the Martian know this? It's not as though you need to actually have a feeling to know a fact in which it figures. (The blind certainly know that light gives rise to visual impressions, and this no doubt plays a role in their understanding of the term.) And anyway, it may be that the Martian *does* have the feeling, but in response to cold things rather than warm ones.

The problem I ~~do~~ want to discuss has to do not with the Martians' understanding of 'heat' but our own. If what a person knows whereby they understand 'heat' is that it stands for whatever feels a certain way, then this should be knowable without a prior understanding of 'heat'. But then the feeling by which heat is identified had better have a name other than 'feeling of heat'. Kripke says the following:

> . . . heat is something we have identified (and fixed the reference of its name) by its giving us a certain sensation, which we call 'the sensation of heat'. We don't have a special name for this sensation other than as a sensation of heat. It's interesting that the language is this way. Whereas you might suppose it, from what I am saying, to have been the other way. (p. 131)

(You might indeed.) And later,

> Some philosophers have argued that such terms as 'sensation of yellow', 'sensation of heat', 'sensation of pain', and the like, could not be in the language unless they were identifiable in terms of external observable phenomena, such as heat, yellowness, and associated human behavior. I think that this question is independent of any view argued in the text. (p. 140)

How *could* it be independent, one might wonder? There is a problem here if the role of a reference-fixing description is to specify the referent in prior and independent terms, thereby conferring understanding. But it is only the rigidifier who assumes that understanding is constituted by knowledge of a reference-fixing biconditional. According to Kripke, as we are reading him, one understands '*n*' by

(1) knowing of the right *x* that '*n*' stands for *x*, while
(2) conceiving of that *x* via a normal idea.

A reference-fixing description can contribute in the first connection by specifying in independent terms which *x* is being referred to; that is how initial baptisms are supposed to work. But it can also contribute in the second connection by reminding us of what counts as a normal idea of *x*. (“It might here be so important to the concept that its reference is fixed in this way . . . ” (p. 131), “The way the reference is fixed seems overwhelmingly important to us in the case of sensed phenomena . . . The fact that we identify light in a certain way seems to us to be *crucial*, even though it is not necessary” (p. 139).) Whether or not it is circular to use the word ‘heat’ in identifying the referent *x* of that very word, it is clearly *not* circular to use the word ‘heat’—which we do after all understand—in our account of how we who understand are expected to conceive of its referent.[8]

## EXPLAINING THE *EPISTEMIC* DATUM

Now let’s consider the rigidifier’s explanation of a priori truths about Neptune, say, or the length *a meter*. It is a priori, we are told, that a meter, supposing there is such a length (the definition has not misfired), is the length of this stick. Both sides agree, I think, that the apriority reflects *something like* immunity to error through misidentification,[9] so let’s talk about that. Error through misidentification happens[10] when

one correctly supposes that *n* and the G exist,
but one wrongly supposes that *n* is the G.

Our judgment that *n* is G is *immune* to error through misidentification if there is no chance at all of this happening; assuming both exist, *n* is bound to be the G. So it is with the reference-fixer’s judgment that “a meter is the length of this stick”. There is no chance at all that there is a length *one meter* and something is *the length of this stick*, but a meter is not the length of this stick. Given how the phrase was introduced, it is this referent or nothing. ‘A meter’ has no other option if it wants to refer.

That much, it seems, Kripke and the rigidifier can agree on. They disagree, though, about *why* ‘a meter’ has only one option if it wants to refer. What is it about our understanding of ‘*n*’ that makes it the case that

(*) ‘*n*’ refers if it does to the G?

[8] Crimmins thinks normality has statistical and normative aspects. It can involve non-representational properties of the idea, say, that it calls up certain associations, or that it is triggered by a certain external phenomenon, as our heat-idea is triggered by heat. Another possibly important feature of normal ideas is that we are *tempted* to ascribe certain properties to the referent so-conceived, like indivisibility to atoms or Tarskian properties to truth. Crimmins (1989) has the fascinating and underappreciated details.

[9] Pryor (1999); Campbell (1999).

[10] Consider this stipulative. See Pryor (1999) for two kinds of error through misidentification.

The rigidifier's story is based on

FN:11 (2DU) To understand '*n*' is to know that NMWWIA,[11] '*n*' refers to *x* iff *x* is the actual G.

This lets her reason her way to (*) as follows:

(1) '*n*' refers
(2) '*n*' is understandable, say by X.
(3) X knows that NMWWIA, '*n*' refers to *x* iff *x* is the actual G.
(4) NMWWIA, '*n*' refers to *x* iff *x* is the actual G.
(5) '*n*' refers to *x* iff *x* is the actual G.
(6) '*n*' refers to *x* only if *x* is the actual G.

Conditional proof gives "if (1), then (6)" which then by basic logic allows the rigidifier to derive (*).

This explanation is not very efficient. To put the point schematically, it derives the conclusion *R only if G* from the premise (among others) that *X knows that NMWWIA, R if and only if G*. An explanation is available at cheaper rates from the Kripkean. Remember the Kripkean theory of understanding:

(KRU) To understand '*n*' is to know in the normal way (using a normal idea) that it stands for *n*.

To understand 'a meter', on this theory, is to know in the normal way—using a normal idea—that it stands for a meter. Now at this point, our *only* idea of a meter is as the length of this stick; so any idea that doesn't figure in a knowledgeable belief that a meter is the length of this stick would have to count as abnormal. And that is just to say that we can't understand 'a meter' without knowing that it stands for the length of this stick. More generally, if our *only* idea of *n*—hence the one whereby we understand '*n*'—is the one figuring in our knowledge that *n* is the G, then the Kripkean can argue as follows:

(1) '*n*' refers.
(2) '*n*' is understandable, say by X.
(3′) X knows that '*n*' refers to the G.
(4′) '*n*' refers to the G.

(*) now follows by conditional proof. Relative to the goal of explaining apriority, the surplus content of (2DU) *vis-à-vis* (KRU) is just wasted.

That was a "using a cannon to kill a mouse" type criticism. Now an example of collateral damage: that is, the rigidifier "explains" things that aren't the case. Distinguish two claims:

● Q5 (A) ●if there's such a length as a meter, it's the length of this stick.

[11] NMWWIA = no matter which world is actual.

This I have agreed is a priori, because ‘a meter’ refers to the length of stick S if to anything. Second,

● Q6 (B) ●if this stick has a length at all, then the length is a meter.

It should be clear that (B)’s apriority follows just as easily from the rigidifier’s notion of understanding as (A)’s. But is (B) in truth a priori? Recall how Kripke sets the case up:

> There is a certain length which he wants to mark out. He marks it out by an accidental property, namely that there is a stick of that length. Someone else might mark out the reference by another accidental property. (p. 55)

Since it is an empirical matter whether stick S is “the length he wants to mark out”, we need to ask what happens if he is wrong and it is a different length than intended. It might be, for instance, that the stick is a millionth of an inch long, but emitting magnification rays that delude us into seeing it as longer. Or maybe the stick is a mile long, but much farther away than anyone had realized. I take it that it is no part of the reference-fixer’s understanding of ‘meter’ that it *continues* to stand for the length of S even if S is much shorter or longer than it appears. Since this cannot be a priori ruled out, we don’t know a priori that the stick is a meter long if it has a length at all.

You can guess the rigidifier’s reply: “That just shows we have not been sufficiently careful about the descriptive condition that defines ‘one meter’. The real meaning of ‘one meter’ is ‘the length of that stick, presuming the stick is roughly as long as it looks’. I agree that this is how the answer has got to go.

But at the same time it *can’t* go that way, because of the rigidifier’s difficulties about real-actuality. The phrase “roughly as long as it looks” cannot mean “as long as it looks to our as-if actual selves”, for then the definition still goes through if the stick is a mile long in *w*, provided there are compensating changes in our perceptual system: we have telescopic vision in *w* so that it takes a mile of stick to FN:12 make true the experience that a much shorter stick answers to here.[12] The phrase has got to mean “as long as it does or would look to us as we really are”. And as we have seen, the two-dimensionalist has no way of capturing that “really”. His only option is to determine empirically that our actual perceptual wiring is

[12] I am skating over various subtleties ~~here~~. Suppose that we can characterize my experience of S: it’s a type-Z experience. Because we don’t want to make a “definition into a scientific discovery rather than a synonymy”, we must take care that the characterization (“type-Z”) not help itself to features of real actuality that concept-users may be in no position to know, features that would have to be discovered empirically. This applies, I submit, to how long in inches a thing must be to answer to my current experience. One reason is that I might not know about inches; I might not yet have any measures of length but perceptual ones. A second reason is that just as I am often surprised by how much bigger a piece of furniture is than I had guessed on the basis of its appearance in the store, I put no great stock in my guess as to the length in inches of a stick that looks to me like this one does now. That an experience is Z-type should be silent on the question of how objectively long it represents its object as being.

Y and then write Y into the definition of "as long as it looks". That may deliver the right extensional results but at the cost of changing the subject, since our concept of *as long as it looks* is clueless about the neurophysiology of vision.

## EXPLAINING THE *MODAL* DATUM

The third phenomenon in need of explanation is the seeming possibility of things that are in fact impossible. The 2D explanation of why it *seems* possible that S is that there are possible worlds *w* such that if *w* is actual, then S. It seems like Hesperus could have been distinct from Phosphorus simply because *the actual evening-visible planet* = *the actual morning-visible planet* is false on certain hypotheses about which world is actual. I want to argue that depending on how you run it, the 2D style of explanation either explains too much (collateral damage) or doesn't explain enough (misses the mouse).

Recall a key feature of Kripke's approach to illusions of possibility. Kripke doesn't just want to show how someone *could* fall under the misimpression that, say, Hesperus could have failed to be Phosphorus, by misinterpreting what was in fact a different possibility. That would be easy, since a sufficiently confused person could presumably misinterpret anything as anything. he wants to show that we plausibly *do* fall under the modal misimpression by misinterpreting a different possibility. It is not just that an intuition of *E*'s possibility *could*, but that our intuition of its possibility plausibly *is*, based on the mistaking of one possibility for another. One should be willing to say: oh, I see, once you point out the difference, it's because *this* really is possible that I supposed *that* to be possible.

The kind of principle I am relying on here is familiar from psychoanalysis. Here is what in my brief ~~(well. . . .)~~ experience psychoanalysts tell you. "You are under the illusion that nobody loves you. A cruder sort of doctor might say, here is how the illusion arises, take my word for it, now you are cured. But I would never dream of asking you take my word for it. No, the test of my explanation is whether you can be brought to accept the explanation, and to accept that your judgment is to that extent unsupported." The analogy is good enough that I will speak of the

*Psychoanalytic Standard* Assuming the conceiver is not too self-deceived or resistant, $\Diamond F$ explains *E*'s seeming possibility only if he/she does or would accept it as an explanation, and accept that his/her intuition testifies at best to *F*'s possibility, not *E*'s.

This is a high standard, but what makes Kripke's approach so convincing is that this is the standard he tries to meet, and mostly *does* meet. Philosophers have been telling us for centuries that this or that common impression is false. And we have for centuries been shrugging them off. What makes Kripke special is that

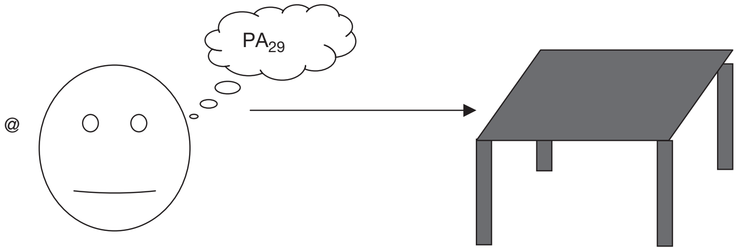

**Fig. 6.1.** "Steve looking at Table in @"

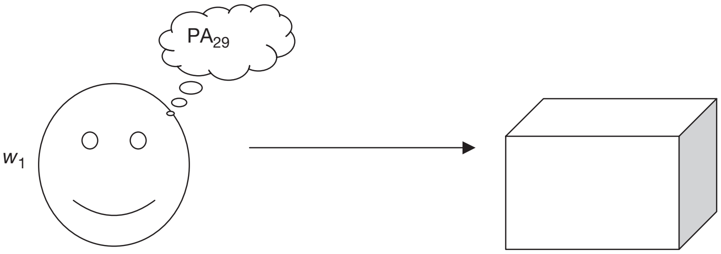

**Fig. 6.2.** "Table$_1$ looks to Steve$_1$ just like Table looks to Steve in @"

he gets you to *agree* that you are making the mistake he describes. Whether the rigidifier can get you to agree that you are making the mistake *he* describes is not so clear.

One way to see the problem is to look at Kripke's explanation of modal illusions in terms of "qualitatively identical epistemic situations": it seems possible that $x$ is P because its counterpart $x^*$ in a qualitatively identical epistemic situation really is P. What does he mean by that phrase *qualitatively identical epistemic situation*? One obvious thought is that to be in the same epistemic situation as I am ~~me~~ in now is to enjoy the same (narrowly individuated) perceptual appearances: to be, say, in perceptual appearance state PA$_{29}$.

But that seems not to be enough. Take the illusion that this table could have been made of ice. One world I am pretty sure is out there is a world $w_1$ whose Steve-character Steve$_1$ is on drugs so powerful that an ordinary old block of ice looks to him just like this table looks to me.

Does the possibility of a world like that explain why it appears to me that this table could turn out to be made of ice? I take it that it doesn't. There is no ‸Oh, temptation to say, "~~OH~~ now I see why this brown table seems like it could be made of ice; it's because there could be a guy to whom regular ice looked like this".

Well, maybe the problem with that first explanation is that Steve$_1$ is *perceptually deluded*. The way things appear to him is not how they are. A second idea, then, is that someone is in my epistemic situation if they enjoy the same (narrowly individuated) perceptual appearances *and their experience is veridical.* This doesn't work either, I think, because my doppelganger need not be deluded even if he is looking at a visibly icy table. All he needs is to be differently wired so that white things produce in him the same perceptual appearances as brown ones produce in me.

Once again, we don't think, "Aha, what seemed like the possibility of this brown-looking table being made of ice was really just the possibility of a spectrum-inverted Steve$_2$ to whom white things look the way brown ones look to me." I am not thinking, "As far as I can tell, I am in the Steve$_2$ situation," because the Steve$_2$ situation is *defined* by a contrast with mine. (It would be like worrying that this has all along been Twin-Earth.)

A third reading, which gets closer, I think, is that someone is in my same epistemic situation if the scene they are experiencing has the same perceptually available properties as this one. (This rules out the $w_2$ scenario, because brown is a perceptible property, and in $w_2$ it's missing.) But remember, Kripke also wants to explain in this way the seeming possibility of *brown* having a different physical nature than it has in fact. That will require a counterpart situation where at least one perceptible property, viz. brown, is changed. So sameness of epistemic situation cannot require sameness of perceptible properties.

I see only one other option, and it's this. Someone is in my same epistemic situation if the scene he is experiencing is a dead ringer for the scene I am experiencing, meaning that the two are for me perceptually indistinguishable. If you quickly substituted his situation for mine, keeping my perceptual systems the same, I would be none the wiser; it would not appear that anything had changed. The picture we want, then, is ^ ~~as shown~~ in Fig. ~~6.3.~~ ^6.4 ^the one

Note that the reactions of my as-if actual self are irrelevant on this picture; it's *real* me to whom the icy table has to look just like the wooden table. The feature of $w_3$ that makes it explanatory—the table there looks the same to *real*-me—is not even expressible in two-dimensionalist terms. The closest the two-dimensionalist can come is to say, it is possible for an icy table to

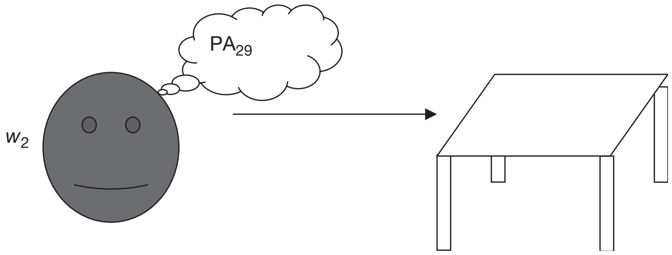

**Fig. 6.3.** "Table$_2$ *veridically* looks to Steve$_2$ just like Table looks to Steve in @"

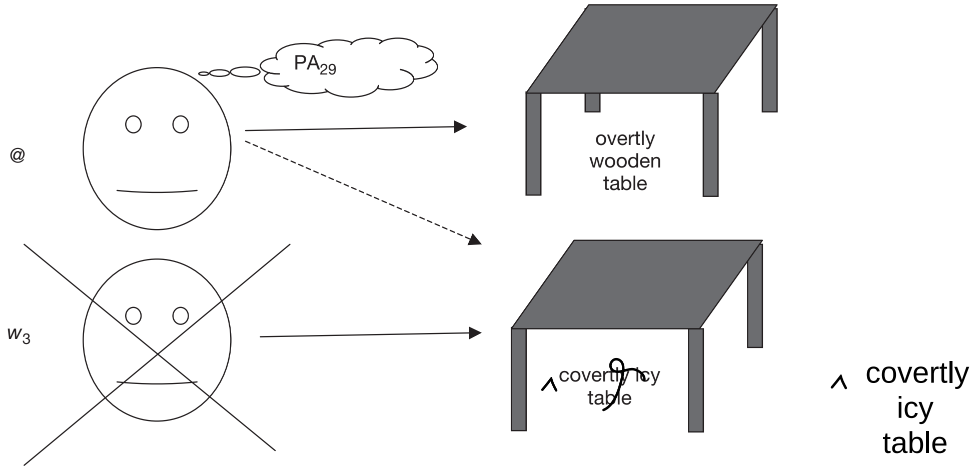

**Fig. 6.4.** "Table$_3$ is a dead ringer for Table"

produce PA$_{29}$ in someone whose perceptual system is—plug in here empirically determined features X, Y, and Z of my perceptual system. There is a world like that—it might even be $w_3$—but it can't explain *my* illusion because what seems possible to me is not that an XYZ observer mistakes this icy table for wood but that I am mistaking this icy table for wood.

So we have the following principle: to explain why *this*, an object of our acquaintance understood to present like so, seems like it could turn out to be Q, one needs a possible scenario in which something *superficially indistinguishable* from it does turn out to be Q. The counterfactual thing has to look the same, not to the counterfactual folks, but to us. I will call that a *facsimile* of the actual thing. And I will refer to the principle as the *facsimile* principle, or the *fool's gold* principle. If you want to explain in a psychoanalytically satisfying way why it seemed possible for gold to be iron pyrites, the explanation should *not* be that there's this perfectly ordinary brownish hunk of rock ("dunce's gold") ~~not looking like gold to us but~~ looking to the people around it as gold looks to us. Since two-dimensionalists cannot express facsimilehood, they drop out of the competition already here.

I said that Kripke respects the psychoanalytic standard and that his explanations often satisfy it. But it seems to me that this is one of those rare cases where the two-dimensionalist error can be traced back to Kripke. Sometimes not even Kripke has a psychoanalytically satisfying explanation. Sometimes he is forced like the two-dimensionalist to appeal to "dunce's gold" when it is fool's gold we want.

Here is some heat; is it HMME (high mean molecular energy)? One has to conduct additional tests. ~~And~~ like any tests, they could come out either way. So caps there's the appearance that heat could turn out to be HMME, and the appearance that it could turn out to be something else, say LMME. The second appearance is an illusion. Kripke would explain it away as follows:

> the property by which we identify [heat] originally, that of producing such-and-such a sensation in us, is not a necessary property but a contingent one. This very phenomenon could have existed, but due to differences in our neural structures and so on, have failed to be felt as heat. (1980: 133)

Let's say, to make it definite, that the difference in neural structures had the result that high MME felt cold, and low MME felt hot. Does this explain in a psychoanalytically satisfying way the illusion that it could have been low MME that was heat rather than high? Does it explain the illusion that heat could have turned out to be low MME to point to possible differences in our neural structures?

Here is the worry. With the table, remember, what seemed possible was not just that ice could have paraded itself in front of *someone or other* who saw it as wood, but that there could have been ice that *I with my existing sensory faculties* would have seen as wood. To explain that seeming, we needed a facsimile of the table—a spitting image of it—that was in fact made of ice. Likewise, what seems possible in the case of low MME is not just that it could have paraded itself in front of *someone or other* who felt it as hot, but that *I* with my existing neural structures could have found it hot. To explain that seeming, we need a facsimile of heat that turns out to be low MME. There should be the possibility of fool's heat which turns out on inspection to be low MME. Similarly to explain the seeming possibility of high MME turning out to be cold, we would need the possibility of fool's cold that was found on inspection to be high MME.

Is there fool's heat of this type, or fool's cold? I don't see how there could be. It may be possible to slip a cleverly disguised icy table in for this wooden one while preserving visual appearances. *But it is not possible to slip cleverly disguised low MME in for high MME and have it feel just the same.* Having substituted low MME for high, there is no other way to preserve the appearances but to postulate observers with different sensory reactions than ours. But then what we are getting is not really *fool's* heat but something more like *dunce's* heat. Because, as already discussed, you would have to be pretty confused to see in the possibility of rewiring on *your* side the explanation of why a switcheroo seems possible *on the side of phenomenon you are sensing*. Whether fool's heat is absolutely impossible I do not know. But what does seem clearly impossible is for *low MME to be fool's heat*, because it by hypothesis feels the opposite of hot; it feels cold.

Kripke is right, or anyway I'm not disagreeing, when he says that "the property of producing such-and-such a sensation in us . . . is not a necessary property", because we could have been wired differently. High MME could have produced what we call sensations of cold. But producing such and such a sensation in us is not the property of interest. The property of interest is *producing such-and-such* *a sensation in us as actually constituted*. And that property would seem to be necessary. There are only three factors in how an external phenomenon is disposed to feel: its condition, our condition, and the conditions of observation. If all these factors are held fixed, as the notion of fool's heat would seem to require, then it is hard to see how the sensory outcome can change.

## REFERENCES

Campbell, J. (1999). "Immunity to Error through Misidentification and the Meaning of a Referring Term". *Philosophical Topics* 26: 89–104

Crimmins, M. (1989). "Having Ideas and Having the Concept". in *Mind and Language* 280–94

Davies, M., and Humberstone, L. (1980). "Two Notions of Necessity". *Philosophical Studies* 28: 1–30

Kripke, S. (1980). *Naming and Necessity*. Harvard University Press

Pryor, J. (1999). "Immunity to Error through Misidentification". *Philosophical Topics* 26: 271–304

## Queries in Chapter 6

Q1. Since we have captured the above line as heading, we have captured this para with no indent. Please confirm Fine

Q2. Please check heading level. Fine

Q3. Please check and confirm whether we have to capture (,) or —here. — please

Q4. Please check and confirm the author correction here. Colon please

Q5. We have captured this in lower case. Please check and confirm. Fine (with

Q6. We have captured this in lower case. Please check and confirm. the author)

# 7

# How in the World?

*. . . the final proof of God's omnipotence [is] that he need not exist in order to save us.*

Peter De Vries, *The Mackerel Plaza*

Is it just me, or do philosophers have a way of bringing *existence* in where it is not wanted? All of the most popular analyses, it seems, take notions that are *not* overtly existence-involving and connect them up with notions that are existence-involving up to their teeth. An inference is valid or invalid according to whether or not there *exists* a countermodel to it; the *F*s are equinumerous with the *G*s iff there *exists* a one-to-one function between them; it will rain iff there *exists* a future time at which it does rain; and, of course, such-and-such is possible iff there *exists* a world at which such-and-such is the case.

The problem with these analyses is not just the unwelcome ontology; it is more the ontology's intuitive irrelevance to the notions being analyzed. Even someone not especially *opposed* to functions, to take that example, is still liable to feel uneasy about putting facts of equinumerosity at their mercy. For various awkward questions arise, of which let me mention three.

How is it that I can tell that my left shoes are equal in number with my right ones just by pairing them off, while the story of how I am supposed to be able to ascertain the existence of abstract objects like functions remains to be told?[1] Pending that story, who am I to say that equinumerosity facts even *correlate* with facts of functional existence—much less that the correlation rises to the level of an analysis?

This paper was written at the National Humanities Center with support from a University of Michigan Humanities Fellowship. I had a lot of help; thanks to Joseph Almog, Louise Antony, Paul Boghossian, Mark Crimmins, Nicholas Georgalis, James Hardy, Sally Haslanger, David Hills, Brad Inwood, David Kaplan, Jerrold Katz, Joe Levine, David Lewis, Bill Lycan, Alasdair MacIntyre, Friederike Moltmann, Thomas Nagel, Calvin Normore, John O'Leary-Hawthorne, Gideon Rosen, Nathan Salmon, Geoffrey Sayre-McCord, Niko Scharer, Stephen Schiffer, Sydney Shoemaker, Scott Soames, Roy Sorensen, Martin Stone, Kendall Walton, and Umit Yalcin. Versions of the paper were read at Duke University, York University, New York University, University of North Carolina at Chapel Hill, East Carolina University, North Carolina State University, Arizona State University, and a conference on the philosophy of Saul Kripke in San Marino.

[1] Compare Benacerraf (1973). See also Katz (1995).

If my left shoes' numerical equality with my right turns on the existence of functions, then in asserting this equality I am giving a hostage to existential fortune; I speak truly only if the existence facts break my way.[2] But that is not how it feels. Am I really to suppose that God can *cancel* my shoes' equinumerosity (and so make a liar out of me) simply by ~~training his or her death gun on~~ knocking out the ~~offending~~ functions, without laying a hand on the shoes themselves?

Assuming that a one-to-one function between my left and right shoes exists at all, there are going to be lots of them. But then, rather than saying that my left and right shoes are equal in number because these various functions exist, wouldn't it be better to say that the functions exist—are *able* to exist, anyway—because my left and right shoes are equal in number? That way we explain the many facts in terms of the one, rather than the one in terms of the many.

All of the analyses mentioned have problems like this. And the reply is the same in each case: The reason these analyses are so popular is that they do crucial theoretical work. If you know of another way of accomplishing this work, terrific; otherwise, though, spare us the handwringing about existence coming in where it is not wanted. This paper explores a *strategy*, only that, for getting the work done without getting mucked up in irrelevant existence questions.

## I

A funny thing happened on the way to the possible-worlds analysis of modality. Or actually, two funny things, of which only the first attracted any notice. The first is David Lewis's well-known “paraphrase” argument for belief in worlds:

> I believe . . . that things could have been different in countless ways. . . . Ordinary language permits the paraphrase: there are many ways things could have been besides the way they actually are. On the face of it, this sentence . . . says that there exist many entities of a certain description, to wit 'ways things could have been'. . . . I believe permissible paraphrases of what I believe; . . . I therefore believe in the existence of entities that might be called 'ways things could have been'. I prefer to call them 'possible worlds'.[3]

If someone wants to know what sort of thing these worlds are,

> I can only ask him to admit that he knows what sort of thing our actual world is, and then explain that other worlds are more things of *that* sort, differing not in kind but only in what goes on at them.[4]

So, other worlds (of the same general sort as our actual world) exist because there are other ways things could have been; and other ways things could have been exist because things could have been different from the way they are in actual fact.

[2] Compare Etchemendy (1990).

[3] Lewis (1979), 182.

[4] Ibid. 184.

Now, everyone knows there is something funny about this argument of Lewis's, because Stalnaker has told them:

> If possible worlds are ways things might have been, then the actual world ought to be *the way things are* rather than *I and all my surroundings. The way things are* is a property or state of the world, not the world itself. The statement that the world is the way it is is true in a sense, but not when read as an identity statement. . . . One could accept . . . that there really are many ways that things could have been . . . while denying that there exists anything else that is like the actual world.[5]

FN:5

The second funny thing is this explanation of Stalnaker's. If it hasn't struck people that way, that's because Stalnaker has packed two quite different points very closely together.

Stalnaker's negative point is that while the paraphrase argument may establish something, it does not establish the existence of Lewis-worlds—that is, (i) concrete, I-and-all-my-surroundings, worlds, which unlike this one are (ii) worlds that do not actually exist. All the argument gets you is "ways things could have been", and ways things could have been meet *neither* condition—not (i), because ways of being are not to be confused with the things that *are* those ways, and not (ii), because if we ask, "Are there *actually* other ways things could have been or is it just that there *could* have been?", the answer is that there actually *are* these other ways.[6]

FN:6

What's so funny about that? Wait, I haven't got to the funny part. ~~That's~~ the part where Stalnaker turns his critique of Lewis's reading of "ways things could have been" into a positive proposal of his own.

I haven't got to

Think of Stalnaker as arguing like this: Ways the world *could* have been are of the same ontological type as the way it *is*. So we need to determine the ontological type of *the way the world is*. What better place to start than with the truism that *the world is the way it is*? Some might read this as saying that *the way the world is* is one and the same entity as the humongous concrete object known hereabouts as *the world*. That, however, would be a mistake: "The statement that the world is the way it is is true in a sense, but not when read as an identity statement."[7] But, if the statement is false read as an identity statement, what is the reading on which it is true? Stalnaker doesn't come right out and tell us, but the usual alternative to an "is" of identity is an "is" of predication. Apparently, then, Stalnaker is saying that the statement is true when the "is" is taken as predicative.

FN:7

This is what I find funny, or at least puzzling. Because the phrase following the "is", namely "the way it is", looks less like a *predicate* than a *singular term*.

[5] Stalnaker (1979), 228.

[6] Otherwise it would seem that blue swans, although they *would* have been possible had things been different, are not possible as matters stand.

[7] Stalnaker (1979), 228. Identity statements are reversible: if $A = B$, then $B = A$ as well. To say that *the way the world is, is the world*, however, sounds wrong. (Except maybe to those who believe with Wittgenstein that the world is all that is the case. Even to them, though, it won't sound truistic, as *the world is the way it is* does. And it should, if the latter is an identity statement.)

And Stalnaker *uses* it as a singular term, when he says that "*the way things are* is a property or state of the world" and that "*the way the world is* could exist even if a world that is that way did not".[8] But, looking at the matter naively, when you've got an "is" between two singular terms, the "is" is not an "is" of predication but one of *identity*.[9] Which is just what Stalnaker denies.

FN:8 FN:9

Of course, if "the way the world is" stands for a property, then there *is* a true predication in the vicinity: one ought to be able to say that the world *has* this property. But "the world *has* the way it is" sounds quite wrong. Why, if "the way the world is" denotes a property that the world as a matter of fact possesses?

## II

What are we to make of this phrase "the way the world is"? According to Stalnaker, it does not stand for me and all my surroundings. But it does not appear to stand for a *property* of me and my surroundings either. What then?[10]

FN:10

The strategy that suggests itself is this. Take the matrix "the way the world is is *X*", plug in a term that makes for a true sentence, and ask what the term stands for. That ought to be what "the way the world is" stands for as well. When we try to carry this strategy out, though, we run into an unexpected problem; the matrix doesn't want a term, it wants adjectives:

> The way the world is is large, complicated, law-governed, mostly uninhabited, shot through with force fields, bathed in radiation, etc.

And now things get *really* confusing. There is no way on earth of interpreting the main "is" in this sentence as an "is" of identity. And yet, if we interpret it as an "is" of predication, we are back with Lewis's concrete worlds—for the thing that

[8] Someone might say that the same word or phrase can play both roles depending on grammatical context; "red" functions as an adjective in "the book is red", a noun in "red is a color". I am not sure what to say about this, but one possibility is that there are *two* words here with the same spelling. (Why *color* nouns in particular should be so often spelled the same as their corresponding adjectives is unclear to me; compare "triangle is a shape", or "tiny is a size".) Another example, suggested by Niko Scharer, is this: "the color it is" functions as an adjective in "the book is the color it is", a noun in "the color it is is a dark color". Again it seems possible that we have two phrases here with the same orthography. The second is a definite description, the first is a concealed question, like "his age" in "I know his age".

[9] Or perhaps of composition, as in "that clay you sold me is now a statue of Goliath". But that is not a likely reading here.

[10] Perhaps "the way it is" is a proadjective or proadverb along the lines of "thus" or "thusly". This would leave us without a reading of Stalnaker's statement that "the way the world is is a property of the world". But Stalnaker exegesis aside, the proform idea seems worth pursuing. Eventually a theory of proadjectival or proadverbial quantification would be needed, for we have sentences like "whatever way things had been, they would have been such and so" to deal with. There is ~~in fact~~ a neglected literature on this, much of it inspired by the work of Lesniewski. See Prior (1971); Kung (1977); and Simons (1985). Also relevant is the program laid out in Grover (1972) and in Grover *et al*. (1975), both reprinted in Grover, (1992).

is large, complicated, bathed in radiation, law-governed, etc., is not a property of me and my surroundings but, well, me and my surroundings.

Here is our puzzle. "The way the world is", "the way it would have been if so and so had happened", "the ways it could have been"—these *look* for all the world like noun phrases. It stands to reason then that they at least *purport* to denote entities of some sort. What sort? This is a puzzle whether you believe in the purported entities or not. Indeed, you don't *know* whether to believe in them until you solve the puzzle—until you figure out what the entities are whose existence is in question.

## III

Suppose we start by beating some neighboring bushes. Ways the world could have been are hardly the only ways countenanced ~~by~~ in ordinary speech. Just from the song "The Way You Do the Things You Do" you could gather a respectable collection.[11] But let's have some more humdrum examples: the ways people *feel* on various occasions (sleepy, happy, jealous, relieved, like a motherless child), the ways birds have of building their nests, the ways of getting from point *A* to point *B*, and so on.

Now, what kind of entity am I talking about in talking about these various ways? Take the way I felt when I got up this morning, viz., sleepy; or the way cockatoos build their nests, from the outside in;[12] or the fastest way of getting from Toronto to Lima, that is, via Tegucigalpa. What *are* these things?

If you are anything like me, the tempting reply is: *What* things? It is hard to think of the phrases "sleepy", "from the outside in", "via Tegucigalpa" as standing for entities at all; their function just does not seem to be referential.

Someone might want to write this off to a lack of imagination. "Sleepy" denotes a state of mind, namely, *sleepiness*; "from the outside in" stands for a property of nest-building events, the property of *centripetality*; "via Tegucigalpa" stands for a path or set of paths or property of paths through space-time.

f

But if sleepiness, happiness, relief, and so on are the entities collectively denoted by "ways of feeling", then it is strange, isn't it, that one can't say "ways of feeling, for example, . . ." and then plug in *names* of these entities. Why is it "ways of feeling, for example, sleepy, happy, relieved, like a motherless child, . . ." rather than "ways of feeling, for example, sleepiness, happiness, relief, similarity to a motherless child, . . ."? Again, if sleepiness and so on are among the entities that are called "ways of feeling", then it ought to make sense to say "sleepiness is the way I felt this morning" and "nobody knows the way I feel: similarity to

[11] The way you smile so bright, the way you knock me off my feet, etc. This is to say nothing of "Fifty Ways to Leave Your Lover".

[12] I am told that cockatoos don't build nests, but say they do.

a motherless child". And it doesn't. What makes sense is "sleepy is the way I felt". And *sleepy* doesn't seem to be an entity at all, not even an entity of a highly abstract sort.

## IV

What is going on? A first clue to the peculiar behavior of "way" is that "*the way* I feel" sounds rather like "*how* I feel".[13] This is no coincidence; "the way" lines up with "how" over a wide range of cases:

(1) the way you put it just now / how you put it just now
the way things work around here / how things work around here
the way they met / how they met
the way she wants to be remembered / how she wants to be remembered

On the right-hand side of (1) we have what ~~grammarians~~ linguists call *indirect questions*. I can't give you an exact definition (I'm not sure anyone could) but intuitively, indirect questions are noun-like counterparts to ordinary or direct questions:

(2) how things work around here / how do things work around here?
what the coach forgot / what did the coach forget?
why he's acting like that / why is he acting like that?
when the swallows return / when do the swallows return?
who invited them / who invited them?
whether it will rain / will it rain?
where she is headed / where is she headed?

The problem we have been wrestling with is, in effect: Are indirect questions referential, or is their semantical contribution to be sought elsewhere?[14]

Now in asking, "Are they referential?" I mean not, "Are there Montague grammarians or other formal semanticists somewhere who have cooked up super-duper semantical values for them, say, functions from worlds to functions from worlds and *n*-tuples of objects to truth-values?"[15] The answer to that is going

[13] Exercise: Compare and contrast (i) "how *S*" / "the way that *S*", (ii) "why *S*" / "the reason that *S*", (iii) "whether *S*" / "the truth-value of *S*", (iv) "which *S*" / "the *S*'s identity", and (v) "how many *S*s" / "the number of *S*s". [It was some remarks by Ian Rumfitt about (iii) that got me thinking about (i).]

[14] The issue is complicated by the distinction between indirect questions and free relative clauses. The distinction is not always well marked in English, but there are some clear cases. "What you think" is a free relative in "I think what you think", an indirect question in "I wonder what you think". Sometimes context does not resolve the ambiguity: "what you know" is a free relative in "I know what you know" interpreted as "whatever you know, I know", an indirect question in the same sentence interpreted as "whatever you know, I know you know".

[15] I certainly don't mean to disparage this kind of approach; it can be powerfully illuminating. See in particular Groenendijk and Stokhof (1989).

to be *yes* almost no matter what part of speech you're talking about—connectives, prepositions, and apostrophe "s" not excluded. I mean: Are they referential in the way that singular terms are, so that someone using an indirect question could reasonably be said to be *talking about* its referent, or purporting to talk about its purported referent?

Are indirect questions referential in *that* sense? Truth be told, some of them seem at first to be, because some of them seem linkable by true identity statements with phrases whose referential status is beyond question. Here are some examples:

(3) what the coach forgot was the keys, the map, and the schedule
March 31 is when the swallows return
who invited them is your friend Becky
FN:16 Albuquerque is where she's going[16]

Taking the "is" in these statements to express identity—and what other option have we, really, with noun phrases on either side?—"who invited them" stands for Becky, "where she's going" stands for Albuquerque, and so on.

And yet, there can be very similar-looking statements where the identity interpretation is unavailable. Newt Gingrich recently had to explain to the mainland Chinese that he hadn't *meant* it about extending recognition to Taiwan, he was "only trying to rattle their cage". Taking him at his word, why Gingrich talks like that is to rattle their cage. Or suppose that pharmacologists studying the effects of tranquilizers on the brain determine that the way Valium soothes our ruffled feelings is by blocking the action of a certain neurotransmitter. Well and good. But, does anyone really take statements like

(4) why he talks like that is to rattle their cage
how Valium soothes is by blocking that neurotransmitter
why we hesitated was out of concern for you
how I want to feel is happy

to assert literal identities? I hope not, because on the face of it there's no such thing as *to rattle their cage* or *out of concern for you* or *happy* to be identical *to*. And

[16] Or are the ~~initial phrases~~ *wh*-phrases in (3) free relatives? Here are some reasons to think not. First, free relatives are interrogated with the relative pronouns they embed—"I once had a drink where Elvis was born" / "you once had a drink where?"—whereas indirect questions are interrogated with "what?"—"I wonder where Elvis was born" / "you wonder what?" And in response to (e.g.) "March 31 is when the swallows return", we say not "*March 31 is when?" but "March 31 is what?" Second, interrogative pronouns take strong stress in a way that relative pronouns do not; "I know WHERE Elvis was born—only not WHEN" but "*I had a drink WHERE Elvis was born—only not WHEN." And we have no problem with "WHO invited them is Becky; WHY I have no idea". Third, plural relatives in subject position take the plural form of the verb—"what Shakespeare regarded as his best plays are nowadays seldom read"—whereas indirect questions take the singular form—"what Shakespeare regarded as his best plays is anybody's guess". And we say, "what the coach forgot is the keys, the map, etc.", not "*what the coach forgot are the keys, the map, etc.". Finally, free relatives according to most authors can only be introduced by "what", "when", or "where". But the construction in (3) works for all *wh*-words other than "whether". I am indebted
Q1 ~~to Baker indebted~~ here to C. L. Baker's (1968) for this.

this raises doubts about the referential interpretation of the examples in (3) as well. The "is" of "how Valium soothes is by blocking that neurotransmitter" seems indistinguishable from the "is" of "*who* Valium soothes is the people who take it". If it expresses identity in the one case, it ought to do so in the other.

Now, of course, the referentialist can simply insist that the two "is"s are different; the one in (3) expresses identity, the one in (4) expresses predication. This would be just as good from her point of view since, predicative "is" functioning to bring the *referent* of the phrase it follows under the descriptive content of the phrase it precedes, "why he talks like that", etc., would again be cast in a referential light.

But the predicative interpretation is hard to make out. The best way to see this is to *allow* the referentialist her contention that the phrases on the left-hand side of (4) are referential. Say, in other words, that an entity *how I want to feel* exists. Is this entity characterized as happy by the sentence "how I want to feel is happy"? Clearly not. No one is saying that a full accounting of the happy things would include (in addition to Dale Carnegie and Barney the dinosaur and your typical sea otter) *how I want to feel*. The claim is rather that if you are asking me how I want to feel, the answer is that I want to feel happy.

With the other examples, matters are even worse. "$X$ is happy" at least has the right *form* to describe $X$. But what property or characteristic is attributed to $X$ by "$X$ is to rattle their cage", "$X$ is by blocking that neurotransmitter", or "$X$ is out of concern for you"? Someone might say that "$X$ is to rattle their cage" describes $X$ as being *done* for the purpose of rattling their cage. But then the problem is the same as before: being done for such and such a purpose is a characteristic not of *why people perform actions* (!?!) but of the actions they perform.

Actually, to the extent that the identity–predication distinction finds a foothold in (4) at all, it may be doubted whether the advantage lies with the predicative approach. What was it that the pharmacologists told us? Not that blocking that neurotransmitter is an *aspect* or *feature* of how Valium soothes; according to them, it *is* how Valium soothes. And you can almost hear Gingrich at the press conference: "You people just don't get it, to rattle their cage *is* why I made those statements—there's no difference between the two." Pressed from the other side by the predicative interpretation, one almost wants to say that there's an *identity* here. There isn't, of course, but the *feeling* of identity is a
FN:17 fascinating datum and one that needs to be taken seriously.[17]

But the case against referentialism needn't be made to rest on these subtleties about predication versus identity. Take "who invited them", as in "who invited them is your friend Becky". If this referred to Becky, we would expect it to be

[17] Not that all indirect-question-embedding statements with main verb "is" have the atmosphere of an identity statement. Some feel downright predicative: for instance, "she is how I used to be" and "the world is how it is". A good theory of indirect questions ought to have something to say about this.

intersubstitutable *salva veritate* with *other* phrases referring to Becky—phrases like "Becky". What we find though is that the two are not even substitutable FN:18 *salva congruitate*.[18] This is illustrated by

(5) I wonder where Becky is ~~has gone~~ / *I wonder where who invited them is ~~has gone~~.
Becky was accepted at Yale / *Who invited them was accepted at Yale.

and, the reverse substitution,

(6) I wonder who invited them / *I wonder Becky.
It doesn't matter who invited them / *It doesn't matter Becky.

These examples point up a final difficulty with the referential approach. Even if "who invited them" *did* refer to Becky, that would explain only a tiny fraction of its semantical behavior. Karttunen in "Syntax and Semantics of Questions" offers the following overview of indirect-question-embedding con- FN:19 texts.[19]

(7) VERBS OF
acquiring knowledge: ask, wonder, learn, notice, discover
retaining knowledge: know, be aware, recall, forget
communication: tell, show, indicate, inform, disclose
decision: decide, determine, specify, agree on, control
conjecture: guess, predict, bet on, estimate
opinion: be certain about, have an idea about
relevance: matter, care, be important, be significant
dependency: depend on, be related to, be a function of

"Who invited them" can occur in almost all of these contexts, yet its semantic FN:20 contribution is purely referential in none of them.[20] Shouldn't we look for an explanation of the *other*-than-referential work done by indirect questions, before we go ahead and assign them referents? That explanation might turn out to apply across the board.

## V

Indirect questions are indirect *questions*; that is the point we keep on losing sight of. Since they are questions, it would not be surprising if their interpretation

[18] Admittedly, the *salva congruitate* test has its limits; "sunny Madrid is a favorite of ours" sounds good, "*sunny the capital of Spain, etc." doesn't. Examples can also be given however where it is truth-value that changes.

[19] I have taken some liberties. See Karttunen (1977).

[20] As always, I am talking about ordinary common-sensical reference, not reference to higher-type objects as in Montague grammar; "who invited them" refers to the one who invited them if it refers at all.

went via the items that *normally* go by that name, viz., *direct* questions. Here is a crude first proposal, using Q for the direct question corresponding to an indirect question IQ:

FN:21 To say that ____ IQ ____ is to offer information about Q's answer or answers.[21]

This admittedly ~~deliberately~~ leaves a lot to the imagination. Among the issues I propose to duck, or settle in whatever way seems most convenient at the time, are these:

What kind of information? Is the information determined by the sentential context (____ . . . ____) alone or do other factors contribute? What are the mechanisms by which these factors operate?

Are answers linguistic in nature or do linguistic items function rather as presentations of the real, extralinguistic, answers?

Are answers one and all sentence-length, as you would think from the fact that we call them true and false, or not, as you would think from the fact that "5" (or perhaps 5) is the answer to "what is 2 plus 3?"

Does each question have a unique complete and correct answer or do some have multiple answers with these features?

All that we need to assume for now is that to each direct Q corresponds a unique complete and correct answer-*set* AQ; there is no official line on how FN:22 many answers AQ contains or what sorts of entities these answers are.[22]

Statements embedding indirect questions IQ are in the business of offering FN:23 information about Q's answer or answers—about AQ.[23] One move in this direction stands out as particularly natural; we might seek to provide AQ outright. The natural way to proceed if that is our goal is to say simply that *IQ is AQ*:

(8) what the coach forgot was the keys, the map, etc.
who invited them is Becky
why Gingrich talks like that is to rattle their cage
how Valium soothes is by blocking that neurotransmitter
why we hesitated was out of concern for you
how I want to feel is happy

[21] "Offering information about Q's answer(s)" is not to be thought of as necessarily involving reference to or quantification over answers, or acceptance of answers into one's ontology. I can give information about the answer to "where is she going?" by saying simply "she's going to Albuquerque". (This fits nicely with Alasdair MacIntyre's suggestion that "where she's going is Albuquerque" is a piece of playacting in which I set myself a question and then respond.)

[22] Both kinds of flexibility will be important later on when we get to questions like "how is that possible?"—the first because such a question might have any number of correct answers, the second because its answers might be understood either as (linguistic or abstract) world representations or as the possibility-conferring worlds themselves.

[23] Wherever convenient, the distinction between AQ and its members will be blurred—so that "what is 2 plus 3?", although strictly speaking {"5"} or {5}, is in practice "5" or 5.

Notice again the equational, identity-like, feeling of these statements. A tempting explanation is that in each case we have an identity in the vicinity: “Becky” *is* (in the identity sense) the answer to “who invited them?”, “to rattle their cage” *is* the answer to “why does he act like that?”, and so on.[24] As for the identity-feeling's curious insensitivity to the grammatical category of the phrase following the “is”—that “happy” is an adjective does not make “how I want to feel is happy” feel any less identity-like—this is only to be expected if the underlying identity is between “happy” and the answer to “how do I want to feel?” rather than (?!?) *happy* and *how I want to feel*.

Another puzzle left over from the last section is this. Why is it that some indirect questions, like “who invited them”, seem at first glance referential, while others, like “how Valium works”, do not? The equational flavor of “IQ is AQ” suggests a two-part explanation. First, some AQs (“Becky”) are referential; others (“via Tegucigalpa”) are not.[25] Second, the “is” of “IQ is AQ” acts as a pipeline transmitting felt referential character from the one side to the other. This leads to the prediction that IQ should strike us as prima facie referential when

(i) “IQ is AQ” makes sense;
(ii) AQ is referential.[26]

Is the prediction borne out? Indirect who-, where-, and when-questions typically satisfy both conditions, so we would expect them to give the impression of referring.[27] And for the most part they do: “who invited them” appears to refer to Becky, “where she's headed” to Albuquerque, and so on.

Indirect how- and why-questions satisfy (i) but not (ii); “how I feel is happy” and “why he talks like that is to rattle their cage” both scan, but their right-hand sides do not refer. So the prediction is that they will *not* feel referential, and this again seems true.

[24] Or perhaps it is the *meanings* of these phrases that constitute the answers; this is one of the issues we're leaving open. Note that the identity-feeling wanes as the material after the “is” goes from *providing* the answer to merely *constraining* it to merely *commenting on* it. Thus “who invited them is Becky”, “who invited them is someone with a strange sense of humor”, “who invited them is no one you want to know”, “who invited them is classified information”, “who invited them is a mystery to me”.

[25] Here again I am blurring the difference between AQs and their members. Also I assume for convenience that answers are linguistic in nature; if not, substitute “AQ's linguistic presentation” for “AQ”.

[26] Jerrold Katz objected that this is circular since AQ might *itself* be an indirect question. One reply is just to stipulate that indirect-question substituends are not allowed. Another is to say that we should read (ii) as: AQ is obviously, convincingly, invincibly referential. Indirect questions fail this condition, so we can safely ignore them.

[27] Some what-questions belong here too, e.g., “what the coach forgot”. But what-questions are incredibly various. Just as often they resemble how- and why-questions in satisfying (i) but not (ii), as for example, “what works best is to dip the brush in turpentine”. Where- and when-questions are tricky too. Prior (1971) observes in effect that these often call for prepositional answers rather than nominal ones: “in Paris”, “to Albuquerque”, “on March 31”, and so on. That leaves only indirect who-questions as *clearly* apparently referential.

Indirect which-questions are a mixed bag, but what *often* happens is that they satisfy (ii) but not (i). The answer to “which door do you pick?”, for example, might be “door number three”, a perfectly good referring phrase. But “which door you pick is door number three” doesn't scan. The prediction then is that “which door you pick” will not strike us as a referring phrase. (The reason is not inherent ungrammaticality, since “which door you pick is up to you” sounds fine.)

Indirect whether-questions feel *highly* nonreferential, finally, as a result of satisfying neither (i) nor (ii). Not only are their AQs lacking in reference, they cannot be plugged into sensical “IQ is AQ” statements: witness “whether you are still grounded is yes”.

Before getting back to possible worlds, consider one last puzzle from the previous section. If the point and purpose of indirect questions is to refer, then what are we to make of

(9) I wonder who invited them
it doesn't matter who invited them
who invited them is none of your business
where guests sit is a function of who invited them?

Read these statements as commenting on the answers to their embedded questions,[28] and all becomes clear. *Something* is said not to matter by “it doesn't matter who invited them”, but it is not Becky, it is the answer to “who invited them?” “Where guests sit is a function of who invited them” does not assign seating authority to any particular person, it says that the answer to “where shall *X* sit?” depends on the answer to “who invited *X*?”[29]

## VI

Assuming that something like this approach to indirect questions is correct, what does it tell us about the possible-worlds account of modality?

Here is the answer you probably expect: It upsets Lewis's paraphrase argument according to which we are committed to worlds in being committed to ways things could have been. The argument doesn't hold up, because each and every way things could have been is a *how* things could have been. And the phrase “how things could have been” is an indirect question with zero referential import.

I see three connected problems with this answer. The first is hinted at by the awkwardness of what I just said: “each and every way things could have been is a *how* things could have been.” A *how* things could have been? What on earth is that? “How things are” makes sense as the translation of “the way

[28] Strictly, the answers to the direct counterparts of their embedded questions.

[29] This section has borrowed freely from the literature on questions, including Karttunen (1977); Engdahl (1986); and Higginbotham (1993).

things are", and "how things would have been" works as the translation of "the way things would have been". But nothing along these lines is available for "some/all of the ways things could have been", because *hows* are not things the language countenances.[30] If the idea was to translate world-talk into way-talk, and way-talk into how-talk, and how-talk into answer-talk, then the idea doesn't work, because *quantificational* way-talk doesn't translate.

Now for the second problem, which has to do with our emphasis throughout on the *irreferentiality* of indirect questions. Isn't this missing the point of the paraphrase argument? Lewis's concern is with ontological *commitment*. And as Quine thought he had made sufficiently clear about half-a-century ago,[31] one makes little progress on matters of ontological commitment by staring at controversial chunks of language waiting for them to yield up the secret of whether they are really referential or not.[32] (You know what they say about a watched mot.) The true and proper test of ontological commitment is *quantification* into the position a given chunk of language occupies. That is why Lewis argues from the fact, not that ordinary language *refers* to ways, but that ordinary language *quantifies* over ways.

Third, the paraphrase argument was never the important one in the first place. The important argument has always been that possible worlds are too useful to be done without. Lewis is crystal clear about this:

> Why believe in a plurality of worlds?—Because the hypothesis is serviceable, and that is a reason to think that it is true.[33]

"Even those who officially scoff", he adds, "often cannot resist the temptation to help themselves abashedly to this useful way of speaking."[34] And to repeat, this is a way of speaking that is up to its neck in ontologically committal quantification.

Wrapping all of this up into a single point: if your one ontologically deflating move concerns indirect questions, and if the real measure of ontological commitment is not these but quantification into the spots they occupy, and if quantification into the spots they occupy is practically speaking unavoidable, then you really haven't gone very far towards diminishing commitment to possible worlds.[35]

[30] Phrases like "whys and wherefores" perhaps reflect some long-ago attempt to go plural while remaining interrogative. Even "hows and whys" is not unheard of. Consider this from *Webster's 3rd New International Dictionary*: "most of the film is devoted to the grim hows and not the difficult whys of battle."

[31] See Quine (1948).

[32] There is a cartoon about the last worker on the Sara Lee assembly line: She sits by the conveyor belt asking herself of each passing pie, "Would I be proud to serve this to my family?" Replace the pies with noun phrases, and you have Quine's picture of the traditional ontologist: "Am I content to think that this refers to a bona fide entity?"

[33] Lewis (1986), 3.

[34] Ibid.

[35] This objection might be answerable for a particular $S$ if we had an inventory of all the ways things could have been which afforded $S$ any chance at possibility: say, the way they would have been if $A_1$, the way they would have been if $A_2$, . . . , and so on. Then we could say that $S$ is possible

## VII

Hold on, though. Because "the *way* that such-and-such" translates into "*how* such-and-such", one naturally supposes that "some/all of the *ways* that such-and-such" has got to translate into "some/all of the *hows* such-and-such"—which of course it can't, because "hows" makes no sense. But this is to insinuate a problem about *plurals* into what was supposed to be a problem about *quantification*.

Imagine someone arguing as follows: "the reason I did it" translates into "why I did it", but no analogous translation is possible for "some/all of the *reasons* I did it", because *whys* are not things the language countenances. Similarly, we have "when the swallows return", "where your story breaks down", and "what really gets my goat", but not "some/all of the *whens* they return", "some/all of the *whereds* your story breaks down", or "some/all of the *whats* that get my goat". There is no grammatical alternative, it seems, to objectual quantification over reasons, times, narrative breakdown points, and whatever sort of thingum it is that gets people's goats.

What this argument overlooks is that one doesn't *need* pluralizations of "why", "when", "where", etc., to carry out the relevant sorts of quantification; that's what words like "whyever", "never", "always", "somewhere", and "whatever" are for. The same applies to "how". "Some/all of the hows" may not make sense, but it doesn't *have* to, for we have "somehow" and "however":[36]

(10) if Valium works how you say, then Librium works similarly
variables: if Valium works *thusly*, then Librium works *like so*
existential: if Valium works *somehow*, then Librium too works *somehow*
universal: *however* Valium works, *so* also works Librium[37]

This is important because it means that quantificational way-talk *can* be rendered in terms of "how", if only we drop the assumption that all quantifiers are (like the logician's "there is an $x$ such that . . .") *entitative* or *objectual*.[38] Of course,

iff there's a way things could have been such that $S$
iff $S$ is how things would have been if $A_1$, or
$S$ is how things would have been if $A_2$, or . . .
iff $S$ figures in the answer to "how would things have been if $A_1$?" or $S$ figures in the answer to "how would things have been if $A_2$?" or . . .

(Or, we could drop questions altogether and say simply that $S$ is possible iff it would have been that $S$ had it been that $A_1$, or it would have been that $S$ had it been that $A_2$, or . . . ) I take it, though, that this kind of inventory is not to be had. The best we can do is: $S$ is possible iff there is *some* way things could have been such that $S$.

[36] See Prior, (1971), 34 ff.

[37] For some reason this sounds better in old English: "however Valium worketh, so also worketh Librium". caps

[38] Better: we need to drop the assumption that the only alternative to an objectual quantifier is a substitutional one.

the embrace of nonobjectual quantifiers gains us nothing unless they share in the freedom from ontological commitment we saw with the corresponding indirect questions. But intuition is absolutely clear on this point. "He blew the house down by huffing and puffing" is a strictly *stronger* statement than "he blew it down somehow". And it would *not* be stronger if the "somehow" carried a commitment to, say, ways, for "by huffing and puffing" is not committal in anybody's book.

## VIII

All of that having been said, let me be the first to admit that colloquial how-quantification is a pretty clumsy semantical instrument compared to direct objectual quantification over ways. Anyone who doubts this is invited to try to render "there is more than one way to skin a cat", or "there are more ways of skinning a cat than of falling off a log", or "some ways of falling off a log resemble some ways of skinning a cat" in the idiom of "somehow", "nohow", and "however". One *could* try to meet these expressive difficulties head on by concocting ever fancier how-quantifiers ("cats can be skinned *doublehow*", "skinning a cat is *howlier* than falling off a log", . . . ).[39] But as a practical matter, there seems little real alternative to quantifying directly over ways, or some approximation to ways.

A second reason why "somehow" and "however" are no automatic panacea is that *all* quantifiers, however nonobjectual in appearance, are caught up in a powerful objectual undertow that threatens to obliterate the distinction here at issue. Talk about objects has a clarity and tractability that few can resist—neither ordinary folk trying to *convey* meanings nor philosophers trying to *explain* them.[40] Things have reached the point that to give the "semantics" of a construction is almost by definition to tell a story about which entities have to behave in which ways for it to make what sort of contribution to truth-value. Anything less and it will be protested that the construction's meaning has still not been ~~rendered~~ made completely clear.[41]

The third reason for refusing to rest content with "somehow" is that an objectual story is not out of the question in this case. Quantified uses of "how" have, it seems, a similar semantical function to indirect interrogative uses. Both

[39] This approach would resemble the *modalist* project of trying to approximate the expressive power of direct quantification over worlds by means of souped-up modal operators, e.g., indexed actuality operators. See Fine (1977) and Forbes (1985), 89 ff.

[40] See Quine (1969).

[41] This is not to forget the occasional brave soul who attempts to do the semantics of, say, tense or adverbs in a tensed or adverbial metalanguage. The brave soul is the exception that proves the rule, because the feeling is bound to be that she is an obscurantist who for some reason refuses to dig down to the deepest semantic levels. If there is no deeper story to be had, the construction itself will be derided as (in a pejorative sense) idiomatic; like "for the sake of . . . " and "believes that . . . " it is better suited to the "the market place or . . . the laboratory" than to precincts more theoretical (Quine (1960), 228).

enable commentary on the answers to direct how-questions; but where indirect "how" does it by, for lack of a better word, *invoking* these answers, quantified "how" does it by *generalizing* over them. "Life after death is possible somehow" says more or less that "how is life after death possible?" has a correct answer. "However you want to do it is fine with me" says that if "like so" is a correct answer to "how do you want to do it?," then doing it like so is fine with me. That the *majority* of nonobjectual quantifiers are built around question-words ("*which*ever", "some*where*", "*why*ever", "*what*ever", "*how*ever many", "*who*ever") only strengthens the case for a semantic link with answers.

## IX

Haven't we now painted ourselves into a corner? Nonobjectual quantifiers, let them be as seriously intended as you like, are not ontologically committal. But they are not semantically primitive either, and our best bet about how to explicate them is in terms of quantifiers over answers that *are* (when seriously intended) committal.

I see only one way out. If a construction that is not committal-when-serious is to be explained by a construction that *is*, then the second construction had better be treated for purposes of the explanation as *non*serious or feigned. Someone who says that the treasure is buried somewhere is saying that "where is it buried?" has an answer, *bracketing* any and all worries about the existence of entities suited to play that role—*stipulating*, if you like, in a spirit of make-believe, that questions that "have answers" in the ordinary-language sense of not being unanswerable have them in a more ontologically loaded sense as well.

This leaves the shape of the make-believe somewhat open. A lot will depend on our views about answers: about what they are in general (linguistic items or their denotata?) and the (grammatical or ontological) forms that they take in connection with specific sorts of questions. If we favor the linguistic conception, for example, then what needs to be imagined is that any answer that was, as we say, "there to be given", *was* given. This is to avoid holding the truth of "he blew the house down somehow" hostage to the issue of whether someone has in fact bothered to put "by huffing and puffing" into words.

But suppose that linguistic so-called "answers" are instead answer-*formulations*; the real answers are the worldly entities to which they refer and other entities of the same sort.[42] Then there are two cases, according to whether the linguistic so-called "answers" are of the right grammatical *form* to refer.

The easy case is the first; here our main imaginative task is to supply each linguistic "answer" with a referent. There will have to be such things as the Easter

[42] E.g., in the case of who-questions, persons rather than their names, and indeed regardless of whether they *have* names.

Bunny, for instance, and the number twelve, to serve as answers to “who does Isaac expect to see at the mall?” and “how much is five plus seven?” Otherwise it will not come out false, as it should, that whoever Isaac expects to see at the mall, Sally expects to see as well, and that however much you get by adding five to seven, you get as much or more by adding three to eight.

If on the other hand we are dealing with linguistic “answers” that do not even *purport* to refer (perhaps they are adjectives or adverbs), then referential purport will have to be projected onto them. The needed make-believe will have two interlocking parts: one in which the phrases in question are seen as referential, another in which their referents are seen to be drawn from the ranks of some real or concocted ontological category. (This paper does not advocate any particular account of ways, but one could do worse than the following: ways are the things we imagine ourselves ^ referring to by use of phrases like “by huffing and puffing”, when we imagine that phrases like that are used to refer.)

^to be

## X

But we are getting ahead of ourselves. It is time to put make-believe to the side for a while and return to the original objection: “If the idea is to translate world-talk into way-talk, and way-talk into how-talk, and how-talk into answer-talk, then the idea doesn’t work, because *quantificational* way-talk doesn’t translate.” Our reply is that quantificational way-talk *does* translate—into colloquial how-quantification—and that colloquial how-quantification translates in turn into (feigned, but we are putting that aside for now) objectual quantification over answers to how-questions. That the middleman here has its expressive limits is not a problem, for we can cut the middleman out and read apparent quantification over ways *directly* into objectual quantification over answers.

Take for instance Lewis’s statement that “there are many ways things could have been besides the way they actually are”. This says that “how could things have been?” has many incompatible answers that do *not* correctly answer “how are things as a matter of fact?”

Next try “there is a way things could have been such that blue swans existed”. One interpretation is that “how could things have been?” has an answer according to which there are blue swans. But this might give the impression of a two-stage process in which we first collect answers to “how could things have been?” at random, only then considering whether we have hit on anything favorable to the blue swan hypothesis.[43] A better interpretation homes in on the scenarios we are

FN:43

[43] It might; I don’t say it must. I prefer the interpretation to be given next because it steers around various tricky issues raised by “according to”, such as the following: can a strictly microphysical answer to “how could things have been?” be such that according to *it* there are blue swans?

actually interested in: “how could it have been that there were blue swans?” *has an answer*, full stop.[44]

FN:44

## XI

Anyone who does modal metaphysics at all has got to feel *some* attraction to the formula: *S is possible iff there is a way things could have been such that S*. To go by what was just said about quantification over ways, the formula means something like this:

(11) $\Diamond S$ iff H $\Diamond$ $S$? has a correct answer.

(“H $\Diamond$ $S$?” is short for “how could it have been that $S$?”) Standard possible worlds semantics has of course grown up around a quite different reading: $S$ is possible iff there is at least one $S$-world, an abstract sort of world in Stalnaker’s version of the semantics,

(12) $\Diamond S$ iff there is an abstract world according to which $S$,

a concrete sort in Lewis’s version,

(13) $\Diamond S$ iff there is a concrete world at which $S$.

Not surprisingly, the three approaches agree in linking possibility to the existence of an appropriate witness. (If necessity is the dual of possibility, they agree too in linking $S$’s necessity to the *nonexistence* of witnesses to the possibility of not-$S$.) But notice a crucial difference between them. (12) and (13) we believe, or try to believe, because we are so impressed by the theoretical work they do. (11) on the other hand comes close to being a conceptual truth about possibility.

How is that? (11)’s right-to-left direction says that $S$ is possible provided that “how is $S$ possible?” has a correct answer. Assuming that a correct answer to “how is $S$ possible?” will be a truth of the form “$S^+$ is possible”, where $S^+$ has $S$ as a consequence, this amounts to the claim that

(11′) $\Diamond S$ if $\Diamond S^+$—where $S^+$ is sufficient for $S$.

[44] Of course, the answer has to be *correct*. Someone might say that worlds reinsert themselves just here, when we try to explain what correctness comes to. But why does the explanation have to be in terms of what does or does not exist, as opposed to what is or is not possible? Suppose that an answer to “how is $S$ possible?” takes the form “$S^+$ is possible”, where $S^+$ is presented as sufficing for $S$. Then the main thing correctness requires is that $S^+$ *is* possible and *does* suffice for $S$. How the correctness of *these* claims is to be understood is another question. Some may opt for homophonic correctness conditions or for Peircean ones; others may insist on something more substantive, up to and including, I suppose, conditions framed in terms of possible worlds. That these various options continue to be available is in a way the point. The worldly semantics is presented as an offer we cannot refuse. But the same analytic advantages are available to those choosing another semantics and, indeed, to those leaving the semantical choice unmade. (This note is one of several places in which I blur the distinction between possible-worlds semantics and the possible-worlds account of modality.)

And on any halfway natural reading of "sufficient", it is true as a conceptual matter that if something sufficient for $S$ is possible, then $S$ is possible as well.[45]

Now consider the direction from left to right: $S$ is possible *only* if "how is $S$ possible?" has an answer. This forces us to speculate a little on the kind of $S^+$ the questioner is looking for. I hear her as issuing a challenge:

> You think that $S$ is possible but I suspect that this is only because you have neglected the matter of $T$. I therefore ask you: is $S$ possible in the $T$ way, or is it possible in the not-$T$ way? According to you, for instance, there can be a town whose barber shaves all and only the town's non-self-shavers. But are we to think of this barber as shaving himself, or as not shaving himself?

Understood like this, the question "how is $S$ possible?" has an answer iff it is possible that $S$ & $T$, or else possible that $S$ & $\neg T$.[46] Since there is no way of telling in advance what unresolved issue $T$ might have attracted the questioner's attention, we arrive at the claim that, pick any $T$ you like,

(11″) $\Diamond S$ only if $\Diamond S^+_1$ or $\Diamond S^+_2$—where $S^+_1 = S \,\&\, T$ and $S^+_2 = S \,\&\, \neg T$.

This says that possibility is *expansive*: nothing is possible which cannot be expanded into a more *inclusive* possibility, inclusiveness being judged along any dimension you like.[47] To come at it from the other side, there can be no refuge from impossibility in refusing to take a stand on matters left open; if an impossibility would result *however* these matters were decided, you've got an impossibility already.[48] Either way, expansiveness looks like a conceptual truth.

[45] Sufficing might be a matter of metaphysical necessitation. Then (11′) is the modal-logical truth that anything necessitated by a possibility is itself possible. Or it might be a subjunctive affair: if $S^+$ were the case, $S$ would be the case as well. Again, from this and the fact that $S^+$ is possible, $S$'s possibility logically follows. Of course, (11′) continues to be a logical truth on stronger readings of "suffices", e.g., $S^+$ necessitates $S$ and understandably so, or it is a priori that $S$ would be the case if $S^+$ were the case.

[46] It might be objected that "possibly, $S$ & $T$" has the wrong form for an answer to "how is $S$ possible?". The questioner does not want to know what could have been the case *together* with $S$, but what could have been the case *to bring it about* that $S$. This assumes that the "how" in "how is $S$ possible?" has got to be one of *means* rather than *manner*. "Possibly $S$ & $T$" answers "how is $S$ possible?" in much the same way as "they fit stacked together like so" answers "how do all of those dominoes fit into that little box?" The emphasis on manner is only natural given that doubts about "$S$ is possible" are prompted by the thought, not that there is no *basis* for $S$'s possibility—some possibilities are surd after all—but that there is a positive *obstacle* to its possibility. This thought takes the form indicated in the text: scenarios according to which $S$ have a way of falling apart when one tries to flesh them out so as to render a verdict about $T$. (For more on this theme, see yablo 1993; Ch. 2 above. caps

[47] Subject to the usual qualifications about semantical paradox.

[48] Otherwise one could say: It's perfectly possible to have a barber who shaves all and only the non-self-shavers; what's impossible is that *along* with a specification of who if anyone shaves the barber.

## XII

Voila! possible world semantics without possible worlds. Because what we have in (11) is a structural analogue of Lewis's (13) in which worlds do not figure[49]—an analogue, moreover, with some claim to be regarded as analytic. But I can imagine various questions and objections, starting with the objection that since (11) is *circular*—modal notions appear on its right-hand side—it fails to provide a reductive analysis of modality.

There is no denying the circularity. "$S^+$ is possible" does not count as a correct answer to "how is $S$ possible?" unless it is *true*, which means that $S^+$ has got to really be possible.[50] But why exactly is this an *objection*? To reply that (11) is up against (13), which being noncircular *can* function as an analysis, just pushes the question back a step. Why should the potential for functioning as an analysis count so heavily in (13)'s favor?

The answer may seem obvious from something already mentioned, that the argument for (13) is in terms of the theoretical services it offers. Lewis ~~offers~~ gives an impressive catalogue of these services in his book *On the Plurality of Worlds*; here we will have to limit ourselves to a single example. How, without an analysis like (13), are we to understand why this, that, and the other should be the laws of modality? True enough, it can be proved in pure mathematics that

> *if* modal operators can be correctly analysed in so-and-so way [as quantifiers over worlds], *then* they obey so-and-so systems of modal logic.[51]

This conditional gets us nowhere, however, unless we are in a position to discharge the antecedent. And to get ourselves into that position, we need to count an analysis like (13) into our belief system.

But although this is often said, it is hard to see how the application depends on (13)'s constituting an *analysis*. As long as (13) is *true* (the left-hand side holds just when the right one does) and *known* to be, the deduction of modal laws from the laws of quantification would appear to go through just the same. And now we begin to lose our grip on where the insistence on a *reductive* correlation is coming from. If the choice of a correlation is to be driven by considerations of theoretical utility ~~alone~~, it should be enough ~~nothing should matter but~~ that

(*a*) the correlation is *there*,
(*b*) it is *comprehensive*,
(*c*) it is not *itself* unduly mysterious, and
(*d*) it can be used to dispel *other* mysteries.

[49] I leave (12) aside for now since Lewis has questioned its claim to be called reductive.

[50] Further circularities creep in later when (11) gives way to (17).

[51] Lewis (1986), 17; italics added.

And, to twist around a famous remark of Lewis's, it is not clear why a nonreductive correlation must have trouble with these conditions—unless you beg the question by saying that it already *is* trouble.

For Lewis, a reductive correlation is the only kind worth having—so, anyway, it is usually assumed, and given that the correlation he defends is the most reductive available, there seems little reason to doubt it. But the fact is that one can read quite a way into Lewis's book without reductiveness coming up as an explicit desideratum.[52] Most of the time it sounds as though the reason for believing in his ''concrete correlation'' (13) is that it exemplifies better than any competing correlation the values expressed above, such as comprehensiveness and theoretical power.[53] That the correlation is reductive besides appears to be just gravy, *except* to the extent that it helps with the other desiderata.

Starting about halfway through the book, however, we find Lewis objecting to certain ersatzist alternatives to (13) that they smuggle modal notions in on the right-hand side. Apparently, then, reductiveness is something that Lewis is prepared to insist on. Why? Is it because he takes the same view of modality that Jerry Fodor does of intentionality, viz., that if it is really real, it must really be something else? I doubt it. No one could be less sentimental than he about the trade-offs philosophers are occasionally forced to make between ideology and ontology. If the price were right, he would be as willing as anyone to buy relief from unwelcome entities by taking on a primitive notion or two.[54] It's just that in *this* case, the price is not right; in fact, the trade-offs play out the other way.[55] A reductive account of modality is so enormously valuable as to more than compensate us for the humongous ontology of worlds.

[52] It does come up in passing, for instance, in the passage just quoted.

[53] So, Lewis objects to linguistic ersatzism that it misclassifies ''alien'' possibilities as impossible, and to magical ersatzism that it postulates a mysterious making-true relation. Lewis's own favored correlation has been charged with falsely ''predicting'' the impossibility of island universes, a charge he takes dead seriously.

[54] Here is a typical expression of his attitude:

> I conclude that linguistic ersatzism must indeed take modality as primitive. If its entire point were to afford an analysis of modality, that would be a fatal objection. But there are many theoretical services left for a version of ersatzism to render; even if it cannot analyze modality away. So it is open to an ersatzer to pay the price, accept modality as primitive, and consider the proposal well worth it on balance. Many ersatzers . . . see the contest between genuine and ersatz modal realism in just that way: there is a choice between unwelcome ontology and unwelcome primitive modality, and they prefer the latter. That seems to me a fair response on their part, but of course not conclusive. (Lewis, (1986), 156)

[55] Thus, he says:

> *If* our work is directed to ontological questions only, we may help ourselves to any primitives we please, so long as we somehow understand them. But if our work is directed to ontological and analytic questions both, . . . then we are trying at once to cut down on questionable ontology and to cut down on primitives; and it is fair to object if one goal is served at too much cost to the other. (ibid. 157)

All right, but now we need a distinction. Is it that a reductive account of modality is so *intrinsically* valuable as to compensate us, etc.? This is hard to take seriously. Faced with a no-strings-attached decision between the humongous ontology of concrete worlds, on the one hand, and letting possibility be what it is and not another thing, on the other, most of us would know which way to jump.

So the claim has got to be that a reductive analysis of modality is so *extrinsically* valuable as to compensate us, etc.—that is, so valuable from the point of view of desiderata *other* than reductiveness. And now we are back where we started: if the concrete correlation is better, that is *not* because it is nonreductive per se but because it outperforms the competition in other respects. Whereupon we're entitled to ask why a nonreductive correlation like (11) couldn't do just as well.

Or indeed better. Because if a correlation is going to do theoretical work, it's very important that it *be* there, that it be comprehensive, and that it not be *itself* unduly mysterious. And between (11) and Lewis's (13), the verdict is clear. (11) is bordering on analytic, which is about as good as you can do in the truth and comprehensiveness and unmysteriousness departments. Whereas (13), on top of being prima facie as improbable as anything ever was, is baffling even on the supposition of its truth. If an oracle convinced us that ours was one of a large number of spatiotemporally isolated universes, each enacting ~~modal facts about~~ scenarios possible in the others while they all the while returned the favor, this would be regarded as the most amazing coincidence on record.[56]

## XIII

Now, the natural and proper reply to this is that Lewis's concrete correlation; its existence momentarily granted, so thoroughly *creams* the competition at

[56] Comprehensiveness may be a problem too, since (13) as Lewis understands it rules out the possibility of spatiotemporally unrelated (''island'') universes. Far from hushing this problem up, Lewis has done a good deal to publicize it:

> The intuitive case that island universes are possible has been much strengthened by a recent argument in John Bigelow and Robert Pargetter, ''Beyond the Blank Stare''. . . . First, mightn't there be a world of *almost* isolated island universes, linked only by a few short-lived wormholes? And mightn't the presence of the wormholes depend on what happens in the islands? And then wouldn't it be true that if the goings-on in the islands had been just a little different, there wouldn't have been any wormholes? Then wouldn't there have been a world of altogether isolated islands? (Lewis (1990), 223)

His rejection of island universes puts Lewis in the prima facie awkward position of maintaining that there is something—the mereological sum of all the various worlds—such that a thing like that cannot be. But reject them he must if he wants to hold on to his definition of worlds as maximal spatiotemporally connected objects. A lot is riding on this definition here since all the alternatives that come to mind are explicitly or implicitly modal. Allow island universes, and it is not clear whether (13) can still be regarded as a reduction. (See Lewis, (1986), 69 ff.)

dispelling modal mysteries that we should take it on board however prima facie improbable and however baffling if true. (13) has been undersold, in other words. This is something we'll get to in a moment; let us consider first a way in which (11) has been *over*sold.

Again and again (11) has been billed as *close* to a conceptual truth, or *bordering* on analytic. Why the hedge, if (11′) and (11″) are conceptual truths and (11) is their conjunction? The hedge is because (11) is *not* their conjunction; it is slightly but crucially stronger. This comes out if we compare what (11) tells us on the supposition that *S* is possible with what (11″) tells us on the same supposition. According to (11), "how is *S* possible?" has a correct answer. But (11″) only as it were *exhibits* this answer,[57] without testifying to its existence. [(11″) does perhaps tell us that "how is *S* possible?" is correctly answerable, in some appropriate sense of that word. But answerability is one thing, having a correct answer another.[58]] The upshot is that (11) is a conceptual truth only modulo the existence of answers with the requisite contents. And that is a very big modulo. It begins to appear that, although on friendly terms with conceptual truths, (11) is not itself actually even true.[59]

Not good. And we have yet to consider the other reply: namely, that (13) is needed *regardless* of (11)'s truth-value on account of its greater effectiveness against modal mysteries. Take again the "mystery" of the laws of modality, using

(14) if $\Diamond S$ and $\Box T$ then $\Diamond\, (S \,\&\, T)$

as a typical instance. Why is it that counterexamples to this never turn up? (13) has an explanation to offer: If *S* is possible, then there is an *S*-world, call it *W*. *W* cannot be a not-*T*-world, since there aren't any; so, worlds being complete, it must be a *T*-world. *W* is accordingly an (*S* & *T*)-world, which means that possibly *S* & *T*. Now try the same thing using (11). Since *S* is possible, "how is *S* possible?" has a correct answer *A*. *A* is clearly *not* an answer to "how is not-*T* possible?", for *T* is necessary. But this still doesn't give us an answer to "how is *S* & *T* possible?", for *A* may well be *silent* on the subject of *T*.

[57] And this only schematically.

[58] One could attempt to deny the distinction, maintaining that Q is correctly answerable iff it has a correct answer iff there is a fact of the matter as to IQ. There is certainly something to be said for this view. " 'What would China do if the United States recognized Taiwan?' has an answer" does not intuitively make an existence claim; it says that there is a fact of the matter as to what China would do. By the same token, when (11) assures us that "how is *S* possible?" has an answer, this means only that there is a fact of the matter as to how *S* is possible. And (11″) gives us the same assurance when it tells us that *S* is possible either *T*-ishly, or (failing that) not-*T*-ishly. Now, though, we have to decide whether "there is a fact of the matter as to . . ." involves genuine quantification over facts. If it does, we lose; the cause of ontology-free theoretical power is hardly served by trading one ontology for another. If it doesn't, we lose again; give up the quantification and the theoretical power goes too.

[59] Another option is to read "there is an answer" in (11) as "there *could* have been an answer, such-and-such conditions holding fixed". Similarly, one could read "there is a world" in (12) and (13) as "there could have been a world"—a type of quantifier discussed in Fine (1977). I will be taking a different line, but this one strikes me as ~~well~~ worth pursuing.

Examples could be multiplied. (13) has a real analytic advantage for the simple reason that worlds are complete while answers to "how is that possible?" questions tend to leave a great deal undecided.

Someone might wonder why completeness should be such a sticking point. Can't possible-world semantics equally well be done in terms of incomplete or "partial" worlds?[60] It is true that these partial worlds have to be conceived as subject to a *refinability* condition: given any *T* you like,

(15) any partial world at which *S* is true has a refinement at which *S* is true and *T* is true or else one at which *S* is true and *T* is false.[61]

But this condition seems very much in the spirit of (11)'s portrayal of possibility as *expansive*: *S* is possible only if it is possible together with *T* or possible together with not-*T*.

Looking a little closer, though, we see that that refinability *adds* something to expansiveness that is necessary for serious modal mathematics; it says not merely that *S* is possible only if its conjunction with *T*, or with not-*T*, is possible, but that any *witness* to S's possibility can be built up into a *witness* to the possibility of one of these conjunctions. If a version of this held for answers—if we could be assured that

(16) any correct answer to H $\Diamond$ *S*? has a refinement that correctly answers H $\Diamond$ (*S* & *T*)? or else one that correctly answers H $\Diamond$ (*S* & ¬*T*)?

—then the analytic gap between (11) and (13) would be significantly narrowed.[62] But unless we have it in mind to shoot ourselves in the foot by indulging at this late date in wishful platonic thinking, (16) is not something we can afford to assume. The only answers we can safely rely on in this context are the ones that have actually cropped up in conversations or on paper. And these, it seems clear, are not closed under refinement; time being short and attention limited, they eventually peter out.

## XIV

All right; people have not in fact gotten around to giving all the answers our approach needs. But having come this far, it seems a shame to retreat before so drearily medical a difficulty. And in fact we don't have to. The insight that (11) is struggling to express is that *S* is possible iff it is possible *somehow*. And

[60] These are sometimes called "situations" or "possibilities".

[61] Cf. Humberstone (1981) and Forbes (1985), 18–22 and 43–7. True refinability is a more complicated affair than (15) suggests, but the differences are not important here.

[62] For example, we could explain (14) by saying that any answer to "how is *S* possible?" is refinable into an answer to "how is *S* & *T* possible?", since given *T*'s necessity "how is *S* & ¬*T* possible?" is unanswerable.

we know from section IX that if "somehow" is going to be understood in terms of quantification over answers, that quantification needs to be seen as *feigned* or conducted in a spirit of make-believe.

As to the form of the make-believe, we can let our present difficulties be our guide. The difficulty about the *existence* of answers to "how is $S$ possible?" is met by supposing that whenever "how is $S$ possible?" is correctly answer*able* (it FN:63 is possible that $S^+$), a correct answer to it actually exists.[63] The difficulty about the *refinability* of these answers is met by supposing that any correct answer to "how is $S$ possible?" has a correct refinement that either affirms $T$ or denies it. These two ideas together can be called the *refinable-answer story*, or RAS. All that remains is to reconceive the quantifier in (11) as "feigned" by prefixing it with an "according to RAS" operator. The biconditional

(17) $\Diamond S$ iff according to RAS, H $\Diamond$ $S$? has a correct answer

that results is a conceptual truth that enables free back-and-forth motion between possibility, on the one hand, and existential quantification over a single matrix of as-determinate-as-you-need witnesses on the other. To the extent that (12) and (13) have their analytic power as catalysts in this sort of transition, (17) can offer FN:64 the same power at a fraction of the ontological cost.[64]

All right, but why stop there? If we are willing to stipulate that incomplete answers are *partly* refinable, why not go whole hog and make them *completely* refinable? The *determinate-answer story* is just like the refinable-answer story, except that correct answers to "how-possible?" questions are always refinable into correct answers leaving nothing unsettled. This gives us a still closer FN:65 approximation to the standard analysis:[65]

(18) $\Diamond S$ iff according to DAS, H $\Diamond$ $S$? has a determinate correct answer.

And now for a final weird twist. "Answer" is a theoretical notion whose proper treatment is to some extent up for grabs. No doubt answers are often best seen as representations. ("*Here* is your answer" I say: "your cousin Giorgio.") Sometimes

[63] As discussed above, $S^+$ should suffice for $S$. There might be "pragmatic" conditions on $S^+$ as well; it should speak to the questioner's doubts. (11″) guarantees that "how is $S$ possible?" is correctly answerable in a way that addresses these doubts whatever they may be, provided that $S$ is indeed possible.

[64] Compare Rosen (1990). The present paper represents one possible development of Rosen's next-to-last paragraph:

> Throughout I have supposed that fictionalism, like modal realism, aims to be a *theory of possibility*. . . . But note that this assumption is not strictly necessary given the modest problem we began with. All Ed ever wanted was license to move back and forth between modal claims and claims about worlds . . . it is one thing to embrace these biconditionals—even to embrace them as a body of necessary truths—and another to regard them as providing analyses. . . . This timid fictionalism of course raises as many questions as it answers. Still it must be granted that many of the objections we have mentioned . . . simply do not arise for this view. (ibid. 233–4).

[65] Albeit to the version (12) that quantifies over abstract representations of concrete worlds rather than the worlds themselves.

though it is the thing *represented* that seems better suited to the role. ("*There* is your answer", I say, with a nod at your approaching cousin.) So far we have been assuming a version of the first approach; more or less determinate answers to "how is $S$ possible?" have been more or less comprehensive representations according to which $S$. But once having made the switch to *fully* determinate answers, and *pretended* ones at that, the second option becomes suddenly attractive.

You want to know how blue swans are possible, in full and comprehensive detail? *There* is your answer, I say, gesturing or pretending to gesture, to the best of my expressive abilities, at a concrete I-and-all-my-surroundings world FN:66 wherein swans really are blue.[66] Reconceive determinate answers like this and the determinate-answer story becomes the *many-worlds story* MWS: $S$ is possible only if "how is $S$ possible?" has an answer taking the form of a concrete world at which $S$ is true. And (18) becomes

(19) $\Diamond S$ iff according to MWS, there is a concrete world at which $S$.

You might think that Lewis would welcome (19) with open arms; isn't the FN:67 many-worlds story *his* story? *It is not.*[67] Both stories tell of an array of concrete worlds. But Lewis's story portrays these worlds as independently constituted, not inherently modal entities which somehow nevertheless contrive to constitute the ground of modal truth. The present story conjures worlds up from within the structure of possibility itself—from what we called its *expansive* quality. Worlds are the ideal objects of our efforts to give more and more specific answers to the question "how could that be?". This is how we can know that (19) is true—that FN:68 $S$ is possible iff according to the story, there is an $S$-world.[68] rom

## XV

Thirty or so years ago, before the campaign to make modal metaphysics honest had gotten seriously under way, the talk was less of *worlds* than of something called the world *metaphor*. One reason this sort of talk fell out of favor was Lewis's Quinean scrupulosity about ontological commitment. Another reason, however, was that it was never quite clear what the talk *meant*. One saw what

[66] Sentential and propositional answers to "how is $S$ possible?" still exist on this view, but they interest us mainly as *presentations* of the fully determinate answers otherwise known as concrete worlds.

• Q2 [67] For the same reasons, it is not Gideon Rosen's (1990) story either. ~~pp. 333–5.~~

[68] Someone might think that MWS was lacking in substantive content. But a story's content is not exhausted by what is explicitly written down (see the next section). If $S$ is possible, then according to MWS, there is a world in which $S$. Since blue swans are possible, according to MWS there is a world at which blue swans exist; since they are possible together with a German victory in the First World War, according to MWS there is a world like that as well. How claims like this make their way into the content of Lewis's story (as elaborated by Rosen) is a nontrivial question. 327-8 See Rosen, (1990), ~~227–8.~~

worlds qua metaphors were supposed to *do*: shed metaphysical light on modality just by making themselves available to theoretical contemplation.[69] But it was never explained what they *were* that this was within their powers.

All the same, it seems to me that the pre-Lewis approach to these matters was onto something. Talk about worlds *is* metaphorical, or close enough not to matter. Some of the ground ~~argument~~ for this has already been laid ~~is already in place:~~ world-talk as it features in (19) is fictional, and so the sort of thing we are to *pretend* or *imagine* is true. The next step is to observe that the pretense is in a quite particular spirit, a spirit characteristic of metaphor.

Almost wherever there is disciplined pretense or imagination, there is something that can be considered a *game of make-believe*.[70] Take for instance the games we play with representational paintings and novels. Standing before Caravaggio's "Bacchus", we are supposed to imagine ourselves meeting the gaze of a figure entreating us with a glass of wine, when all that is really there is marks on a canvas. Reading "The Speckled Band" by Arthur Conan Doyle, we are supposed to imagine ourselves reading (not sentences strung together for dramatic effect but) reports of a detective's activities compiled by one who knows whereof he speaks.

Both of these games can lay claim to some sort of official sanction; the reason we are supposed to imagine in such-and-such ways is that *that* is what the author intended, or that is how the institution of painting works. Other games derive their shape and authority from humbler sources. Some are grounded in ad hoc arrangements ("these clumps of mud can be the pies") or the understanding of a moment ("look out, I've got your nose"). Some are grounded in nothing at all, arising among like-minded pretenders of their own accord. (Finding ourselves in an unexpectedly swanky hotel room, we begin putting on airs and acting the part.) The common thread here—the factor that links all make-believe games together—is that they call upon their participants to pretend or imagine that certain things are the case. These "to-be-imagined" items make up the game's *content*, and to elaborate and adapt oneself to this content is often the game's very point.[71]

Often, but not always; an alternative point suggests itself when we reflect that all but the most boring make-believe games are played with *props*, whose game-independent properties help to determine what it is that the players are to imagine or pretend. Nowhere in the rules of mud pies does it say that Sam's

[69] See the epigraph.

[70] "Almost" because the pretense has to be disciplined in the right way. I'm not sure what the right way is, but at least this much is true. There is no make-believe game if imaginings are *forbidden* but none are *prescribed* (Albanians under Enver Hoxha were told not to imagine life in the West) or if they are prescribed but on a basis having not enough to do with what they are imaginings *of* (as in a biofeedback game where contestants try to lower ~~raise~~ their heart rates just by the exercise of their imaginations).

[71] Better, such-and-such is part of the game's content if "it is to be imagined . . . *should the question arise*, it being understood that often the question *shouldn't* arise" (Walton (1990), 40). Subject to the usual qualifications, the ideas about make-believe and metaphor in this and the next few paragraphs are all due to Walton. See Walton (1993).

pie is too big for the oven; we are to imagine this because Sam's clump of mud doesn't fit into the hollow stump. Nowhere in the rules of the Holmes game does it say that Holmes lives nearer to Hyde Park (in London) than to Central Park (in New York). If this is fictionally the case, the facts of nineteenth-century geography deserve part of the credit.[72]

Now, a game whose content reflects in part the properties of worldly props can be seen in two quite different lights. What ordinarily happens is that we take an interest in the props because and to the extent that they influence the game's content; one tramps around London in search of 221B Baker Street for the light it may shed on what is true according to the Holmes stories.

*But in principle it could be the other way around*: we could be interested in a game's content because and to the extent that it informed us about the props. This would not stop us from *playing* the game, necessarily, but it would tend to confer a different significance on our moves. Pretending within the game to assert that blah would be a way of giving voice to a fact holding *outside* the game: the fact that the props are in such-and-such a condition, viz., the condition that makes blah a proper thing to pretend to assert. One can even imagine there being *advantages* to this style of expression. It might be the only way open to us of putting the indicated fact into words. Or, it might be the *best* way of putting the fact into words, the one with the happiest cognitive and motivational effects.[73]

Using games to talk about game-independent reality makes a certain in-principle sense, then. But is such a thing ever actually done? A case can be made that it is done all the time, not indeed with explicit self-identified games like "mud pies" but with impromptu everyday games hardly scratching the surface of consciousness. Some examples of Walton's suggest how this could be so:

> Where in Italy is the town of Crotone? I ask. You explain that it is on the arch of the Italian boot. 'See that thundercloud over there—the big, angry face near the horizon,' you say; 'it is headed this way.' . . . We speak of the saddle of a mountain and the shoulder of a highway. . . . All of these cases are linked to make-believe. We think of Italy and the thundercloud as something like pictures. Italy (or a map of Italy) depicts a boot. The cloud is a prop which makes it fictional that there is an angry face. . . . The saddle of a mountain is, fictionally, a horse's saddle. But our interest, in these instances, is not in the make-believe itself, and it is not for the sake of games of make-believe that we regard these things as props. . . . [The make-believe] is useful for articulating, remembering, and communicating facts about the props—about the geography of Italy, or the identity of the storm cloud . . . or mountain topography. It is by thinking of Italy

[72] The example is adapted from Lewis (1978).

[73] The kind of point I am gesturing at is familiar from ~~might be guessable from~~ the literature on indexicality, for instance, Perry (1979) and Kaplan (1989). See also Stern (1985). A metaphor that shares its *content* with some literal paraphrase might still be indispensable due to its special *character*; that the content is arrived at by way of a contextually salient make-believe game might make a world of difference to its cognitive reception.

or the thundercloud . . . as potential if not actual props that I understand where Crotone is, which cloud is the one being talked about.[74] FN:74

Games of make-believe, Walton says, can be ''useful for articulating, remembering, and communicating facts about [their] props''. He might have added that they can make it easier to reason about such facts, to systematize them, to visualize them, to spot connections with other facts, and to evaluate potential lines of research. That similar virtues have been claimed for metaphors is no accident, if Walton is right in his account of how metaphor works:

> The metaphorical statement (in its context) implies or suggests or introduces or calls to mind a (possible) game of make-believe. The utterance may be an act of verbal participation in the implied game, or it may be merely the utterance of a sentence that *could* be used in participating in the game. In saying what she does, the speaker describes things that are or would be props in the implied game. [To the extent that paraphrase is possible] the paraphrase will specify features of the props by virtue of which it would be fictional in the implied game that the speaker speaks truly, if her utterance is an act of verbal participation in it.[75] FN:75

Stripped to essentials, the account is this: A metaphor is an utterance *U* that portrays its subject as of a kind to make *U* pretense-worthy in a game that *U* itself suggests. The game is played not for its own sake but to make clear *which* game-independent properties are being attributed; they are the ones that do or would confer legitimacy upon the utterance construed as a move in the game.

Is it just me, or do utterances about possible worlds appear to fit the bill pretty exactly? ''There are worlds in which blue swans exist'' suggests a game in which we pretend that all and only the things that *could* happen in this world *do* happen in some world or other. The point of the utterance is to say that the modal facts are such as to make ''there are blue~~s~~ swan worlds'' pretense-worthy in the game—to say, in other words, that blue swans could have existed. ~~I conclude that~~ even if the tradition did not know quite what it *meant* in calling worlds metaphors, that is what they plausibly are.[76] caps FN:76

[74] Walton (1993), 40–1.

[75] Ibid. 46. (I should say that Walton does *not* take himself to be offering a general theory of metaphor.) Walton goes on to say that unparaphrasable metaphors ''may still amount to descriptions of their (potential) props'' (ibid.). If he is right, then it becomes suddenly clear how, even if modal reality had nothing to *do* with worlds, there could still be modal truths requiring quantification over worlds for their expression. The point generalizes. Ineliminable quantification over blahs does not count in favor of blahs *existing* unless it can be shown that the quantifier is not metaphorical. Add to this that the metaphorical–literal distinction is deeply and irremediably inscrutable, and the whole project of Quinean ontology is thrown into ~~considerable~~ doubt, while Carnap's position that there is no worthwhile activity of trying to puzzle out what ''really'' exists begins to look ~~notably~~ less insane. These matters are discussed in ~~''Does Ontology Rest on a Mistake?'' (Yablo, S., ''Does Ontology Rest on a Mistake?'' *Aristotelian Society* (1998) Supp (72) pp 229–261).~~ rom Yablo 1998.

[76] Some of the many additional topics that need attention are: iterated modalities, transworld identity, grades of modality, impossible worlds, and modal epistemology.

## REFERENCES

Baker, C. L. (1968). ''Indirect Questions in English'' (Ph.D. dissertation, University of Illinois).

Benacerraf, Paul (1973). ''Mathematical Truth''. *Journal of Philosophy*, 19: 661–79.

Engdahl, E. (1986). *Constituent Questions*. Dordrecht: Reidel.

Etchemendy, John (1990). *The Concept of Logical Consequence*. Cambridge, Mass.: Harvard University Press.

Fine, Kit (1977). ''Postscript: Prior on the Construction of Possible Worlds and Instants''. In Arthur Prior and Kit Fine (eds.), *Worlds, Times, and Selves*, London: Duckworth, 116–61.

Forbes, Graeme (1985). *The Metaphysics of Modality*. Oxford: Clarendon Press.

Groenendijk, J., and Stokhof, M. (1989). ''Semantics of Interrogatives''. In Gennaro Chierchia, Barbara Hall Partee, and Raymond Turner (eds.), *Properties, Types, and Meaning*, Dordrecht: Kluwer, 21–68.

Grover, Dorothy (1972). ''Propositional Quantifiers''. *Journal of Philosophical Logic*, 1: 111–36; repr. in Grover (1992).

____ (1992). *A Prosentential Theory of Truth*. Princeton: Princeton University Press.

____ Camp, Joseph, and Belnap, Nuel D. Jr (1975). ''A Prosentential Theory of Truth''. *Philosophical Studies*, 27: 73–124; repr. in Grover (1992).

Higginbotham, J. (1993). ''Interrogatives''. In K. Hale and S. J. Keyser (eds.), *The View from Building 20*, Cambridge, Mass.: MIT Press, 195–228.

Humberstone, I. L. (1981). ''From Worlds to Possibilities''. *Journal of Philosophical Logic*, 10: 313–39.

Kaplan, David (1989). ''Demonstratives''. In Joseph Almog, John Perry, and Howard Wettstein (eds.), *Themes from Kaplan*, Oxford: Oxford University Press, 481–564.

Karttunen, Lauri (1977). ''Syntax and Semantics of Questions''. *Linguistics and Philosophy*, 1: 3–44.

Katz, Jerrold (1995). ''What Mathematical Knowledge Could Be''. *Mind*, 104: 491–552.

Kung, G. (1977). ''The Meaning of the Quantifiers in Lesniewski''. *Studia Logica*, 26: 309–22.

Lewis, David (1978). ''Truth in Fiction''. *American Philosophical Quarterly*, 15: 37–46.

____ (1979). ''Possible Worlds''. In Michael J. Loux (ed.), *The Possible and the Actual: Readings in the Metaphysics of Modality*, Ithaca, NY: Cornell University Press, 182–9.

____ (1986). *On the Plurality of Worlds*. New York: Basil Blackwell.

____ (1990). ''Review of Armstrong, *A Combinatorial Theory of Possibility*''. *Australasian Journal of Philosophy*, 70: 211–224.

Perry, John (1979). ''The Problem of the Essential Indexical''. *Noûs*, 13: 3–21.

Prior, A. N. (1971). *Objects of Thought*. Oxford: Oxford University Press.

Quine, W. V. O. (1948). ''On What There Is''. *Review of Metaphysics*, 2: 21–38.

____ (1960). *Word and Object*. Cambridge, Mass.: MIT Press.

____ (1969). ''Speaking of Objects''. In *idem, Ontological Relativity and Other Essays*, New York: Columbia University Press, 1–25.

Rosen, Gideon (1990). ''Modal Fictionalism''. *Mind*, 99: 327–54.

Simons, P. (1985). ''A Semantics for Ontology''. *Dialectica*, 39: 193–216.

Stalnaker, Robert C. (1979). ''Possible Worlds''. In Michael J. Loux (ed.), *The Possible and the Actual: Readings in the Metaphysics of Modality*, Ithaca, NY: Cornell University Press, 225–34.

Stern, Josef (1985). ''Metaphor as Demonstrative''. *Journal of Philosophy*, 92: 677–710.

Walton, Kendall (1990). *Mimesis and Make-Believe*. Cambridge, Mass.: Harvard University Press.

——(1993). ''Metaphor and Prop-Oriented Make-Believe''. *European Journal of Philosophy*, 1: 39–57.

Yablo, S. (1993). ''Is Conceivability a Guide to Possibility?''. *Philosophy and Phenomenological Research*, 53: 1–42; Ch. 2 above.

^

^Yablo, S. (1998). "Does Ontology Rest on a Mistake?" *Proceedings of the Aristotelian Society*, Supp(72): 229–261.

**Queries in Chapter 7**

Q1. Please check author edit not clear.

Q2. Please check and confirm the author correction here.

# 8

# Mental Causation

## 1.

Writing to Descartes in 1643, Princess Elisabeth of Bohemia requests an explanation of "how man's soul, being only a thinking substance, can determine animal spirits so as to cause voluntary actions".[1] Agreeing that "the question which your Highness raises [is] one which can most reasonably be asked", Descartes launches with his reply a grand tradition of dualist apologetics about mind–body causation that has disappointed ever since. Apologetics are in order because, as Descartes appreciates, his conception of mental and physical as metaphysically separate invites the question, "how, in that case, does the one manage to affect the other?"; and because, having invited the question, he seems unable to answer it. Much as the Cartesian epistemology breeds skepticism, then, the metaphysics breeds epiphenomenalism: the theory that our mental lives exercise no causal influence whatever over the progress of physical events.

That was the price Descartes paid for his dualism, someone might say. Why should epiphenomenalism concern anyone today? Part of the answer is that dualism is not dead, only evolved. Immaterial minds are gone, it is true, but mental *phenomena* (facts, properties, events) remain. And although the latter are admitted to be physically *realized*, and physically *necessitated*, their literal numerical *identity* with their physical bases is roundly denied.[2]

Thanks to Louise Antony, Simon Blackburn, Paul Boghossian, Donald Davidson, Graeme Forbes, Sally Haslanger, Jaegwon Kim, Vann McGee, Sarah Patterson, Gideon Rosen, Larry Sklar, William Taschek, David Velleman, Ken Walton, Catherine Wright, Crispin Wright, and two anonymous readers for reactions and advice. Versions of the paper were read at Chicago Circle, the University of North Carolina at Chapel Hill, and the University of Western Ontario; discussions there were extremely useful, and I'm grateful to all who took part. Research was supported by the National Endowment for the Humanities and the Social Sciences and Humanities Research Council of Canada.

[1] Descartes (1969), 373. In the "Fifth Objections", Gassendi puts a similar question: "How can there be effort directed against anything, or motion set up in it, unless there is mutual contact between what moves and what is moved? And how can there be contact without a body . . . ?" (Descartes (1984), 236 ff.).

[2] In case it seems odd to describe the picture just outlined as dualist, bear in mind that all I mean by the term is that mental and physical phenomena are, contrary to the identity theory, *distinct*, and,

Surely, though, it is hard to imagine a dualism more congenial to mental causation than this! So it would seem. But epiphenomenalism has been evolving too; and in its latest and boldest manifestation, this is all the dualism it asks for. As a result we find ourselves in a somewhat paradoxical situation. Just when the conditions for accommodating mental causation have become little short of ideal, epiphenomenalist anxiety rages higher than ever. Nor is this a pretended anxiety, put on for dialectical purposes but posing no genuine danger to established views. Some say we must simply make our peace with the fact that "the mental does not enjoy its own independent causal powers".[3] Others would renounce (distinctively) mental phenomena altogether, rather than see them causally disabled.[4] Radical as these proposals are, they are backed by a straightforward line of reasoning.

"How can mental phenomena affect what happens physically? Every physical outcome is causally assured already by preexisting physical circumstances; its mental antecedents are therefore left with nothing further to contribute." This is the *exclusion argument* for epiphenomenalism. Here is the argument as it applies to mental events; for the version which applies to properties, replace 'event *x*' with 'property *X*'.[5]

(1) If an event *x* is causally sufficient for an event *y*, then no event $x^*$ distinct from *x* is causally relevant to *y* (*exclusion*).[6]

(2) For every physical event *y*, some physical event *x* is causally sufficient for *y* (*physical determinism*).[7]

contrary to eliminativism, *existents*. That this much dualism is acceptable even to many materialists is in a way the point: having broken with dualism's Cartesian version over its vulnerability to epiphenomenalism, they find to their horror that epiphenomenalism lives equally happily on the lesser dualism latent in their own view.

[3] Kim (1983), 54. Kim does allow the mental a role in what he calls *epiphenomenal* causal relations, and he says that macrophysical causation is epiphenomenal in the same sense. My position is that neither sort of causation is epiphenomenal in any interesting sense.

[4] This is particularly clear in Schiffer (1989, ch. 6), who rejects mental *properties* on the ground that they would be causally superfluous, and makes mental *events* a subspecies of physical events on the theory that they would *otherwise* be causally superfluous.

[5] So '*x*' and '$x^*$' become '*X*' and '$X^*$', and where either is prefixed by 'event', this becomes 'property'; 'event *y*' and 'event *z*' are unaffected. Although causes and effects are events, properties as well as events can be causally relevant or sufficient. I try to remain neutral about what exactly causal sufficiency and relevance amount to (e.g., causal sufficiency could be sufficiency-in-the-circumstances, or it could be absolute). Versions of the exclusion argument are found in Feigl (1970); Malcolm (1982); Goldman (1969); Campbell (1970); Kim (1979) and Kim (1989); Sosa (1984); Honderich (1988); and Macdonald and Macdonald (1986). Objections similar in spirit to the exclusion argument are sometimes raised against the causal claims of other phenomena apparently unneeded in fundamental physical explanation (e.g., macroscopic and color phenomena). This paper offers a potentially general strategy of response.

[6] Some authors use a slightly weaker premise: if *x* is causally sufficient for *y*, then unless *y* is causally overdetermined, every distinct event $x^*$ is causally irrelevant (see note 53).

[7] (2) could obviously be questioned, but I take it that physical determinism isn't the issue. For one thing, the conviction that mind makes a causal difference is not beholden to the contemporary opinion that determinism is false, and would remain if that opinion were reversed. Second,

(3) For every physical event *x* and mental event $x^*$, *x* is distinct from $x^*$ (*dualism*).
(4) So: for every physical event *y*, no mental event $x^*$ is causally relevant to *y* (*epiphenomenalism*).

This is bad enough—as Malcolm says in "The Conceivability of Mechanism" (1982), it means that no one ever speaks or acts—but a simple extension of the argument promises to deprive mental phenomena of all causal influence whatsoever. Every event *z* of whatever type is metaphysically necessitated by some underlying physical event *y*, whose causally sufficient physical antecedents are presumably sufficient for *z* as well. But then by the exclusion principle, *z*'s mental antecedents are irrelevant to its occurrence. So, mental phenomena are *absolutely* causally inert. And now it is not only speech and action that are chimerical but also thinking.

Note ~~well~~ that the exclusion argument raises *two* problems for mental causation, one about mental particulars (events), the other about mental properties.[8] Strangely, philosophers have tended to treat these problems in isolation and to favor different strategies of solution.[9] In Malcolm's original presentation, he emphasizes problem one. Given a neurophysiological theory rich enough to

> provide sufficient causal conditions for every human movement, . . . there would be no cases at all in which [the] movement would not have occurred if the person had not had [the] desire or intention . . . [thus] desires and intentions would not be causes of human movements.[10]

Here the mystery is how mental *events*, desires for example, can be making a causal difference when their unsupplemented neurophysiological underpinnings are already sufficient to the task at hand. To reply with the majority that mental events just *are* certain physical events, whose causal powers they therefore share,[11] only relocates the problem from the particulars to their universal features:

> the being of a desire by my desire has no causal relevance to my extending my hand . . . if the event that is in fact my desire had not been my desire but had remained a neurological event of a certain sort, then it would have caused my extending my hand just the same.[12]

nothing essential is lost if '*x* is causally sufficient for *y*' is replaced throughout by '*x* determines *y*'s objective probability'. So unless the argument can be faulted on other grounds, mental causation is problematic under indeterminism too.

[8] C. D. Broad (1975) was perhaps the first to emphasize epiphenomenalism's double-sidedness: "[it] asserts . . . that mental events either (a) do not function at all as cause-factors; or (b) that, if they do, they do so in virtue of their physiological characteristics and not in virtue of their mental characteristics" (p. 473).

[9] Kim (1984*b*) is an important exception.

[10] Malcolm (1982), 136.

[11] See Feigl (1970), 36 ff.; Smart (1970), 54, 65–6; and Davidson (1980). Note that Davidson advances the token identity theory in response to a slightly different problem. His aim is to reconcile the following assumptions: singular causal claims need always to be backed by strict causal laws; strict laws are physical laws; every event subsumable under a physical law is a physical event; and mental events are efficacious.

[12] Sosa (1984), 278.

Mental events are effective, maybe, but not by way of their mental *properties*; any causal role that the latter might have hoped to play is occupied already by their physical rivals.[13] Although someone *could*, following the line above, attempt to *identify* mental properties with (certain) physical properties, say, being a desire with instantiating such-and-such a neurophysiological type, this approach is now discredited, because of the well-known multiple realizability objection.[14] Properties are identical only if each necessitates the other; but any physical property specific enough to necessitate a mental property is inevitably *so* specific that the converse necessitation fails. Since (as I'll maintain) the objection applies, *mutatis mutandis*, to mental *particulars*, the identity response is unworkable in either case.[15]

So I find no fault with dualism, or with the associated picture of mental phenomena as necessitated by physical phenomena which they are possible without. Rather than objecting, in fact, to the asymmetric necessitation picture, I propose to go it one better. Traditionally, the paradigm of one-way necessitation was the relation of *determinate* to *determinable* (sections 2 and 5). What if mental phenomena are determinables of physical phenomena in something like the traditional sense (sections 3 and 6)? Then since a determinate cannot preempt its own determinable, mental events and properties lose nothing in causal relevance to their physical bases (sections 4 and 7).[16] If anything, it is the other way around. Overladen as they frequently are with physical details far beyond the effect's causal requirements, it is the *physical* phenomena which are liable to disqualification on grounds of superfluity (section 8).

## 2.

Before asking what determinates and determinables might be, consider the "easier" question of when properties are identical. Probably no one would quarrel with

(**I**) *P* is identical to *Q* iff: for a thing to be *P* is for it to be *Q*,

[13] Again, this needs to be distinguished from a somewhat different worry directed primarily at Davidson's anomalous monism: singular causal claims need always to be backed by strict causal laws; *x*'s causally relevant properties *vis-à-vis y* are those figuring in the antecedent of some such backing law; strict causal laws never involve mental properties; so *x*'s mental properties are causally irrelevant. For discussion, see Stoutland (1980); Honderich (1982); Sosa (1984); Loewer and Lepore (1987); Fodor (1989); Loewer and Lepore (1989); Macdonald and Macdonald (1986); and McLaughlin (1989) (some of these papers discuss the exclusion objection also). Note that the exclusion objection, the subject of the present paper, assumes nothing about the role of laws in causation or in the characterization of causally relevant properties.

[14] See, for example, Putnam (1980) and Block and Fodor (1980).

[15] This is hardly a cause for regret. Identifying mental phenomena with physical phenomena, we saddle the former with the causal properties of the latter; but common sense sees mental phenomena as possessed of *distinctive* causal properties (see sections 8 and 9).

[16] About mental and physical *properties*, the Macdonalds (1986) reach a similar conclusion; however, their argument depends on treating mental *events* as identical to, rather than determinables of, physical events (see note 32 for the problems this causes).

on at least some interpretation. But, apart from its possible circularity, (**I**) explains one obscurity with another; and it has become customary to seek relief from both complaints in the modal idiom. That idiom permits no sufficient condition for property identity, unfortunately; so something is sacrificed. But we're repaid with the necessary condition that

(I) $P = Q$ only if: necessarily, for all $x$, $x$ has $P$ iff $x$ has $Q$.[17]

Properties are identical, in other words, only if it is impossible for a thing to possess either without possessing the other.

Among (I)'s attractions is that we *know* it is true since it follows from Leibniz's Law, the indiscernibility of identicals. Or better: it follows if the modality is read as *metaphysical*. Whether because they conflated conceptual with metaphysical necessity, or because they construed the properties themselves as concepts, philosophers *used* to think that properties were the same only if it was *conceptually* or *a priori*[18] true that their instances could not differ.[19] (Thus they felt justified in arguing from purely conceptual considerations to a distinction between, say, being salt and being sodium chloride.) This stronger condition can of course claim no support from Leibniz's Law.[20] But that isn't what led to its rejection: it was rejected because it proved unable to cope with the discovery of identical properties, such as the ones just mentioned, whose necessary coextensiveness was knowable only *a posteriori*.[21] So the mutual conceptual necessitation requirement is now defunct; its metaphysical kernel (I), although insufficient for property identity, is the only game in town.

According to a still reputable traditional doctrine, some properties stand to others as *determinate* to *determinable*—for example, *crimson* is a determinate of the determinable *red*, *red* is a determinate of *colored*, and so on.[22] Since the distinction is relative, one does better to speak of a determination *relation* among properties, where

(**Δ**) $P$ determines $Q$ iff: for a thing to be $P$ is for it to be $Q$, not *simpliciter*, but in a specific way.

[17] Treating necessary coextensiveness as also *sufficient* for property identity would lead to various unwanted results, for instance, that there is only one universally necessary property.

[18] I lump these two together not out of conviction but just as an expedient.

[19] This, the condition $(\mathrm{I}_1)$ that properties are identical only if their *necessary* coextensiveness is conceptually guaranteed, entails (I) trivially; (I) does not entail $(\mathrm{I}_1)$ conversely because some necessary coextensiveness claims are not *a priori* knowable, for example, that necessarily, the extension of identity-with-Hesperus is the same as that of identity-with-Phosphorus. Note the contrast between $(\mathrm{I}_1)$ and the weaker condition $(\mathrm{I}_2)$ that $P = Q$ only if their *actual* coextensiveness is knowable *a priori*. $(\mathrm{I}_1)$ and $(\mathrm{I}_2)$ fail for essentially similar reasons (see note 21), but it is $(\mathrm{I}_1)$ that I have in mind in the text.

[20] Reason: 'it is *a priori* that . . . ', like 'Jones believes that . . . ', generates an opaque context.

[21] Kripke (1980). Likewise, the weaker condition $(\mathrm{I}_2)$ cited in note 19 was overturned by the discovery of identical properties whose *actual* coextensiveness was not knowable *a priori* (e.g., identity-with-Hesperus and identity-with-Phosphorus).

[22] Two classic discussions are Johnson (1964), i. ch. 11, and Prior (1949).

Except for the 'not *simpliciter* . . . ', (**Δ**) would describe identity; and like identity, determination as traditionally understood involves conceptual and metaphysical elements jumbled confusingly together. Metaphysically, the central idea is that

(Δ) $P$ determines $Q(P > Q)$ only if:

(i) necessarily, for all $x$, if $x$ has $P$ then $x$ has $Q$; and

(ii) possibly, for some $x$, $x$ has $Q$ but lacks $P$.

Not always distinguished from this is a requirement of asymmetric conceptual entailment: there is no conceptual difficulty about a world in which some $Q$ lacks $P$, but the converse scenario is excludable on *a priori* grounds.

Now, just as the discovery of *a posteriori* necessities upset the traditional presumption of a conceptual equivalence condition on property *identity*, it also makes trouble for the conceptual entailment condition on *determination*. Take the property of being at temperature 95°C, and some highly specific micromechanical property $K$ chosen so that necessarily whatever has $K$ has the temperature property, though not conversely. Since $K$s which are warmer than 95°C cannot be ruled out on *a priori* grounds alone, traditional determination fails. Yet the relevance of this to the properties' strictly *metaphysical* relations is obscure; and since it is only the metaphysics that matters to causation, we should discount the traditional doctrine's conceptual component and reconceive determination in wholly metaphysical terms.[23] What justifies the continued use of the word 'determine' is that (**Δ**) holds essentially as before. To be in the micromechanical condition of this steaming tea, for instance, is to be at temperature 95°C *in a certain micromechanical way*.

## 3.

As I write, I am in a certain overall physical condition, and I am also thinking; presumably the one fact about me has quite a lot to do with the other. Suppose the pertinent aspects of my physical condition to be encoded in some physical property $P$. Could it be that $P$ is a *determinate* of thinking? Barring some unsuspected conceptual entailment from physics to thought, the full-scale traditional doctrine answers in the negative. On the other hand, traditional determination incorporates elements visibly irrelevant to how the properties

[23] So $P$ determines $Q$ just in case the traditional relation's first, metaphysical component is in place, where this consists primarily in the fact that $P$ necessitates $Q$ asymmetrically. Probably it goes too far to identify determination with asymmetric necessitation outright; ~~otherwise, for example,~~ we don't want conjunctive properties to determine their conjuncts and universally impossible properties ~~are~~ to be all-determining. For dialectical reasons, I try to remain as neutral as I can about where determination leaves off and "mere" asymmetric necessitation begins (Prior (1949) reviews some of the fascinating history of this problem).

are related in themselves; so the interesting question is whether *P* determines thinking in the *metaphysical* sense.[24] I say that it does. And I hold further that there is this sort of physical determination whenever a mental property is exemplified.

Such a view is in fact implicit in the reigning orthodoxy about mind–body relations: namely, that the mental is *supervenient* on, but *multiply realizable* in, the physical.[25] Because neither thesis concerns determination directly, the point is easily missed that in combination their effect is to portray mental properties as determinables of their physical realizations. Take supervenience first, the claim that a thing's mental properties are fixed by how it is physically:

(S) Necessarily, for every $x$ and every mental property $M$ of $x$, $x$ has some physical property $P$ such that necessarily all $P$s are $M$s.[26]

Now, thinking is a mental property, and I possess it. By supervenience, then, I have a physical property *P* given which thinking is metaphysically guaranteed. Of course, *P* can be considered a determination of thinking only if it is possible to think *without P*, which is to say otherwise than by way of the physical property

[24] "But if there is no conceptual entailment from *P* to thinking, then unthinking *P*s are conceivable, and to that extent possible; thus *P* doesn't determine thinking in the metaphysical sense either." I grant that the conceivability of a proposition $\phi$ is *prima facie* evidence of its possibility. But this *prima facie* evidence is defeated if there is not improbably a proposition $\psi$ such that (a) $\phi$ is true, (b) if $\psi$ is true, then $\phi$ is impossible, and (c) $\phi$ is conceivable only because one was unaware of (a) and/or (b). The ancients, for instance, were able to conceive Hesperus as existing without Phosphorus only because they were unaware of their identity; and if I find it conceivable that something should be in the micromechanical condition of this steaming tea but with a different temperature, that is for ignorance of the temperature's microphysical explanation. But I take it that there may also be an explanation of how thinking arises out of neurophysiology, such that if I knew it, then I would find it *in*conceivable, and consider it impossible, that something should be *P* without thinking. What's more, the prospect of such an explanation makes the hypothesis of an unthinking *P* only dubiously conceivable *today*. So the complaint is questionable on two counts. First, from a proposition's conceptual coherence, from the fact that its denial is not conceptually false, its conceivability does not follow—witness the Hesperus/Phosphorus example. Even where conceptual difficulties are absent, conceivability can be inhibited by the knowledge or suspicion of a defeater; and this is how it is, for many of us, with the proposition that there could be *P*s that did not think. Second, any conceivability intuition I *might* muster in this area I regard as unreliable, because liable to defeat by the progress of science. (For the (a), (b), (c) model of modal error, see Yablo (1990) and (1993)●; Chs. 1 and 2 above.) ?

[25] "All but explicit" would not be much of an exaggeration; determination lies so near the surface and so neatly organizes received opinion that one wonders why it is not already a standard theme.

[26] This is Kim's "strong supervenience" (1984*a*). Perhaps not everyone accepts supervenience in quite this strong a form; perhaps I don't myself (Yablo, 1990). Yet for two reasons I have thought it better to formulate the thesis as in the text: (i) strong supervenience is seen nowadays not as the *answer* to epiphenomenalism but rather as the context in which the problem as currently discussed arises (avoiding epiphenomenalism may indeed have been part of the original impulse behind (S), but that is what makes its reappearance *under* (S) all the more troubling); (ii) it focuses the essential line of thought to work within relatively strong assumptions. How much supervenience the approach really needs, and whether that much is plausible, are questions for another paper. For now I just state my hope of getting by with a form of supervenience that allows for the possibility of nonphysical thinkers (see note 47).

that *does* realize my thinking; and this is where the official story's second element comes in.

When philosophers abandoned the hope of finding for every mental property an identical physical property, the reason was that mental properties seemed intuitively to be multiply realizable in the physical.[27] However, some care should be taken about what this means. Is the claim that for *any* pair of properties, one mental and the other physical, something could have the first without the second? Really, this is stronger than intended, or needed. Imagine someone who holds that necessarily every thinker is spatially extended. Surely such a person could accept multiple realization, intuitively understood, without falling into inconsistency; yet since the necessitation of extension by thinking is the necessitation of a physical property by a mental one, her view actually runs contrary to multiple realization as just explained. Provided that they are suitably unspecific, then, physical properties *can* be necessitated by mental properties compatibly with multiple realization—which suggests as the thesis's proper formulation that $M$ necessitates no physical $P$ that is *specific enough to necessitate M in return*:

(M) Necessarily, for every mental property $M$, and every physical property $P$ which necessitates $M$, possibly something possesses $M$ but not $P$.[28]

For purposes of refuting the identity theory, note, (M) is all that's required. If $M$ were $P$, then $P$ would necessitate it. But then by (M), it could not necessitate $P$ in return, contrary to their assumed identity.

Together, (M) and (S) make it a matter of necessity that something has a mental property iff it has a physical property by which that mental property is asymmetrically necessitated. But this is extremely suggestive, for with 'determines' substituted for 'asymmetrically necessitates', it becomes

(D) Necessarily, something has a mental property iff it has also a physical determination of that mental property;

and (D) is an instance of the standard equation for determinables and determinates generally: namely, that something has a determinable property iff it has some determinate falling thereunder. This calls out for explanation, and the

[27] See Putnam (1980) and Block and Fodor (1980).

[28] "Now you contradict yourself, for (M) is incompatible with supervenience. Let $\vee P_i$ be the disjunction of all $M$-necessitating physical properties (alternatively, the second-order property of possessing some $P_i$ or other); then (S) entails that $M$ and $\vee P_i$ necessitate each other, contrary to (M)'s claim that physical properties necessitate mental properties only asymmetrically." To respond by denying the reality of disjunctive properties, on the principle that co-possessors of *real* properties are thereby similar, forgets that the $\vee P_i$s *are* similar in that they have $M$ in common. However, a related point still holds good: sharing of *physical* properties should make for *physical* similarity, and unless the multiple realizability thesis can be faulted on other grounds, the $\vee P_i$s are only mentally alike. (The tendency to think of the physical properties as closed under disjunction may owe something to a confusion of wide- and narrow-scope readings of '$x$ exemplifies a $P_i$'. What is true is that for each $P_i$, whether $x$ possesses *it* is a physical question; this does not make it a physical question whether $x$ has some $P_i$ or other.)

one that comes first to mind is that mental/physical relations are a species of determinable/determinate relations. "Can you really be saying that mental properties stand to their physical realizations in the relation that rectangularity bears to squareness, or that colors bear to their shades?"[29] Yes. At least that is my conjecture, to be evaluated like any other by the evidence for it and by its theoretical fruitfulness. The evidence is as just described; its consequences for mental causation are considered next.

## 4.

Imagine a pigeon, Sophie, conditioned to peck at red to the exclusion of other colors; a red triangle is presented, and Sophie pecks. Most people would say that the redness was causally relevant to her pecking, even that this was a paradigm case of causal relevance. But wait! I forgot to mention that the triangle in question was a specific shade of red: scarlet. Assuming that the scarlet was causally sufficient for the pecking, we can conclude by the exclusion principle that every *other* property was irrelevant. Apparently, then, the redness, although it looked to be *precisely* what Sophie was responding to, makes in reality no causal contribution whatever. Another example concerns properties of events. Suppose that the structures in a certain region, though built to withstand lesser earthquakes, are in the event of a *violent* earthquake—one registering over five on the Richter scale—causally guaranteed to fall. When one unexpectedly hits, and the buildings collapse, one property of the earthquake that seems relevant to their doing so is that it was violent. Or so you might think, until I add that this particular earthquake was *barely* violent (its Richter magnitude was over five but less than six). What with the earthquake's *bare* violence being

[29] "There is a crucial difference: My mental properties *result* from my physical condition, but in no sense does a thing's redness result from its being scarlet." Actually this raises a subtle interpretive question about supervenience. On the *emergence* interpretation, a thing's physical properties are metaphysically prior to its mental properties and bring them into being. To caricature emergentism just slightly, supervenience is a kind of "supercausation" which improves on the original in that supercauses act *immediately* and metaphysically *guarantee* their supereffects (the supervenience/causation analogy is common; see, e.g., Kim (1984*a*)). Another view is that the supervening mental properties are *immanent* in their physical bases; rather than giving rise to thought by some obscure metaphysical motion, certain material conditions are inherently conditions of thinking. Now, as the objector suggests, immanentism is clearly correct in standard cases of conceptual entailment, for example, scarlet and red, squareness and rectangularity. Surely, though, this ought to make us suspicious about emergentism as an interpretation of the other cases—for how can the properties' conceptual relations bear on the metaphysical character of the supervenience? That the emergentist thinks they do hints at an unconscious appeal to the neo-Humean prejudice that regularities divide into the conceptual and the causal, or causal-like. But the dilemma is unreal: 'whatever is in the micromechanical condition of this tea is at temperature 95°C' fits into neither category, and I see no reason to treat 'whatever is in the physical condition of this person is thinking' differently. On the immanence model, ~~of course,~~ the alleged disanalogy with colors and their shades evaporates.

*already* causally sufficient for the effect, that it was *violent* made no causal difference.

Surprising results! To the untrained eye, the redness and the violence are *paradigm cases* of causal relevance, but only a little philosophy is needed to set matters straight. Now, though, one begins to wonder: if even paradigm cases of causal relevance fail the exclusion test, what passes it? Not much, it turns out. Almost whenever a property $Q$ is *prima facie* relevant to an effect, a causally sufficient determination $Q'$ of $Q$ can be found to expose it as irrelevant after all.[30] Applying the argument to $Q'$, $Q''$, etc. in turn, it appears that only ultimate determinates—properties unamenable to further determination—can hope to retain their causal standing.

Or, on second thought, maybe not them either. Not everything about a cause contributes to its effect; and even where a property does contribute, it need not do so in all its aspects. From the examples it is clear that such irrelevancies do indeed creep in, as we pass from determinable to determinate (e.g., registering less than six); and if the determination process is continued *ad finem*, they may be expected to accumulate significantly. So any ultimate determinate seems likely to incorporate causally extraneous detail. But then, abstracting some or all of this detail away should leave a determinable which, since it falls short of the original only in irrelevant respects, is no less sufficient for the effect.[31] By the exclusion principle, this robs even ultimate determinates of their causal powers. And now it begins to look as though no property ever makes any causal difference.

At least as it applies to properties, then, the exclusion principle is badly overdrawn. Not that there is nothing right about it. In *some* sense of 'separate', it stands to reason, separate properties *are* causal rivals as the principle says. Then what if someone identifies the appropriate notion of separateness and reformulates the exclusion principle accordingly? Suppose it done. Even without hearing the details, we *know* that the corrected principle does not apply to determinates and their determinables—for we know that they are not causal rivals. This kind of position is ~~of course~~ familiar from other contexts. Take for instance ~~example~~ the claim that a space completely filled by one object can contain no

[30] Depending on what exactly the exclusion principle demands in the way of causal sufficiency, $Q'$ might be a determination of $Q$ only in a fairly relaxed sense (see notes 5 and 23). Those uncomfortable about this should remember the dialectical context: we are trying to show that the assumption needed to disempower mental properties—namely, that determinates are causally competitive with their determinables—would, if true, disempower virtually *all* properties. But if they are causally competitive on a *strict* reading of the determination relation, then when it is *loosely* construed they should be competitive also; and the argument in the text, with determination read the second way, shows that this results in a basically unmeetable standard of causal relevance.

[31] Although it contributed nothing to the earthquake's destructiveness that it registered under Richter six, a determinate of its violence that omitted this would *ipso facto* not be ultimate. Hence the ultimate determinate, whatever exactly it may be, sets a causally idle upper bound on the earthquake's violence; abstracting this upper bound away, we arrive at a determinable still sufficient for the buildings' collapse. (Again, in some cases, this might be a determinable of the ultimate determinate only in a fairly relaxed sense—but see the previous note.)

other‸object. Then are even the object's *parts* crowded out? No. In this competition wholes and parts are not on opposing teams; hence any principle that puts them there needs rethinking. Likewise any credible reconstruction of the exclusion principle must respect the truism that determinates do not contend with their determinables for causal influence.[32]

With the exclusion principle neutralized, the application to mental causation is anticlimactic. As a rule, determinates are tolerant, indeed supportive, of the causal aspirations of their determinables. Why should it be different, if the determinate is physical and the determinable mental? Inferring the causal irrelevance of, say, my *dizziness*, from the causal sufficiency of its physical basis, is not appreciably better than rejecting the redness as irrelevant on the ground that all the causal work is accomplished already by its determinate scarlet. Or, if someone thinks it *is* better, then she owes us an explanation of what the metaphysically important difference is between the cases. That there is a conceptual difference is granted, but it is not to the point; there is no conceptual entailment either from the tea's micromechanical condition to its high temperature, yet this occasions little skepticism about the role of the tea's temperature in its burning my tongue. If there is a metaphysical difference, then someone should say what it is, and why it matters to causation.

## 5.

According to our guiding principle (**Δ**) for property determination, *P* determines *Q* iff to possess the one is to possess the other, not *simpliciter*, but in a certain way. But this way of putting things comes naturally, too, in connection with particulars, and especially events. If *p* is the bolt's *suddenly* snapping, for example, and *q* is its snapping *per se*, then for *p* to occur is for *q* to occur in a certain

[32] This is the Macdonalds' view also, but I question their rationale. Sometimes they seem to be arguing as follows: properties derive their causal powers from their instances; if one property determines another, an instance of the first is an instance of the second; so whenever a determinate is efficacious, its determinables are too. However, the conclusion is much too strong. Imagine a glass which shatters if Ella sings at 70 decibels or more. Tonight, as it happens, she sang at 80 db, with predictable results. Although it was relevant to the glass's shattering that the volume was *80* db, it contributed nothing that it was *under 90* db. Therefore, an efficacious determinate can have an irrelevant determinable. Another reading of the Macdonalds' position might be that the determinate's instances are instances of the determinable only *sometimes*, and that it is only in *these* cases that the determinable is efficacious if the determinate is. But notice what this requires: Ella's singing at 80 db is *identical* to her singing at over 70 db, but *distinct* from her singing at under 90 db. Apart from its intrinsic implausibility, such a view is untenable for logical reasons. *P* and its determinable *Q* are efficacious not absolutely, but only relative to some specified effect; whether their instantiations are identical, though, has to be decided once and for all. So the strategy of identifying the *P*- and *Q*-events iff both *P* and *Q* are efficacious leads to inconsistent results: they *can't* be the same event, because there are effects (the glass's shattering) to which only *P* is relevant; at the same time they *must* be, to accommodate effects (the neighbor's turning up her hearing aid) to which *Q* is relevant too.

this is correct, but most of the later lower-case greek letters should be plain

way, namely suddenly; and my *slamming* the door consists in my shutting it, FN:33 not *simpliciter*, but with significant force.[33] This suggests the possibility of a determination relation for events:

bf **(δ)** *p* determines *q* iff: for *p* to occur (in a possible world) is for *q* to occur (there), FN:34 not *simpliciter*, but in a certain way.[34]

If the relation can be made out, then in addition to the examples mentioned, Icarus's flying too near the Sun determines his flying *per se*, Brutus's killing FN:35 Caesar determines his stabbing Caesar,[35] Gödel's discovering the incompleteness of arithmetic determines his realizing that arithmetic was incomplete, and so on indefinitely.

There is a complication. Determination involves the idea that the requirements associated with one thing include the requirements associated with another; and although properties are requiremental on their face, particulars are not. Hence the need for a notion of individual essence.

By a thing's *essential* properties, I mean those it cannot exist without. And its *essence* is a certain selection of its essential properties. But which essential properties does it make sense to include? The simplest proposal, obviously, would be to include *all* of them. For two related reasons, though, that won't do. Naively, the "what-it-is" of a thing—its identity and kind—should be *in virtue of* its essence. Yet if identity- and kind-properties are allowed into essences, this requirement becomes quickly trivialized: a thing does not get to be identical to Brutus's stabbing Caesar, or of the kind *stabbing*, by having the property of so being, but by having certain *other* properties and by their dividing along appropriate lines between essential and accidental. Second, the essence of a thing is supposed to be a measure of what is *required* in order to be that thing. Thus if more is required to be *y* than to be *x*, this should be reflected in an inclusion relation between their essences. The problem is that identity-properties, kind-properties, and the like are liable to disrupt these inclusion relations. Allowing *identity-with-x* into *x*'s essence precludes the possibility of a *y* whose essence includes everything in *x*'s essence, and more besides; and the effect of allowing *x*'s kind into its essence is to kill the chances for a thing *y* whose essence exceeds FN:36 *x*'s by properties which things of that kind possess at best accidentally.[36]

Both problems have the same solution: essences are to be drawn from a pool of properties such that any particular such property's modal status—essential

[33] Here and throughout 'events' are event tokens, not types; my slamming the door is something that happens at a specific time, in a specific place, and in a particular way.

[34] Where this is understood fairly generally, so that, for example, Poindexter's lying to Congress is his speaking to Congress in a certain way, ~~to wit~~ viz. falsely.

[35] Killings need not be stabbings, and Brutus could have killed Caesar without stabbing him; but this *particular* killing, I assume, could not have occurred except by way of the associated stabbing (this is important if the killing is to be a determination of the stabbing).

[36] For example, to *stabbings*, unlike *killings*, it is not essential that someone die.

or accidental—is without undue prejudice to the modal status of the others. Dubbing these the *cumulative* properties, *x*'s *essence* will be the set of cumulative properties that it possesses essentially. When *q*'s essence is a subset of *p*'s essence, *p* is said to subsume $q(p \geqslant q)$; and *p determines* $q(p > q)$ when the inclusion is FN:37 strict.[37]

Explaining determination by essence has three points in its favor: it fits the intuitive examples; it supports the analogy with property determination; and it predicts the principle that *p* determines *q* only if for *p* to occur is for *q* to occur in a certain way. Take the example of Gödel's *discovering*, versus his simply *realizing*, that arithmetic was incomplete. Though identical on some accounts, there is in fact a subtle difference between them. Speaking first of Gödel's *realizing* that arithmetic was incomplete, this *could* have been the realization of a result already widely known (in that case, it would not have made Gödel famous). To Gödel's *discovering* arithmetic's incompleteness, though, some degree of priority is essential. Otherwise one could ask, would it still have made Gödel famous, if incompleteness had been common knowledge? But this is like asking, of Brutus's killing Caesar, what Caesar would have done to Brutus if he had not died of it. So the essence of Gödel's discovering that arithmetic was incomplete *adds* something to the essence of his realizing that it was.

For the analogy with property determination, we need a distinction: a property is *categorical* if its possession by a thing *x* at a possible world is strictly a matter of *x*'s condition in that world, without regard to how it would or could have been; other properties, such as ~~for example~~ counterfactual and modal properties, are FN:38 *hypothetical*.[38] This gives the idea of categoricity, but as a definition it would be circular. To see why, suppose it is a categorical property of this piece of wax to be spherical. How can this depend on the wax's condition in other worlds? In a

[37] Here is the basic condition on cumulative properties stated more formally: ($\kappa$) for all *x*, for all possible worlds *w*, for all sets *S* of cumulative properties [*x* exists in *w* and possesses there every member of $S \leftrightarrow$ there exists in *w* an $x^+ \geqslant x$ to which every member of *S* belongs essentially]. To see how this works to exclude identity properties, suppose that *x* possesses some cumulative *P* accidentally in some world *w* where it exists. If *identity-with-x* were cumulative, by ($\kappa$) there would be an $x^+$ in *w* to which *identity-with-x* and *P* were both essential—a contradiction, since nothing can be both identical to *x* and essentially possessed of a property which *x* possesses only accidentally. Likewise for kind-properties: if *x* is accidentally *P* and of such-and-such a kind, it will normally be impossible to strengthen *x* into an $x^+$ still of that kind but possessing *P* essentially. Thus, no *person* is *essentially* born on a certain day, no *stabbing* is *essentially* fatal, no *landslide* is *essentially* between nine and ten seconds long, and so on. (Terminological note: subsumption is called 'refinement' in ~~"Identity, Essence, and Indiscernibility", *Journal of Philosophy* 84 [1987]: 293–314~~ Yablo (1987), and 'strengthening' in "~~Cause and Essence," *Synthese* [1992] 403–449).~~ Yablo (1992)

[38] More familiar are the notions of an *occurrent* property: one whose possession by a thing at a time is insensitive to how matters stand at other times; and an *intrinsic* property: one which a thing possesses wholly in virtue of how it is in itself, irrespective of what goes on around it. Within limits we can think of categoricity as standing to the modal dimension as occurrence stands to time and intrinsicness to space (~~see "Identity, Essence, and Indiscernibility,"~~ Yablo (1987) and Yablo (1999) ~~Yablo, S. (1999) "Intrinsicness, Occurrent, Categorical," manuscript *Philosophical Topics* xxvi 1/2: 479–505~~).

way, though, it does, for the wax cannot be spherical in this world without being possibly spherical in every other world it inhabits. More generally, sensitivity to its possessors' *hypothetical* characteristics in other worlds should not make a property noncategorical, or *no* properties will be categorical. What we *meant* to say, it seems, is that a property is categorical iff it attaches to its objects regardless of how they would or could have been in *categorical* respects. And now the circularity is apparent.

Luckily the categorical properties can be approached from another direction. When *p subsumes q*, their difference (if any) comes down ultimately to the fact that they possess different of their shared properties essentially. Such a difference is *merely* hypothetical if any difference is; so

not bf, make plain

**(γ)** $C$ is categorical only if: necessarily, for all $p$ and $q$ such that $p \geqslant q, p$ has $C$ iff $q$ does.

This, although only a necessary condition on categoricity, is all that the announced analogy requires.[39] For it entails that in worlds where both exist, the subsuming particular $p$ and the subsumed $q$ are categorically indiscernible, or as I will say *coincident*. And since $p$ cannot exist *without* $q$[40] (the bolt's suddenly snapping is impossible without its snapping) we have:

FN:39

FN:40

not bf, make plain

**(μ)** $p \geqslant q$ only if: necessarily, if $p$ exists, then $q$ exists and is coincident with $p$.

[39] Assuming that the logical space of particulars is *full* in a sense I discuss elsewhere, the stated condition is sufficient also (Yablo (1987) ~~"Identity, Essence, and Indiscernibility,"~~ secs. 4 and 5). Fullness is a sort of plenitude principle whose point is to ensure that there are particulars enough to witness the hypotheticality of every hypothetical property; that is, that for each hypothetical $H$, there exist in some possible world $\geqslant$-related $p$ and $q$ such that $H$ attaches to exactly one of them. To illustrate, part of the assumption is that for any particular $q$ and any non-empty set $W$ of worlds in which it exists, there is a $p \geqslant q$ which exists in the $W$-worlds exactly. Now suppose we agree that to be, say, flexible, a thing must be at least capable of flexing, that is, it needs to flex in at least some worlds. By fullness, any flexible $q$, provided only that there are worlds in which it never flexes, will have a determination $p$ which metaphysically *cannot* flex. This shows that flexibility is hypothetical. (Some say that if dispositional properties are hypothetical, then *all* properties are, for it is essential to every property, however categorical it might otherwise seem, to confer on its possessors correlative causal dispositions, for instance, flexibility, corrosiveness, visibility. But the idea that even seemingly categorical properties are essentially disposition-conferring is, in the context of the fullness assumption, quite implausible. For instance, it detracts not at all from a thing's actual-world *roundness* to restrict or otherwise adjust its counterfactual career, but its dispositions can be varied almost at will by the same operation. What *might* be essential to roundness is to confer appropriate dispositions on particulars meeting *further hypothetical conditions*, conditions aimed at ruling out unusual hypothetical coloration such as we saw above. Yet since roundness "entails" these dispositions only over its hypothetically *ordinary* possessors, the objection is analogous to the following: no *ordinary* thing moves discontinuously; so being at such-and-such a location at a given time "entails" the non-occurrent property of not being at every *other* time a million miles away; so, location properties are not occurrent!)

[40] That is, if $p \geqslant q$, then necessarily if $p$ exists then so does $q$. *Proof*: Run ($\kappa$) from right to left with $S =$ the empty set. (Another proof uses the assumption that $x$ exists in $w$ iff $x$'s essence is satisfied there, that is, something possesses there all its member properties: $w$ contains $p$ only if $p$'s essence is satisfied in $w$ only if $q$'s smaller essence is satisfied in $w$ only if $w$ contains $q$.)

This divides into two subconditions, according to whether $p$ is identical to $q$ or determines it.[41] By Leibniz's Law, or a double application of ($\mu$), plain, good

not bf, make plain FN:41

(ι) $p = q$ only if: necessarily, $p$ exists iff $q$ exists, and if existent, they are coincident.

When $p$ determines $q$, the condition holds in one direction only:

plain, good

(δ) $p > q$ only if:

(i) necessarily, if $p$ exists, then $q$ exists and is coincident with $p$;

(ii) possibly, $q$ exists and $p$ does not exist.[42]

FN:42

That we get these analogues for particulars of (**I**) and (**Δ**) is the second attraction of using essence to explain determination.

Now for the fact that reflects most favorably on the essence approach: that it predicts (**δ**)'s intuitive description of determination. From (δ) we know that a determinate $p$ exists in some, though not all, of the worlds where its determinable $q$ is found. But how does $p$ decide in *which* of these $q$-worlds to put in its appearances? For instance, what separates the worlds in which the bolt's suddenly snapping accompanies its snapping *per se* from those in which it does not? In the former worlds, presumably, the snapping is sudden; and as it turns out, this answer holds good in general:

plain, good

bf

not bf, make plain FN:43

(ϵ) $p > q$ only if: necessarily, $p$ exists iff $q$ (both) exists and exemplifies the difference $S$ between its own essence and $p$'s larger essence.[43]

*Mirabile dictu*, this is just what (**δ**) *says* about determinates and their determinables: for $p$ to occur is for $q$ to occur, not *simpliciter*, but $S$-ly.

bf

## 6.

Identicals are indiscernible; so an argument that mental events have different essential properties from physical events is an argument that they are not

[41] That these exhaust the possibilities is not trivial; but it can be proven from ($\mu$) and the assumption ($\sigma$) that distinct particulars either exist in different worlds or are noncoincident in some world where they exist together. *Proof*: It suffices to show that $p = q$ if they have the same essence. Suppose they do. Then each subsumes the other. By ($\mu$), they exist in the same worlds and are coincident in all of them. By ($\sigma$), $p = q$.

[42] *Proof*: (i) is immediate from ($\mu$) and the fact that determination entails subsumption. (ii) If $p$ existed in every world in which $q$ did, then by ($\mu$) and (i) they would exist in the same worlds and be coincident in all of them. Given ($\sigma$) that would make them identical, contrary to the assumption that $p$ determines $q$.

[43] *Proof*: Suppose that $q$ exists in a world $w$ and exemplifies $S$ there. By ($\kappa$)'s left-to-right direction, there exists in $w$ a $q^+ \geqslant p$; it follows from ($\mu$) that $p$ exists in $w$. For the converse, run ($\kappa$) from right to left with $S$ as before.

identical. According to one popular line of thought, this essential difference can be established in the following simple form: only mental events possess mental properties (e.g., phenomenal and content properties) essentially. Thus Kripke:

> Let '*s*' name a particular pain sensation, and let '*b*' name the corresponding brain state, or the brain state some identity theorist wishes to identify with *s*. *Prima facie*, it would seem that it is at least logically possible that *b* should have existed (Jones's brain could have been in exactly that state at the time in question) without Jones feeling any pain at all, and thus without the presence of *s*.[44]

*Prima facie*, Kripke says, *b* could have occurred without there being any pain, and presumably he would say the same about other physical events *p* and mental properties. Unless these *prima facie* appearances can be overcome, mental properties are at best accidental to physical events.

Are these really the *prima facie* appearances, though? Remember that all it takes for *p* to have a mental characteristic essentially is for its essential physical properties to necessitate one—and that the dominant modal intuition in recent years has been that mental properties *supervene* on physical properties and so are necessitated by them *all the time*.[45] Someone might of course ask why any physical *p* should have the mentally consequential *kind* of physical property, but this is easily explained. Consider the bearing of supervenience on *mental* events: for each of *m*'s mental properties, supervenience assigns it a necessitating physical property. But it is hard to think what *m*'s physical properties could be if not those of some physical event *p* which subserved it. Thus, among *p*'s physical properties are some with *m*'s mental properties as necessary consequences. Only if *p* somehow managed to have *all* of these physical properties contingently could it avoid having at least some mental properties essentially.

Instead of insisting that *p* has *no* essential mental properties, perhaps the token dualist should say that it doesn't have *all* the essential mental properties of its

[44] Kripke (1980), 146, with inessential relettering. Note that if "logically possible" is taken literally, as covering everything permitted by logic, then even identicals can differ in what is logically possible for them—for example, it is logically possible that Hesperus, but not that Phosphorus, should exist in Phosphorus's absence. Obviously this would make logical possibility useless in applications of Leibniz's Law; so I assume that Kripke is using "logical possibility" for metaphysical possibility.

[45] Two remarks. First, the point of calling this an *intuition* is that Kripke's argument might be read as objecting to supervenience itself (1980, 155). So read, the argument assumes that the weight of modal intuition favors the antisupervenience position. This I deny; there are many reasons for supervenience's popularity, but one, surely, is its enormous modal intuitiveness. (Such antisupervenience intuitions as may exist I would hope to explain away in the manner of note 24.) Second, someone might complain that "the dominant intuition" is only that the mental characteristics of *objects*, or perhaps *worlds*, are necessitated by their physical properties; *events* are another story. Otherwise supervenience entails, as it surely should not, that every mental event is a physical event. But the objection assumes that events with mental properties are thereby mental events; and I am in the process of questioning whether even *essential* mental properties are enough to make an event mental. ●(For the idea that strong supervenience presupposes token identity, see Haugeland (1982) and Kim (1988).)

alleged mental identical. Here is a bad way to argue for that result: since no mental event is physical, *p* lacks mental kind-properties, for example, being of the kind *after-image, sensation*, or indeed *mental*; therefore it doesn't have these properties essentially. Dialectically, of course, this begs the question against the token identity theory. But there is a deeper problem: it says nothing about what *makes* a mental event *m* different from a physical event *p*, to be told that only the former is (essentially) mental, or of some specific mental kind. Mental events are mental rather than physical not because mentality is essential to them alone, but because of some *prior* fact about them—the sort of fact that essences were designed to capture. Thus *m*'s essential mental advantage over *p*, if it exists, should be that its *essence* contains mental properties beyond those in *p*'s essence.

Yet supervenience opposes this weakening of the essential mental advantage view as much as the original. The reason is this. Every mental property $M_k$ in *m*'s essence is backed by a necessitating physical property $P_k$; and as before, these physical properties attach also to some realizing physical event (this time called *q*). Even if some or all of the $P_k$s are only accidental to *q*, we can imagine a more determinate physical event *p* to which they are all essential. But then *p* has essential physical properties to necessitate every mental property in *m*'s essence; and it follows that these mental properties are in *p*'s essence too. Not only does this rule out an essential mental advantage for mental events, it puts us in sight of an intriguing parallel between the ways that mental events and properties relate to their physical underpinnings. For assuming that *p* can be chosen determinate enough to essentially possess such few *non*mental properties as might be found in *m*'s essence, we have

(s) Whenever a mental event *m* occurs, there occurs also a subsuming physical event *p*, that is, a physical event whose essence includes *m*'s essence[46]

—an analogue for events of the supervenience thesis.

From (s) it is clear that if there is an essential difference between mental events and physical ones, it is *not* that physical events' essences are mentally impoverished. Instead, I suggest, it is the other way around: the essences of *mental* events are *physically* impoverished. For those who believe, with Descartes, that their mental lives could have proceeded just the same in a wholly immaterial world, this hardly requires argument.[47] Events which can occur in such a world presumably have *none* of their physical properties essentially. But Cartesian dualism is only the most dramatic expression of a thought which seems probable

[46] Notice what (s) doesn't say: that every property *essential* to *m* is essential to *p*. For all we know so far, no mental event is physical; in that case *m*'s mental identity- and kind-properties are not properties of *p* at all.

[47] Of course, whoever accepts supervenience in form (S) will find the Cartesian hypothesis hard to swallow, for (S) implies that in, and across, immaterial worlds, everyone is thinking exactly the same thing! This has led some authors (e.g., David Lewis and Frank Jackson) to seek more permissive interpretations of supervenience.

in any case: namely, that in comparison with their physical bases, mental
FN:48 phenomena are exceedingly modally elastic.[48]

Take for example the pain sensation *s*, and the underlying brain event *b* whose identity with *s* is in question; and grant the identity theorist that *b* at least subsumes *s* and so necessitates it. The problem is that as *b* takes on the degree of essential physical detail that this requires, it becomes intuitively irresistible that the pain is possible even in *b*'s absence. Something like this is Kripke's second argument against the identity theory:

> [*B*]*eing a brain state* is evidently an essential property of *b* (the brain state). Indeed, even more is true: not only being a brain state, but even being a brain state of a specific type is essential to *b*. The configuration of brain cells whose presence at a given time constitutes the presence of *b* at that time is essential to *b*, and in its absence *b* would not have existed. Thus someone who wishes to claim that the brain state and the pain are identical must argue that the pain could not have existed without a quite specific type of configuration
> FN:49 of molecules.[49]

*Prima facie*, it seems obvious that the pain could still have occurred, even if that specific arrangement of molecules hadn't, and as Kripke says, the *prima facie*
FN:50 appearances aren't easily defeated.[50] But if the molecular arrangement is essential

[48] This is a particular theme of Richard Boyd (1980).

[49] Kripke (1980), with inessential relettering.

[50] "Granted that *a* pain could still have occurred in the absence of that molecular configuration, what makes you think that it is the same pain that occurred actually?" Among the lessons of *Naming and Necessity* is that to find a thing *x* capable of existing in some counterfactual condition, one imagines this *directly*—as opposed to imagining something *y* in that condition whose transworld identity with *x* must then be established. This is crucial if imaginability is to be a source of knowledge about *de re* possibility. For (i) having imagined *y* in the indicated condition, verifying that *y* is *x* requires appeal to transworld identity criteria which, if they are available at all, are typically *more* controversial than the *de re* attributions they are called on to support; and (ii) without reliance on direct *de re* imagination there would be no way to justify these criteria in the first place. Stripped then of its reference to transworld identity, the question is, Is *m* really imaginable in the absence of *b*, or is the only imaginable scenario one in which a distinct if similar pain occurs in *b*'s absence? Here I can do no more than echo Kripke in claiming the former intuition. Such intuitions are of course defeasible by reference to unnoticed complications, but they are *prima facie* credible, and the burden of proof is on the critic (see Boyd (1980) for pertinent thought experiments, and note 24 for the defeasibility of modal intuition). ~~On~~ At a deeper level, perhaps the objection reflects not any particular attachment to a picture of mental events as bound to their physical underpinnings, but a more general malaise attending *all* modal thinking about events. Whereas objects fall into more or less settled kinds, which then guide us in our assessment of what counterfactual changes they will tolerate, with events our commonsense sortal apparatus is relatively primitive and modally inarticulate; that something is a pain, or an explosion, tells us enormously less about its possibilities than that it is a person or a ship. Hence our admitted squeamishness about events' potential for contrary-to-fact behavior—which hardens all too easily into the positive thesis that that potential is extremely limited (i.e., that events are inherently modally inflexible). This last, though, is surely an overreaction. What the squeamishness really signifies is the inadequacy of everyday event-sortals to the task of identifying just which of various coincident-but-hypothetically-different items one has in mind. Small wonder, then, if the identificatory task falls partly to the *de re* modal attributions themselves; and some of the more dogmatic-sounding attributions in the text may seem less so when understood in this spirit: as partial specifications of their subject matter rather than as attempts to describe an already singled-out particular.

to *b* alone, then *b*'s essence is physically richer than *s*'s essence. Therefore *b* subsumes *s properly*; and this, extended across mental events in general, gives an analogue for particulars of the multiple realizability thesis:

plain face (**m**) For every mental event *m*, and every physical event *p* which subsumes *m, p* subsumes *m* properly and so determines it.

Token dualism follows: if *m* were identical to *p*, then *p* would subsume *m*; hence by (m) it would determine *m*, contrary to their assumed identity.

Drawing these various threads together, we find that the relation between mental and physical events effectively duplicates that of mental to physical properties. Whenever a mental event *m* occurs, (s) guarantees a subsuming physical event *p*, which by (m) is not identical to *m* but determines it. Thus with FN:51 every mental *m* comes a determining physical *p*.[51] Since for *p* to occur is just for *m* to occur in a certain physical way, the converse is trivial; so we can say that

(d) A mental event *m* occurs iff some physical determination *p* of *m* occurs.

This is our analogue for events of the mental/physical determination thesis for properties.

## 7.

Haven't we now made mental events causally irrelevant? By the exclusion principle, *m* can influence an outcome only to the extent that *p* leaves that outcome causally undecided. Results which *p* causally guarantees, therefore, it

[51] This may seem doubtful, if one insists on seeing *p* as (i) a localized brain event, (ii) capable of occurring in isolation from anything like its actual neural context. Imagine a C-fiber stimulation, *b*, and a pain sensation, *s*, with the following properties. First, they are both occurring in me right now; second, *b could* have occurred in isolated C-fibers afloat in agar jelly; third, had *b* occurred in the latter environment, *s* would not have accompanied it. Then since determination entails necessitation, *b* does not determine *s*. The moral is that (i) and (ii) ask too much. Most mental events *m* seem not to be localizable in any specific portion of the brain; determination entailing coincidence, their physical determinations *p* will not be localizable either (thus *p* might be the event of my falling into a certain overall neural condition). Perhaps no mental event is localizable, but if *m* is an exception, its physical determination *p* will have a partly extrinsic essence (thus *p* might be my C-fibers' firing in normal neural surroundings). So-called "wide content" mental events raise related but different problems which I don't discuss. Possibly they will have to be allowed as exceptions to the physical/mental determination thesis; in that case, the paper should be read as defending the causal potency of *other*-than-"wide content" mental events. Two remarks, though, to put this in perspective. First, it is controversial how often such events are genuinely efficacious, in particular because their "narrow" counterparts seem ordinarily to be more commensurate, in the sense of ● Q3 section 8, with their supposed effects ●(see Fodor (1987), ch. 2, and (1991) Second, determination is only the most obvious of a number of intimate identity-like relations equally unsupportive of the "$x_1$ was sufficient, so $x_2$ was irrelevant" reflex. Neither of Beamon's outjumping the competition and his jumping 29′ 2 1/4″ determines the other; but nobody would think the latter irrelevant to his being awarded the gold medal because the former was sufficient (see Heil and Mele (1991)).

renders insusceptible to causal influence from any other source, *m* included. Assuming, for example, that all it took for me to wince, clutch my brow, and so on, was my antecedent physical condition, everything else was strictly by the way. Since my headache is a different thing from its determining physical basis, it is not a *bona fide* causal factor in my headache behavior.

By now the deficiencies of this line of argument must be apparent. Suppose that we think of the exclusion principle as saying that for every irreflexive relation *R* (every "form of nonidentity"), and every *R*-related pair *x* and $x^*$, *x*'s causal sufficiency for an effect entails $x^*$'s causal irrelevance. Though there may be irreflexive relations *R* whose relata *do* contend for causal influence as the principle says, for many *R*s this competition arises only sometimes, and for others it *never* arises. Ironically, $R =$ causation is a case in point. Let *x* be causally sufficient for *y*. Then taken at its word, the exclusion principle predicts that *y* owes nothing to the causal intermediaries by which *x* brings *y* about. When *R* is causation's converse, the prediction is different but still absurd: events causally antecedent to *x* can claim no role in *y*'s production.[52] Of course, the case that interests us is $R =$ the determination relation. Remember Archimedes' excited outburst on discovering the principle of displacement in his bath. Assuming that his shouting "Eureka!!" was causally sufficient for his cat's startled flight, nobody would think that this disqualified his (simply) shouting from being causally relevant as well. And it would be incredible to treat Socrates' *drinking* the poison as irrelevant to his death, on the ground that his *guzzling* it was causally sufficient.

Thinking of causal influence as something that an effect's would-be causal antecedents compete over in a zero-sum game, the exclusion principle looks not unreasonable. If the causally sufficient antecedent monopolizes *all* the influence, then the others are left with none. To judge by the examples, though, causation is not like that: rather than competing for causal honors, determinables and their determinates seem likelier to share in one another's success. Again the application to mental and physical events is anticlimactic. Unless an arbitrary exception is to be made of them, it is no argument at all for the causal irrelevance of, say, a sensation that its occurring in some specific physical way was causally sufficient.[53] With events, as with properties, physical determinates cannot defeat the causal pretensions of their mental determinables.[54]

[52] Goldman (1969) and Kim (1989) make related observations.

[53] Lately there has been a tendency to argue that *p*'s causal sufficiency for an effect, though it does not *directly* entail *m*'s irrelevance, limits *m*'s role to that of a causal overdeterminant at best (see note 6); that *m* is indeed irrelevant then emerges from the fact that the effect is not overdetermined. With as much or little plausibility, one could argue that Ella's singing at over 70 db was irrelevant to the glass's breaking, since the latter was causally guaranteed, but not overdetermined, by her singing at 80 db exactly.

[54] Suppose that causal sufficiency is read in some fairly demanding way, say, as requiring the strict nomological impossibility of *x*'s occurring without *y*'s doing so. Then no physical event *p* with hopes of determining a mental event *m* is likely to be itself causally sufficient for *m*'s apparent effect *y*. For *p* can determine *m* only if they are the same size, and nothing that small—assuming

## 8.

To this point our position is wholly negative: for all that the exclusion argument shows, mental phenomena *can* be causally relevant compatibly with the causal sufficiency of their physical bases. It is a further question whether they *will* be in any particular case. And even if some mental antecedent *is* causally relevant, it is a further question yet whether it actually *causes* the effect.

Notice some important differences between causal relevance and sufficiency, on the one hand, and causation, on the other: $x$ can be causally sufficient for $y$ even though it incorporates enormous amounts of causally extraneous detail, and it can be causally relevant to $y$ even though it omits factors critical to $y$'s occurrence. What distinguishes causation from these other relations is that causes are expected to be *commensurate* with their effects: roughly, they should incorporate a good deal of causally important material but not too much that is causally unimportant. And this makes causation special in another way. Although determinables and determinates do not compete for causal *influence*, broadly conceived as encompassing everything from causal relevance to causal sufficiency, they *do* compete for the role of *cause*, with the more commensurate candidate prevailing. Now I argue that the effect's mental antecedents often fare *better* in this competition than their more determinate physical bases.[55]

Inspiring the commensuration constraint is a certain platitude: the cause was the thing that "made the difference" between the effect's occurring and its not. Had the cause been absent, the platitude seems to say, then (i) the effect would have been absent too, but (ii) it *would* have occurred if the cause had. Thus effects are *contingent* on their causes:

(C) If $x$ had not occurred, then $y$ would not have occurred either;[56]

anyway that its essence is not unconscionably extrinsic—can nomologically guarantee any but the most trifling and immediate results. Let it be granted, then, that $p$ is not causally sufficient for $y$; that honor falls instead to a spatially more extensive physical event $p'$, whose occurrence essentially requires, in addition to $p$'s occurring, that the surrounding physical conditions be approximately as they are in fact. This affects the question of $m$'s causal potency, *only* if there is more causal rivalry between $m$ and $p'$ than we found between $m$ and $p$ (namely, none). But, how could there be? What dispelled the illusion of rivalry between $m$ and $p$ was that $p$'s occurrence consisted, in part, in $m$'s occurrence, and that is as true of $m$ and $p'$ as it was of $m$ and $p$: for $p'$ to occur is for $m$ to occur in a certain physical way, and in a certain physical environment. So $p'$ poses no greater threat than $p$ to $m$'s causal aspirations.

[55] To keep things simple, I'll focus on mental events; there is a related story about mental properties.

[56] For definiteness, we interpret would-counterfactuals Stalnaker's way: 'if it had been that $P$, then it would have been that $Q$' is true iff $Q$ is true in the $P$-world best resembling actuality; where it is indeterminate which $P$-world that is, the condition must hold on all admissible ways of resolving the indeterminacy. Might-counterfactuals, 'if it had been that $P$, then it might have been

and causes are *adequate* for their effects:

(A) If $x$ had not occurred, then *if it had, $y$* would have occurred as well.[57]

Without mentioning determination explicitly, these conditions do nevertheless discover causal differences between unequally determinate events. Suppose we stipulate that it contributed nothing to Socrates' demise that he guzzled the hemlock rather than simply drinking it. Then Xanthippe is mistaken when, disgusted at Socrates' sloppy habits, she complains that his *guzzling* the hemlock caused his death. Assuming that the drinking would still have occurred, if the guzzling hadn't, (C) explains the error nicely. Even without the guzzling, the death would still have followed on the drinking. So while Socrates' death may have been contingent on his drinking the hemlock, it was *not* contingent on his guzzling it.[58]

Here the contingency condition exposes an overly determinate pretender; sometimes, though, the pretender's problem is that it is not determinate enough. Safety valves are designed to open quickly under extreme pressure, thus easing the burden on the equipment upstream. This particular valve has begun to operate as advertised when a freak molecular misalignment stiffens the mechanism; this decelerates the opening to just past the point of endurance and the boiler explodes. Assuming that the explosion does *not* result from the valve's opening *per se*, I ask why not. Because the contingency condition is violated? But we can arrange it so that the explosion *was* contingent on the opening, say, by stipulating that if the opening had not occurred, rather than the boiler's exploding the connecting pipe would have burst. Adequacy does better: given the unlikelihood of the molecular mishap, had the opening failed to occur, it might easily have been quicker if it had.[59] Speaking then of how things *would* have been if not for

that $Q$', are true just in case their associated would-counterfactuals, 'if it had been that $P$, it would have been that not-$Q$', are *not* true. Equivalently, a might-counterfactual holds iff on at least one admissible selection of closest $P$-world, the closest $P$-world is a $Q$-world. ●(See Lewis (1981) and Stalnaker (1981*a*) and (1981*b*))

[57] Rasmussen (1982) contains the only explicit reference to (A) that I have seen. There it is argued, fallaciously I think, that (A) follows from (C) on the assumption that $x$ and $y$ actually occur. Another erroneous criticism, encountered mostly in conversation, is that (A) is trivial given just the occurrence of $x$ and $y$: (A) is true iff $y$ occurs in the nearest $x$-containing world $w$ to the nearest $x$-omitting world $v$ to actuality; but since $x$ actually occurs, the nearest $x$-containing world $w$ to $v$ is the actual world, which contains $y$ by hypothesis. This forgets that $w$ is the actual world only if no $x$-containing worlds are nearer to $v$ than the actual world is, and that some are bound to be if, as seems likely, the actual world sits in the interior of a neighborhood of $x$-containing worlds. (What *does* follow trivially from the occurrence of $x$ and $y$ is the condition that if $x$ had occurred, so would have $y$; this is why we use (A) despite its greater complexity.)

[58] David Lewis puts contingency to similar use in his (1986*a*).

[59] I emphasize that the decelerating stiffness sets in only *after* the opening gets under way because I want it to be clear that *that very opening* could have been less protracted (as opposed to: a slower opening could have occurred in its place). To deny this would be to hold that the opening, once begun, *could not* have continued apace, that is, that the approaching deceleration was essential to it. As for the further claim that it *might* have been less protracted, suppose if you like that

the opening, it cannot be said that, *were* it to have occurred, it would still have brought the explosion in its wake.

Important as they are, contingency and adequacy capture the commensuration intuition only partly. Imagine that Socrates, always a sloppy eater, had difficulty drinking without guzzling, to such a degree that if the guzzling hadn't occurred, the drinking wouldn't have either. Then Socrates' death *was* contingent on his guzzling the hemlock; and so more than contingency is needed to explain why it was not the effect of his doing so. Intuitively, it appears that not *all* of the guzzling was needed, because there occurred also a lesser event, the drinking, which would still have done the job even in the guzzling's absence. By hypothesis, of course, without the guzzling this lesser event would not have taken place; but that doesn't stop us from asking what would have happened if it had, and evaluating the guzzling on that basis. Suppose we call *x required* for *y* just in case

(R) For all $x^- < x$, if $x^-$ had occurred without $x$, then $y$ would not have occurred.

Then what disqualifies the guzzling is that, given the drinking, the death did not require it.

Symmetry considerations suggest the possibility of a condition complementary to (R), and a variation on the valve example shows that one is in fact needed. Imagine that the mechanism stiffens, not extemporaneously as above, but because of a preexisting structural defect that would have decelerated the opening in any case. Presumably this means that if the opening had not occurred, it would still have been protracted if it had, and the explosion would still have ensued. Since now the opening *is* adequate for the effect, the problem with taking it for the cause lies elsewhere; and the obvious thought is that the effect required something more. Thus define *x* as *enough* for *y* iff no more than *x* was required:

(E) For all $x^+ > x$, $x^+$ was not required for $y$.

Because the valve's *slowly* opening was required for the explosion, its opening *per se* was not enough; and that is why it was not the cause.

When all of the conditions are met—that is, *y* is contingent on *x*, and requires it, and *x* is adequate, and enough, for *y*—*x* will be called *proportional* to *y*. FN:60 Without claiming that proportionality is strictly necessary for causation,[60] it seems clear that faced with a choice between two candidate causes, normally the more proportional candidate is to be preferred. Which of the contenders

indeterminism holds, and that the misalignment's objective probability, conditional on preceding events, was extremely low. (The relation between 'would' and 'might' is described in note 56.)

[60] Because of the problems of preemption, overdetermination, and so on, strictly necessary conditions on causation are extremely hard to find. As far as I know, philosophers have not succeeded in turning up even a single one, beyond the trivialities that cause and effect should both occur and be suitably distinct (see Lewis (1986*a*)).

proportionality favors depends, of course, on the effect in view; Socrates' drinking the hemlock is better positioned than his guzzling it to cause his death, but relative to other effects proportionality may back the guzzling over the drinking.

More to the present point is the following example: I arrive on your doorstep and, rather than knocking, decide to press the buzzer. Epiphenomenalist neuroscientists are monitoring my brain activity from a remote location, and an event *e* in their neurometer indicates my neural condition to be such-and-such. Now, like any mental event, my decision *m* has a physical determination *p*, and the question arises to which of these the neurometer reading *e* is due. The scientists reason as follows: Because the neurometer is keyed to the precise condition of his brain, *e* would not have occurred if the decision had been taken in a different neural way, in particular if it had occurred in *p*'s absence. So *m* was not enough FN:61 for *e*;[61] *p* on the other hand looks *roughly* proportional to *e*, and so has the better claim to cause it. Another triumph for epiphenomenalism!

Everything is all right except for the last step. What is true is that *this* mental event did not cause *that* effect. But who would have thought otherwise? When an effect depends not simply on an event's occurring, but on its occurring in some specific manner, one rightly hesitates to attribute causation. Taking the meter reading to result from my decision would be like attributing Zsa Zsa's speeding citation to her driving through the police radar *per se*, or the officer's ~~abrasions~~ scratches to her touching his face.

Then when *do* we attribute effects to mental causes? Only when we believe, I can only suppose rightly, that the effect is relatively insensitive to the finer details of *m*'s physical implementation. Having decided to push the button, I do so, and the doorbell rings. Most people would say, and I agree, that my decision had the ringing as one of its effects. Of course, the decision had a physical determination *p*; but, most people would also say, and I agree again, that it would still have been succeeded by the ringing, if it had occurred in a different physical way, that is, if its physical determination had been not *p* but some other physical event. And this is just to say that *p* was not *required* for the effect.

Remember that this makes no prediction about what would have happened if the decision had occurred in *whatever* physical way, but speaks only of what transpires in the *nearest* world where its physical implementation was not as actually—the world in which it undergoes only the minimum physical distortion required to put its actual implementation out of existence. Maybe, of course, we were wrong to think that the ringing would still have occurred in that world; if so, then let us hurry to withdraw the assertion that the

[61] Strictly speaking this assumes that *p* was required for *e*—in other words, that *each* of *p*'s determinables, not just *m*, is such that if it had occurred in *p*'s absence, *e* would not have ensued. (For the interpretation of (R) and (E)'s event-quantifiers see ~~Yablo, S. "Cause and Essence,"~~ Yablo (1992), sec. 11.)

decision caused it (the real cause is some physically more determinate event). But if not, then our conclusions should be these (where $r =$ the doorbell's ringing):

(i) *m* is a counterexample to *r*'s requiring *p* (for *r* would still have occurred, if *m* had occurred without *p*);
(ii) *p* is not proportional to *r* (since *r* does not require it);
(iii) *p* does not cause *r* (since it is not proportional to *r*);
(iv) *p* is not a counterexample to *m*'s enoughness for *r* (it could be a counterexample only if *r* required it);
(v) *p* is not a counterexample to *m*'s proportionality with *r* (by inspection of the remaining conditions);
(vi) *p* poses no evident threat to the hypothesis that *m* caused *r*.

Here are the beginnings, at least, of a story wherein a mental event emerges as better qualified than its physical basis for the role of cause. I believe that this *kind* of story is enacted virtually wherever common sense finds mental causation.

## 9.

Indeterministic scruples aside, everything that happens is in strict causal consequence of its physical antecedents. But causally necessitating is a different thing from causing, and the physical has no monopoly on causation. Among causation's prerequisites is that the cause should be, as far as possible, commensurate with its effect; and part of commensuration is that nothing causes an effect which is essentially overladen with materials to which the effect is in no way beholden. This, though, is a condition of which would-be physical causes often fall afoul, thus opening up the market to less determinate events with essences better attuned to the effect's causal requirements. Sometimes, these events are mental; and that is how mental causation happens.

In a "Concluding Unscientific Postscript" to "The Conceivability of Mechanism," Malcolm remarks that

> it is true for me (and for others, too) that a sequence of sounds tends to lose the aspect of speech (language) when we conceive of those sounds as being caused neurologically. . . . Likewise, a sequence of movements loses the aspect of action . . . ;

FN:62 and he asks, "Is this tendency due to a false picture or misleading analogy?"[62] Many philosophers, anxious to defend the possibility of speech and action, have struggled to articulate what the analogy is which so misleads us. But maybe we are *not* misled to think that outcomes caused ~~effected~~ by their physical antecedents are

[62] Malcolm (1982), 149.

neither speech nor action, nor expressions of any sort of human agency. Maybe the mistake was to think that outcomes of the kind normally credited to human agency are caused by their physical antecedents.[63]

FN:63

## REFERENCES

Block, N., and Fodor, J. (1980). "What Psychological States are Not". In N. Block (ed.), *Readings in Philosophy of Psychology*, Cambridge: Cambridge University Press, Vol. 1 pp. 237–50.

Boyd, R. (1980). "Materialism without Reductionism". In N. Block (ed.), *Readings in Philosophy of Psychology*, Cambridge: Cambridge University Press, i. 67–106.

Broad, C. D. (1925). *Mind and its Place in Nature*. London: Routledge & Kegan Paul.

Campbell, K. (1970). *Body and Mind*. New York: Macmillan.

Davidson, D. (1980). "Mental Events". In *Essays on Actions and Events*, Oxford: Oxford University Press, 207–24.

Descartes, R. (1969). *The Essential Descartes*, ed. M. Wilson. New York: New American Library.

____ (1984). *The Philosophical Writings of Descartes*, ii, ed. J. Cottingham, R. Stoothoff, and D. Murdoch. Cambridge: Cambridge University Press.

Feigl. H. (1970). "Mind-Body, Not a Pseudo-Problem". In C. V. Borst (ed.), *The Mind–Brain Identity Theory*, New York: St Martin's Press, 33–41.

Fodor, J. (1987). *Psychosemantics*. Cambridge, Mass.: MIT Press.

____ (1989). "Making Mind Matter More". *Philosophical Topics*, 17: 59–79.

____ (1991). "A Modal Argument for Narrow Content". *Journal of Philosophy*, 88: 5–26.

Goldman, A. (1969). "The Compatibility of Mechanism and Purpose". *Philosophical Review*, 78: 468–82.

Haugeland, J., (1982). "Weak Supervenience". *American Philosophical Quarterly*, 19: 93–103.

Heil, J., and Mele, A. (1991). "Mental Causes". *American Philosophical Quarterly*, 28: 49–59.

Honderich, T. (1982). "The Argument for Anomalous Monism". *Analysis*, 42: 59–64.

____ (1988). *Mind and Brain: A Theory of Determinism*. Oxford: Oxford University Press.

Johnson, W. E. (1964). *Logic*. New York: Dover.

Kim, J. (1979). "Causality, Identity, and Supervenience in the Mind–Body Problem". *Midwest Studies in Philosophy*, 4: 31–50.

____ (1983). "Supervenience and Supervenient Causation". *Southern Journal of Philosophy*, supp. vol. 22: 45–6.

____ (1984*a*). "Concepts of Supervenience". *Philosophy and Phenomenological Research*, 45: 153–76.

[63] Obviously these remarks cannot hope to resolve *all* the problems that physical determinism has been thought to raise for agency; they are directed only at the outright contradiction between agency and determinism's alleged consequence epiphenomenalism. There my solution is: deny epiphenomenalism.

Kim, J. (1984*b*). "Epiphenomenal and Supervenient Causation". *Midwest Studies in Philosophy*, 9: 257–70.

——— (1988). "Supervenience for Multiple Domains". *Philosophical Topics*, 16: 129–50.

——— (1989). "Mechanism, Purpose, and Explanatory Exclusion". *Philosophical Perspectives*, 3: 77–108.

Kripke, S. (1980). *Naming and Necessity*. Cambridge, Mass.: Harvard University Press.

Lewis, D. (1981). "Counterfactuals and Comparative Possibility". In W. L. Harper, R. Stalnaker, and G. Pearce (eds.), *Ifs*, Dordrecht: Reidel, 57–86.

——— (1986*a*). "Causation" with "Postscripts". In *Philosophical Papers*, ii, Oxford: Oxford University Press, 159–213.

——— (1986*b*). "Events". In *Philosophical Papers*, ii, Oxford: Oxford University Press, 241–69.

Loewer, B., and Lepore, E. (1987). "Mind Matters". *Journal of Philosophy*, 84: 630–42.

——— ——— (1989). "More on Making Mind Matter More". *Philosophical Topics*, 17: 175–91.

Macdonald, C., and Macdonald, G. (1986). "Mental Causation and Explanation of Action". In L. Stevenson, R. Squires, and J. Haldane (eds.), *Mind, Causation, and Action*, Oxford: Basil Blackwell, 35–48.

McLaughlin, B. (1989). "Type Epiphenomenalism, Type Dualism, and the Causal Priority of the Physical". *Philosophical Perspectives*, 3: 109–35.

Malcolm, N. (1982). "The Conceivability of Mechanism". In G. Watson (ed.), *Free Will*, Oxford: Oxford University Press, 127–49.

Prior, A. (1949). "Determinables, Determinates and Determinants (I, II)". *Mind*, 58: 1–20, 178–94.

Putnam, H. (1980). "The Nature of Mental States". In N. Block (ed.), *Readings in Philosophy of Psychology*, Cambridge: Cambridge University Press, 223–31.

Rasmussen, S. (1982). "Ruben on Lewis and Causal Sufficiency". *Analysis*, 42: 207–11.

Schiffer, S. (1989). *Remnants of Meaning*. Cambridge, Mass.: MIT Press.

Smart, J. (1959). "Sensations and Brain Processes". *Philosophical Review*, ii 68: 141–56.

Is this right?

Sosa, E. (1984). "Mind–Body Interaction and Supervenient Causation". *Midwest Studies in Philosophy*, 9: 271–81.

Stalnaker, R. (1981*a*). "A Defense of Conditional Excluded Middle". In W. L. Harper, R. Stalnaker, and G. Pearce (eds.), *Ifs*, Dordrecht: Reidel, 87–104.

——— (1981*b*). "A Theory of Conditionals". In W. L. Harper, R. Stalnaker, and G. Pearce (eds.), *Ifs*, Dordrecht: Reidel, 41–55.

Stoutland, F. (1980). "Oblique Causation and Reasons for Action". *Synthese*, 43: 351–67.

Yablo, S. (1990). "The Real Distinction between Mind and Body". *Canadian Journal of Philosophy*, supp. vol. 16: 149–201; Ch. 1 above.

——— (1993). "Is Conceivability a Guide to Possibility?". *Philosophy and Phenomenological Research*, 53: 1–42; Ch. 2 above.

TO BE ADDED

Yablo, S. (1987). "Identity, Essence, and Indiscernibility." *Journal of Philosophy*, 84:293–314.

Yablo, S. (1992). "Cause and Essence". *Synthese,* 93:403–449.

Yablo, S. (1999). "Intrinsicness". *Philosophical Topics*, 26(1/2):479–505.

## Queries in Chapter 8

Q1. Please check and confirm the author correction here.

I put what made sense to me but it's a question for the copy-editor

Q2. Please provide closing parentheses.

Q3. Please provide closing parentheses

Q4. please provide closing parentheses.

# 9

# Singling out Properties

Colors have characteristic causes and effects—that we do know.

Wittgenstein, *Remarks on Color*

## I

Epistemologists used to be an exasperating bunch. ~~Who can forget the story of how epistemologists used to do business?~~ Having offered to help you clarify *how* you know that *p*, they would gasp at your reasoning and declare that you *don't* know it—or rather *wouldn't*, if not for a reconstruction of your procedures invented by themselves. Then they would be gone, leaving you struggling to reconceive your relations to *p* along recommended lines.

Of course this, the style of epistemology Putnam once satirized as "intellectual Walden Two",[1] is now defunct. But the spirit animating it lives on. What the metaphysician offers to clarify is not how *p*'s truth is *known*, but what *makes p* true. Trouble is (you can guess the rest), a review of all likely truth-makers reveals that *nothing* does. Or rather nothing *would*, if not for a certain *substitute* truth-maker identified by the metaphysicians themselves. Whoever would persist in counting *p* true is thus forced to reconceive its truth as flowing from unexpected sources.

With apologies to Putnam, this approach to metaphysics might be called "ontological 1984", in view of the Party's penchant for tampering with the truth grounds of everyday statements. Statements about the past, O'Brien explains, are true in virtue of what is preserved in records and memories; numerical claims

This paper began life as comments on Mark Johnston's contribution to the 1989 Colloquium on Color at the University of Michigan. Later versions were read at Queensland University, Australian National University, Monash University, UC Riverside, and UC San Diego; thanks to all who participated, especially Andre Gallois, Peter Menzies, Philip Pettit, and Michael Smith. Closer to home I had help from Mark Crimmins, Allan Gibbard, Sally Haslanger, David Hills, Ted Hinchman, Joseph Levine, Gideon Rosen, David Velleman, and José Zalabardo. I learned of John Campbell's related ideas ("A Simple View of Color") from Crispin Wright.

[1] "Why Reason Can't Be Naturalized" in Putnam (1983). Perhaps I'm not using the phrase in exactly his sense.

owe their truth values to the Party's stipulations. "You are no metaphysician, Winston," O'Brien says when his prisoner boggles at some such revelation. Using the term in O'Brien's sense, this paper explores some strategies for not being a metaphysician.

## II

Here is the scenario. There is a predicate "F", and we have various things we want to say with it—things we regard as quite likely true, or even certainly true. But we are troubled. Granted that "F" applies to roughly the objects we suppose, what property or relation or condition or what have you does it apply *in virtue of*? Information about the property is not lacking; there are the various statements we are inclined to make using "F", and these, being presumably correct, add up to a considerable data base. What bothers us is that the information feels *circumstantial*; we learn what the property is *like*, not what it *is*. The urge thus arises to *identify* Fness,[2] to single it out and elucidate its nature. Only then will we know what makes our predications true.

This sort of scenario is enacted everywhere in philosophy, as a few examples will bring out. Naturalized semanticists have charged Tarski with promising, but failing, to explain what truth is. Tarski *partly* explained truth when he reduced it to denotation; it falls to us, though, to complete Tarski's task by finding the relation, presumably naturalistic, that words bear to the objects they denote.[3] Or take the debate in the philosophy of mind as to which feature of a person makes her correctly describable as "believing that *p*". Is it a relation she bears to a sentence of mentalese, or the attribution's making best sense of her behavior, or the fact that we, in a similar state, would declare that *p*? The pattern even extends to ethics, where the search is on for the property that "perfectly deserves the name 'value' ".[4] Whether value attributions can be true is disputed, of course, but if they can, the thought is, it ought to be possible to produce for inspection the property that gives them this status.

So: by the looks of things, there are lots of properties standing in need of further identification. However, it's in connection with *qualitative* properties—secondary and phenomenal—that the identification problem becomes really acute. This is because qualitative properties can seem not merely undiscovered but positively *hidden*.[5]

According to one way of drawing the primary/secondary distinction, primary qualities are well represented by our ideas of them. When veridical, which is not

FN:2 FN:3 FN:4 FN:5

[2] I use "Fness" stipulatively as the property in virtue of which "F" applies; it's not a foregone conclusion that this will be the same property in every case.

[3] Field 1972. Field might not endorse this project now.

[4] Lewis 1989, 136. See also Railton 1986, Boyd 1988, Johnston 1989, and Smith 1989.

[5] Oddly, they can also seem entirely open to view; see J. Campbell 1993.

always, these ideas portray shape, size, number, and so on as they really are. With secondary qualities, it is nearly the reverse: our ideas of taste, smell, color, and so on, though seldom false,[6] convey a very *poor* sense of their associated properties. Thus Reid:

> our senses give us a direct and distinct notion of the primary qualities, and inform us what they are in themselves: but of the secondary qualities . . . [they] inform us only, that they are qualities that effect us in a certain manner . . . as to what they are in themselves, our senses leave us in the dark.[7]

For this reason, "the nature of the secondary qualities is a proper subject of philosophical disquisition".[8] Happily, "philosophy has made some progress" in the matter.[9] But philosophy would not be needed if experience had not left these qualities' identities so obscure.

Now for Reid, the problem about secondary qualities grows out of a contrast between the qualities themselves, which are unknown, and the sensations they produce in us, which are known full well. So he would not himself see the problem as extending to *phenomenal* properties, like that of sensing greenly or suffering pain. Others however find the cases analogous—some, like Descartes, mentioning them in the same breath:

> If someone says he sees color in a body or feels pain in a limb, this amounts to saying that he sees or feels something there of which he is wholly ignorant . . . [10] CSM(1)217.

Should this be in the footnote?

Pain-experience, though generally veridical in the sense of occurring only when one is really in pain, gives little indication of what that condition really amounts to. But if we cannot know pain just on the basis of our experience of it, how to proceed? Research will have to be conducted into pain's identity; somewhere a state lies waiting such that identifying pain with *it* honors, in Dennett's words, "all, or at any rate most, of our intuitions about *what pain is*".[11]

## III

These examples remind us of what is fast becoming standard operating procedure in metaphysics: gather a body of more or less central preconceptions

[6] Aristotle says, "I call that sense-object special that does not admit of being perceived by another sense and *about which it is impossible to be deceived*, as sight is connected with color, hearing with sound, and taste with flavor" (*De Anima*, II, 6; my emphasis). Likewise Locke: "blue and yellow, bitter or sweet can never be false ideas" (*Essay Concerning Human Understanding*, II, 32, 16). This view of Locke's is noted in Curley 1972, 463.

[7] Reid, *Essays on the Intellectual Powers of Man*, II, 17. According to Locke, "the ideas produced in us by these secondary qualities have no resemblance of them at all" (*Essay*, II, 8, 15).

[8] *Essays*, II, 17.

[9] Ibid.

[10] *Principles*, I, 68. Note that Descartes speaks not of pain, the feeling, but pain the thing felt; I ignore the many issues this distinction raises.

[11] Dennett 1978, 224.

associated with some predicate, and then, guided by whatever clues philosophy and science have to offer, strike out in search of the associated state or property. That a procedure is standard doesn't make it right, of course. But there *do* seem to be reasons for wanting to know the worldly correlates of our predicates.

Nothing counts as a theory of *pain*, Dennett says, unless it honors enough of our intuitions about what pain is. Intuitions being the impromptu, unexamined things they are, though, "a prospect that cannot be discounted is that these intuitions do not make a consistent set".[12] How might we ward that prospect off? The obvious strategy would be to *produce* a state making the intuitions come true. And this seems little different from identifying pain itself.

Rather more worrisome, because less remote, is the prospect that our intuitions clash not among themselves, but with views we hold about the larger world. There is nothing *internally* incoherent, according to Mackie, in the idea of "objectively prescriptive" value properties, or "colors as we see them belong[ing] intrinsically to the . . . surfaces of objects";[13] it's just that other things we think rule out anything's actually *possessing* such properties. Again, the obvious response would be to identify redness with (say) $R_{295}$, a property conspicuously at peace with our larger theory. This is a second reason we might want Fness identified.

"[D]oes it not appear a contradiction", Reid asks, "to say we know that fire is hot, but we know not what heat is?"[14] Even to *understand* the statement that fire is hot, some would say, requires knowledge of what heat is.[15] These ~~formulations~~ ways of putting the point can be faulted, ~~no doubt,~~ for treating knowledge and understanding as all-or-nothing affairs. But the underlying idea seems right: the *better* one knows what Fness is, the better one understands "X is F", and the less superficial one's knowledge that it is F. This gives a third motive for the project of singling Fness out.

Next is the motive of simple curiosity. Questions may arise about a property whose answers depend, or seem to depend, on the property's further identification. Is pain *essentially* painful? Kripke says yes and concludes that pain is not identical to anything physical. But matters arguably belong the other way around: "an opinion on [the] truth or falsity [of the essential painfulness claim] waits upon a theory of what pain *is*".[16]

What, finally, if we have *no* intrinsic interest in Fness, and want only to *communicate* using the predicate "F"? Not even this, it seems, can excuse us from the identificatory project. For communication has a metaphysical side; unless both parties are using "F" in reference to *the same property*, they are talking past

[12] Dennett 1978, 224.

[13] Mackie 1977, 35; Mackie 1976, 19.

[14] Reid, *Essays*, II, 17.

[15] See the literature on Russell's principle that lack of acquaintance with any constituent of a proposition prevents understanding; for instance, Russell 1912, 58, and Evans 1985*b*, ch. 4.

[16] Jackson *et al.* 1982, 216.

each other. Whether this proviso holds, however, depends on (i) the property that we ourselves attribute by use of "F", and (ii) the property that our interlocutors do. This gives us an interest in (i); and an interest in (i) is an interest in the identity of Fness.

With so many reasons for seeking after Fness's identity, how could anyone object? Nonetheless I do object, if not to the project's goal, then in some cases to the project itself. Seeking after a property's identity makes sense only if its identity is not yet known. That is, there has to be a *better* way of conceiving Fness than what we have already, such that knowing what Fness is means conceiving it like *that. But the existence of a better idea of Fness is not something that can be assumed in advance*. The next three sections look at some of the trouble this assumption s ~~can~~ cause; the rest of the paper experiments with dropping the assumption and getting out of the identification racket.[17]

## IV

Remember the scenario: we have a body of doctrine involving some predicate "F", and we seek a better idea of the property we attribute with it. Asked what's wrong with our existing idea, we'd complain of knowing the property only indirectly, as the whatever-it-is meeting certain conditions. These conditions might take a number of forms, but in the usual examples they are causal. Pain is the state that is brought about by tissue damage, that prevents concentration, that prompts avoidance behavior, and so on. Red is the property producing a certain type of experience in suitably placed observers. While all of this is useful information, the complaint goes, none of it tells us what the property *is* as opposed to what it *does*.

To which the obvious reply is, given enough information about what something *does*, it ought to be possible to track it down and make a positive identification. This is the strategy Armstrong employs in *A Materialist Theory of Mind*. Having argued on philosophical grounds that "the concept of a mental state is the concept of a state of the person apt for the production of certain sorts of behavior", he proceeds to ask, "What in fact is the nature of these inner states?" This last is said to be a matter not for philosophy but for "high-level scientific speculation".[18] Armstrong envisages a two-part procedure then. Calling a property *F-ish* if it satisfies the main presumptions about Fness, philosophical analysis reveals that

(1) Fness = the F-ish property.

[17] This oversimplifies the eventual proposal; see section VII.

[18] Armstrong 1968, 89–90. He treats color similarly. Redness is identified "by reference to the way it happens to affect us and by mentioning objects that happen to be red" (ibid. 276). And "just as there arises the question what, as a contingent matter of fact, a mental state is, so there arises the question what, as a contingent matter of fact, the property of redness actually is" (ibid. 277).

Next, the scientists are brought in to tell us which property is in fact F-ish. Not that we can't make an educated guess. Armstrong himself thinks that "the identification of [mental] states with physico-chemical states of the brain is, in the present state of knowledge, nearly as good ^ as the identification of the gene ^a bet with the DNA molecule".[19] And "from the standpoint of total science, the most *plausible* answer is that redness is a purely physical property".[20]

Now, though, we run into a famous problem. When Armstrong says that the red-role is played by a physical property, he is only playing a scientific hunch; there *could*, he admits, be "*irreducibly* diverse causes in the physical surfaces bringing about identical colour-appearances for human observers".[21] But where Armstrong puts this forward as a sort of a doomsday scenario in which colors are reduced to the status of pseudo-qualities, nowadays it is thought to be more or less the situation: "apart from their radiative result, there is nothing that blue things have in common . . .".[22] Something similar is of course the standard line on suffering things—they too have nothing but a causal syndrome in common. It begins to seem that Armstrong-style concepts of sensations and colors are concepts of precisely nothing.

## V

What is the property that red things, suffering things, or what have you possess in common, and in virtue of which they are rightly called by those names? While we know, or think we do, how the intended property *behaves*—its causal role—we cannot seem to find a property that covers all and only the cases where the role is actually played.

Or can't we? If what is wanted is a property covering just the cases where a causal role gets played, why not the property of having a property that plays it? This is the solution urged by *dispositionalism* in the philosophy of color,[23] and the souped-up dispositionalism called *functionalism* in the philosophy of mind. The general format is

(2) Fness = the property of having an F-ish property.

Because (2) represents Fness as the property of having a property playing a certain role, let's call it the *role theory* of Fness. ((1) was the *realizer* theory.) Redness, this theory claims, is the property of having a property playing the red-role; pain is the state of being in a state playing the pain-role; etcetera.

[19] Armstrong 1968, ~~91~~ 90. [20] Ibid. 277. [21] Ibid. 289.

[22] Hardin 1984, 496; he bases the conclusion on Nassau 1980. K. Campbell 1969 made the point earlier.

[23] I assume the widely accepted second-order property treatment of dispositions; see Prior 1985. Alternative theories treat dispositions as counterfactual properties or as categorical ones. Against the first alternative see Shope 1978 and Wright 1992; Johnston 1993 attempts a fix. Against the second see Prior *et al.* 1982 and Prior 1985.

Having been crafted with an eye to just this result, the properties championed by the role theory have a large advantage over those championed by the realizer theory, viz. satisfaction of the following key perconception about Fness:

*commonality*: it is shared by all and only Fs.

But this is not the only preconception in play. The thing we know best about Fness, our main intuitive grip on the property, is that Fness is *F-ish*, with all the causal powers that entails. Hence a second key preconception is

*causality*: it has F-ish causal powers.

The question is whether the property of *having* an F-ish property—the property that Fness *is* according to the present theory—is *itself* in relevant respects FN:24 F-ish.[24]

When Reid refers to the colors as the causes of well-known experiential effects, he is only echoing common sense. Unschooled by philosophy, anyone would say that our experiences of a ripe tomato as red are caused by the tomato's *redness*. But is this a view that the dispositionalist can accept? Experiences of color are, for her, *manifestations* of color; they stand to it as sleep stands to dormitivity or death to deadliness. While we may indeed cite dormitivity in the *explanation* of sleep, to indicate that the sleep occurred not by chance but thanks to a sleep-inducing ~~property~~ feature of the dormitive substance, it is famously *not* plausible to say that the sleep-inducing feature was the dormitivity itself. Still less would we seek a place for the mushroom's *deadliness* in the causal ancestry of the ensuing death.

Is the claim that dispositions lack causal powers altogether? Not at all; the mushroom's deadliness might well influence ~~the~~ some property-owner to post warning signs. The claim is not even that dispositions cannot be responsible for the effects they are dispositions to produce. There could be a sleeping potion that FN:25 worked partly, or even entirely, through our recognition of its dormitivity.[25] And part of the magic of the true charmer is his ability to win you over by his very charmingness. (This is to say nothing of the phenomenon of being famous for being famous.) But the lengths one must go to to find such cases brings it home that, special fiddling aside, dispositions do not cause their FN:26 manifestations.[26] Dispositions to produce color-sensations are no exception;

[24] Here we ask whether the role property is F-ish in causal respects, but the question could be generalized. Part of the role of evaluative properties, for instance, is to merit an appreciative response. But does the property of having an appreciation-meriting property *itself* merit appreciation?

[25] Block's nice example; see his 1990.

[26] This is also the conclusion of Prior *et al.* 1982, Prior 1985, Jackson and Pargetter 1987, and Block 1990. Based on his and Jackson's "program model" of causal relevance, Pettit holds dispositions to be causally relevant to their manifestations (Jackson and Pettit 1990). But this model, which counts a higher-level property causally relevant if it "effectively ensures" the instantiation of some causally efficacious lower-level property, seems overly permissive; it makes Brutus's property of being Caesar's killer causally relevant to Caesar's death, and Drano's property of being plumber-recommended causally relevant to the unclogging of my drain.

sensations of redness might be due to various properties of an object, but the property of having a property productive of such sensations is not one of them.[27]

Sure as common sense might be about the causal powers of colors, it is adamant about those of occurrent mental states: pain causes avoidance behavior, itchiness makes us scratch, and so on. Part of the knock against Rylean behaviorism was its reluctance to acknowledge such facts. And part of the attraction of early functionalism was the way it seemed to welcome them and indeed build them into the essences of the relevant states.

But although this is a story functionalists love to repeat, it doesn't really hang together; given the difficulty of getting a state to *cause* the effects by which it is defined, the part about building various outcomes into pain's essence sits ill with the part about preserving pain as their causal basis. What really happened is that the functionalist succeeded rather too well. Determined not to slight pain's effects, she created a state so tightly bound up with those effects that it could no longer bring them about.[28]

## VI

Back to the original problem. When Armstrong proposed to identify Fness as *the* state playing a certain role, the reply came that such a thing might not exist, different states playing the role on different occasions. But why did we assume that a *single* state was required? Is it written in stone that "F" must be attributable for the same reason in every case? No, says Lewis. Pain might be

> one brain-state in the case of men, and some other . . . in the case of mollusks. It might even be one brain state in the case of Putnam, another in the case of Lewis. . . . The seeming contradiction (one thing identical to two things) vanishes once we notice the tacit relativity to context in one term of the identities.[29]

That is, we can agree with the realizer theory that

(1) Fness = the F-ish property

always holds true,[30] without supposing that it expresses the *same* truth in connection with different objects.[31] Rather, (1) resembles "the winner = whoever came in first" in maintaining a constant truth value through coordinated fluctuation in the references of its parts. But then, just as it would be clearer to

[27] Some dispositionalists embrace this result gladly. Secondary qualities, McDowell says, "cannot be credited with causal efficacy" (1985, 188). See also McGinn 1983, 15.

[28] On the causal powers of functional states see Block 1990.

[29] Lewis 1980, 233.

[30] Provided that context supplies a unique F-ish property.

[31] Or different objects at different times—I ignore this complication.

say "the winner of a given race = whoever came in first in that race", we should clarify (1) to

(1*) Fness in a given thing = the property that is F-ish w.r.t. that thing.

This is Armstrong's original realizer theory amended to take account of the fact that different properties might play the F-role in ~~with respect to~~ different objects. But as written, the theory is incomplete.

Imagine that itchiness in Pam is I-fiber firing while itchiness in Sam is I-sac fibrillation. Are we to conclude that there is nothing that Pam and Sam have in common when both feel itchy? No such conclusion follows, says Lewis. What is true is that Pam and Sam don't have their *realizer*-properties in common. But there is also (and still) the *role*-property of having *some* property or other playing the itchy-role for a creature like yourself. This latter property may not be itchiness, but it is a related property, and it's one that Pam and Sam share. For FN:32 the sake of a label, why don't we call it *the property of having itchiness*?[32] And in general why don't we contrast *Fness*, the realizer-property that varies from object to object, with *the property of having Fness*, the role-property that Fs share? Then in addition to replacing (1) with (1*), Lewis would replace (2) with

(2*) the property of having Fness = the property of having an F-ish property.

By this distinction between the property of Fness and that of having Fness, Lewis appears to get the best of both worlds. For whatever names you call them by, he has an itchiness-property that makes him scratch, and an itchiness-property he FN:33 shares with all itchy creatures of whatever physiological make-up.[33]

Where are we? Armstrong's realizer-properties had the right causal powers but were not common to all of the right things; the role theory (once) championed by Putnam had the opposite virtues. Only with Lewis's mixed theory, it seems, do we get both *commonality*—properties shared by all and only F-things—and *causality*—properties with F-ish causal powers. What could be better?

Here is what. Better than a theory offering *two* F-properties, one shared by all FN:34 and only Fs and the other with the intuitive causal powers of Fness,[34] would be a theory offering a *single* F-property, common to all and only Fs *and* possessing the intuitive causal powers of Fness. Otherwise we cannot say that Pam is scratching for the same reason Sam is, viz. itchiness. The closest that a Lewisian can come to this is: Pam and Sam both scratch because both have the property of having

[32] "I mean to deny all identities of the form ⌜*a* is identical with the attribute of having *a*⌝ where *a* is an experience-name definable as naming the occupant of a specified causal role . . . I take 'the attribute of having pain' . . . as a noncontingent name of that state or attribute Z that belongs, in any world, to whatever things have pain in that world" (Lewis 1983, 101).

[33] Similarly, in addition to the pencil's redness, which causes redness's intuitive effects, there is also its higher-order property of having redness, the property it shares with other red things.

[34] Lewis 1986, p. xi: ". . . there is a state common to all who are in pain—'being in pain,' I call it—but it is not pain, and it does not itself occupy the role of pain."

itchiness. But this is like saying that mescaline and Audrey Hepburn movies are enjoyable for the same reason, viz. possession of properties conducive to enjoyment.[35] I conclude that the mixed theory is giving us the reverse of what we want. Rather than similar effects due to a common property, we are getting a common property built around the fact of similar effects.[36]

## VII

Originally we sought to identify "F" 's referent as *the F-ish property*, whatever that might turn out to be. This is recognizably a descendant of Kripke's method in *Naming and Necessity*, except that where we speak of F-ishness, Kripke speaks of the conditions fixing "F" 's reference.[37] That the method works more or less as advertised in connection with terms like "hot" and "gold" can be taken for granted here. But the attempt to extend it to color terms—to identify the referent of "yellow" as

> that (manifest) property of objects which causes them, under normal circumstances, to be seen as yellow (i.e., to be sensed by certain visual impressions)[38]

—runs, it appears, into a familiar sort of trouble; for, as Crispin Wright puts it,

> there may simply fail to be any interesting physical essence underlying the manifestations which have a salient similarity for us . . . we hold out a hostage to fortune in attempting reference fixing of this kind, and the hostage may not be redeemed.[39]

Reference fixing is pointing where one *hopes* an interesting physical essence lies. But some hopes are better founded than others, and it is all too easy to believe that objects making a similar impression on human color sense do so for physically different reasons.

[35] A similar point applies to Jackson and Pargetter on color. For them, "when I say that an object in [circumstances] C3 is red and another in C4 is red also, I am saying that they . . . have something in common. Both have what is redness for me now in their circumstances" (1988, 135). But given their Lewisian framework, *that* commonality consists merely in the fact that both are such as to cause red experiences—which cannot be the reason *why* they would cause red experiences.

[36] That is what we are getting if we insist on commonality. Insist on causality and the problem is different: itchiness does not cause scratching because it's *itchiness*, rather it qualifies as itchiness because (among other things) it causes scratching. This is on top of the fact that it is not shared.

[37] Reference-fixing conditions are generally conceived as concise (this is suggested by the examples), indefeasible (so as to enable a priori knowledge), antecedently graspable (on pain of circularity) specifications of a term's referent. I'm not sure that any of these conditions is strictly demanded by what Kripke (1980) says. As characterized on pp. 34 ff., for example, apriority looks quite *compatible* with defeasibility. And Kripke seems to show ambivalence about antecendent graspability when he describes as "independent of any view argued in the text" the view that "such terms as 'sensation of yellow', 'sensation of heat', 'sensation of pain', and the like, could not be in the language unless they were identifiable in terms of external observable phenomena, such as heat, yellowness, and associated human behavior" (p. 140 n. 71).

[38] Kripke 1980, 140.

[39] Wright 1992, 131.

How could Kripke have missed this problem? A possible reply is: what problem? To speak of the "manifest property of objects which causes them . . . to be seen as yellow" is not to say *anything* about physical essences; the issue of physicality is not addressed.[40] Kripke does tell us that "it is up to the physical scientist to identify the property so marked out in any more fundamental physical terms that he wishes".[41] But this sounds more like a burden-shifting remark ("identify it if you can") than a profession of faith that the scientist will succeed, much less an insistence on physical specifiability as a condition of successful reference fixing.

Whether handing the identificatory ball off to appropriate experts is sound methodology is not the question; we can assume it is. The question is: what if the experts can make no yardage, and the ball is handed back? Emboldened by their reading, or misreading, of Kripke, many philosophers would take a hard line on this, rejecting not-further-identifiable properties as unreal. ("If there is no *saying* what Fness is, most likely it isn't *anything*.") But this is because they have allowed the sensibly unpretentious policy of deferring to expert opinion where it exists to harden into the absurdly self-effacing one of *automatically* discounting ordinary, nonexpert, ways of conceiving properties as superficial. Assume that all properties possess hidden depths, and Fness, which refuses to reveal any (or indeed to reveal much of anything about itself not already imputed by common opinion), takes on the feel of a projective fantasy.

Well, suppose we do *not* allow the sensible policy to harden into the absurd one. Then the fact that Fness is impervious to experts admits of a new interpretation: the way the rest of us conceive Fness is the *right* way if you want to know what Fness is. (The rest of us *become* the experts, if you like.) This is what the "hostage to fortune" line overlooks. Fortune might have had a hand to play if yellowness were a theoretical posit, an I-know-not-what postulated to explain the familiar and known. (Depending on whether the explanation could be made to work, yellowness's claim to reality might or might not hold up.) But the claim here is that yellowness is *itself* something familiar and known; our ordinary, nonexpert ways of conceiving it tell as good a story as any about what it is.

So, then: yellowness is the intrinsic, categorical feature that objects *appear* to have when they look yellow to us, that *makes* them look yellow to us, that yellow things have in common, and so on. Pressed for its "true identity", the best we can do is reiterate the preconceptions (intrinsic, categorical, yellow-look-making, etc.) while insisting that it is not laziness or any other human failing that prevents a fuller answer, but the property itself.

[40] Not in this passage, anyway; see p. 128 n. 66. *Perhaps* Kripke's use of "physical" in note 66 can be read as expressing commitment to an "interesting physical essence" underlying manifestations of yellowness. But so little weight is put on the word that it is hard to feel sure; it might equally be functioning to bring out yellowness's objectivity or intrinsicness. "Perhaps I am rather vague about these questions, but further precision seems unnecessary here."

[41] Ibid. 140.

"Naive objectivism" is the usual name for a view like this—the word "naive" functioning partly to identify the view and partly to mark it as ludicrously simple-minded. This assessment is so deeply ingrained that it is surprising to realize how little explicit argument there is to back it up. "What is beyond dispute," Dennett says,

> is that there is no simple, nondisjunctive property of surfaces such that all and only the surfaces with that property are red . . .[42]

His evidence boils down to the fact that science offers no simple, nondisjunctive, conception of redness. But this is agreed all around. The most that follows is something else agreed all around: naive redness, if it exists, is not a property that scientists have much truck with.

That they should present an attractive face to science is hardly a core presumption about the colors. But it *is* generally presumed that objects *look* colored as a result of being so; and this might seem equally damaging:

> One view about [secondary qualities] seems clearly ruled out. Colors, for example, can't be properties of substances over and above the microstructural properties of them that account for the ways they influence the physical features of the light that impacts on our visual systems . . . To suppose [otherwise] is either to embrace a view about the causation of our perceptual experiences which is known to be false, namely the view that they are caused by something other than the microstructural physical properties of objects, or to embrace the view that secondary qualities are epiphenomenal and play no role in the production of our perceptual experiences . . .[43]

Because our color experiences are fully accounted for in terms of the relevant microstructure, colors understood as over and above that microstructure would be causally otiose. (At best they could aspire to "seconding" whatever causal messages were being sent by the underlying physics.[44]) Moral: if you want your colors causally active, better make them microphysical.[45]

Before accepting this result, consider a parallel argument. The scale at a certain weigh-station is adjusted to sound an alarm whenever a truck weighs in at over 70,000 pounds—in a word, whenever a truck is *heavy*. Enter yourself, on an overloaded semi, and the buzzer sounds. Given how the scale is adjusted, it would seem that your truck's property of being heavy was highly relevant to the alarm's sounding. But think again. I forgot to mention that your truck was *barely* heavy, in the sense of weighing *just* over 70,000 pounds. With the truck's *bare* heaviness

[42] Dennett 1991, 376.

[43] Shoemaker 1990, 116.

[44] Johnston 1992 speaks of "a bizarre pre-established harmony of redundant causes of our visual experience" (pp. 227–8).

[45] Likewise, apparently, for nonphysical *mental* properties. If neurophysiology sufficed to explain behavior, "we would be forced to say that the extra mental properties postulated . . . are causally idle; and that the characteristically Parallelist thesis that the mental is unable to affect the physical order in any way is completely correct" (Armstrong 1968, 47). This is a particular theme of Malcolm 1968.

being *itself* sufficient for the effect, every *other* aspiring cause is left with nothing to do. Apparently, then, the truck's heaviness ("another" aspiring cause after all) made no causal difference to the buzzer's sounding. Moral: if you want your weight-properties causally relevant, make them as determinate as possible.

How can this be, though? To be heavy is *part of what it is* to be barely heavy; and how can a part be crowded out by its containing whole? What the truck story really shows is that when properties are so related that to possess one is part of what it is to possess the other—when they are related as determinable to determinate—they do not compete for causal honors.[46]

Imagine that the scale is constructed on a balance-beam model; if a truck weighs enough to lift the 70,000 pound counterweight, then a circuit is broken and the buzzer sounds. So the mechanism is absolutely insensitive to weight differences above 70,000 pounds. With this in mind, what is the property of the truck whose instantiation resulted in the buzzer's sounding? While it is true that the truck's bare heaviness was sufficient for the effect, if we had to name a property as the one responsible, it would be the heaviness pure and simple.[47] For only the latter is *commensurate* with the effect, in the sense of including what the effect needed with a minimum of irrelevant extras.

So a determinable property, far from being preempted by its determinate, is often *better* placed to function as cause. Couldn't this be how it is with the tomato's surface microstructure and its surface redness? If colors were nonphysical determinables of the microproperties thought to preempt them, then no causal competition would be possible. And because color-properties would be *better* proportioned to our perceptual responses than their microphysical determinates—a rose whose color were otherwise microphysically implemented would look as red—they'd be better placed to play the role of cause.[48]

Does it even make sense, though, to think of redness as a determinable of its physical underpinnings? Normally determinates are taken to *entail* their determinables as a conceptual matter. So even if, as seems plausible, colors are

[46] This is not to say that determinables *inherit* their determinates' causal powers; see Yablo 1992, n. 32. The claim is that determinables are not *preempted* by their determinates.

[47] To speak of a property as causing, or being causally sufficient, for an effect, is short for a similar claim about the property's instantiation; so the truck's *bare* heaviness was sufficient, but its *heaviness* was the cause.

[48] "The position is inconsistent; properties with physical determinates are themselves physical, yet you say color is *not* physical." I deny the assumption that only physical properties can have physical determinates. *Determinates* are properties such that to have them is to have the original property, not simpliciter, but in a certain way; and *physical* properties are properties whose actual and possible possessors have something physical in common, or form a physically natural class. The assumption in question is therefore this: if for at least one way G of being F, the Gs have something physical in common, then the Fs do too. This is wildly implausible. And it remains so even if we strengthen the premise to: every x that is F at all is F in some specific way $G_x$, where the $G_x$s have something physical in common. Why shouldn't a physically unnatural class decompose into physically natural subclasses? (Analogy: Whenever a first-order statement is provable, it is provable in some specific way. But although provability-in-such-and-such-a-way is decidable, provability is not; the decidable parts add up to an undecidable whole.)

*necessitated*, in the metaphysical sense, by their microphysical underpinnings, talk of determinates and determinables is out of place. For plainly, no microphysical state conceptually entails redness.

Yet on second thought, it's the conceptual entailment requirement that's out of place. Determination is above all a relation between *properties*. But as we know from Kripke's examples of "synthetically identical" properties, conceptual entailment is not; a single property, conceived in alternative ways, will have different conceptual consequences. (To put the point in the usual misleading way, entailment relates not properties, but properties-under-a-conception.) Accordingly, we drop the entailment condition and explain determination in wholly metaphysical terms: Fness is a determinate of Gness iff to be F is a *way* of being G. To have your molecules arranged *thusly* is a *way* of being red, so redness is a determinable of the given microproperty.

## VIII

Standing back for a minute from the example of color, what is the picture we have arrived at? The goal was a property Fness that was common to all Fs, and that played an F-ish causal role. Surmising that only a physical property could play the desired role, we were dismayed to learn that *different* physical properties $P_1, P_2$, etc. played it in different objects. But the surmise was wrong: the F-role is best played by a nonphysical determinable with $P_1, P_2$, etc. as determinates. Surmising that coverage of the Fs required a higher-order property existentially generalizing over $P_1, P_2$, etc., we were dismayed to learn that this property had the wrong causal powers. But the second surmise was also wrong: a determinable with $P_1, P_2$, etc. as determinates covers the same extensional ground as its higher-order alternative.

Notice that the very same property, what we might call *Fness as such*, gets overlooked both times, despite offering the only real hope of harmonizing some fairly basic convictions, e.g., that red objects have something in common on account of which they *look* red. The tempting conclusion is that philosophers don't really like this property; they dislike it to the point of suppressing it even in contexts where nothing else will do. Because the reasons for such an attitude will be different in different cases, let's consider the reasons for hostility to properties like *redness* as such, that is, to the naive colors. And let's start with the worry that, despite my optimistic noises above, what these *are* remains to be explained.

ital

## IX

Zinc is explained by showing where it falls in the table of elements; radio waves are explained by pointing to a segment of the electromagnetic spectrum; pokeweed is explained by locating it in the kingdom of plants. To explain what a thing is, then, apparently, one blocks off its ontological neighborhood, enumerates the inhabitants according to some illuminating principle, and indicates which of the enumerated items it is. All that remains is to apply this model to the case of the colors. Explaining what they are would be a matter of spelling out "which properties colors are . . . in a particular set that is acknowledged on both sides to exhaust the properties of material objects".[49]

FN:49

But there is no such set, that I know of. To assume that there is, is to assume that the colors can be caught in a net designed with other sorts of properties in mind. And why should the naive objectivist accept this? *Sui generis*, unscientific, and of minuscule causal impact, colors as she conceives them have nothing to draw the enumerator's attention.

If redness isn't to be picked out on a master list of properties, what *can* be done to calm our concerns about what it is? Attempting to *define* those concerns would be a good start. Wh-questions notoriously require a *context*: some sort of gap or defect in one's information that one is seeking to remedy. This is why it makes little sense to wonder, apropos of nothing, who Frank Sinatra is, or where North America might be found. But suppose that one's information about Sinatra was hearsay from what turned out to be a defective source. Then it *would* make sense to wonder who Sinatra was. Similarly, it would make sense to wonder what redness was if our "source" on it proved defective. But this is exactly the situation, according to some philosophers: our "source" on the colors is color experience, and color experience does not portray the colors as they really are. Either it *conceals* the colors, or (worse) it *deceives* us about them.

Does our experience of color fall short in either of these ways? The charge of deception is leveled by Mackie in his book on Locke. Science has left us with only so many candidates for the role of color, Mackie thinks, and the colors as presented by color experience are not among them. Since there are no properties that *are* as the colors *look*, color experience tells a false story. How the story goes can be gleaned from Mackie's remark that

> it is most improbable that there is any single quality, an objective 'resemblance' of, say, my sensation of a particular shade of green, in all the things . . . that give me this sensation.[50]

FN:50

What forest green *looks* to be, then, is

(a) the common cause of our experiences of forest green, and
(b) an objective resemblance of those experiences.

But if this is the story experience tells, it is not obvious why it should be thought false. The only argument offered against (a), that science does not *postulate* a common cause, applies equally to our experiences of things as jagged, loopy,

[49] Boghossian and Velleman 1991, 67.

[50] Mackie 1976, 36.

FN:51 heaped, frizzy, or tangled;[51] science does not postulate common causes here either. The worst that follows is that scientifically speaking, color is in the same boat as these. (And why should the naive objectivist not agree?) How to interpret "resemblance" in (b) is famously unclear, but suppose we take it Mackie's way: "an objective quality resembles the idea of that quality [iff] in this respect FN:52 things are just as they look."[52] Then to say that reality contains no property "resembling" my experience of forest green is to say that it contains no property that is as forest green looks. This is not an argument for the deception thesis but just a restatement of it.

What about the charge that, while color experience may not lie, it leaves out important parts of the truth? That this should be raised against objectivism is ironic. Dispositionalism and physicalism would seem far more natural targets. Not FN:53 even Locke thought that red *looked* like a power to produce experiences,[53] and still less does it look microstructural in nature. But what is the argument that redness does not look to be what the naive view says it is: an intrinsic nondispositional *sui generis* color property? This would seem to be *exactly* how it looks.

Not so fast, you might say. Even if naive objectivism does not make the colors microphysical, they do come out *determinables* of microproperties. And redness does not *look* like a determinable of microproperties. This invites the question of how determinables of microproperties may be *expected* to look. It is true that the experience of a color does not suggest the myriad ways in which that color is liable to be microphysically implemented. But if a property's liability to be implemented in thus and such ways is the kind of thing a revealing experience of the property ought to register, then *primary* quality experience is unrevealing too. For roundness, no less than redness, is implementable in myriad microphysical ways (having outermost molecules arranged like *so*) of which the experience of roundness gives no hint. Yet roundness is the paradigm of a *revealed* property, the kind that redness was supposed to be contrasted with.

## X

So far we have found no basis for the complaint that color experience is unrevealing. Neither, though, have we attached much content to the contrary claim: that the experience of a color gets the color *right*.

[51] "Sesquiary" qualities these might be called. Thanks to David Hills for making me see their relevance.

[52] Mackie 1976, 49.

[53] Although compare McDowell: "What would one expect it to be like to experience something's being such as to look red, if not to experience the thing in question (in the right circumstances) as looking, precisely, red?" (1985, 112). What bothers me here is the substitution of "experience a thing *as looking red*" for "experience it as *red*". If these are different, then the unrevealingness charge—which concerned the experience of things *as red*—goes unanswered. But to assume their identity is to prejudge one of the main questions before of us: namely, does our experience of redness represent it as ontologically visual?

What standard experiences of color *do* seem to suggest is that redness (e.g.) is intrinsic and categorical. But this much is true of lots of properties—roundness, for instance. To be revealing, shouldn't our experience of a color inform us of features *peculiar* to it? There is even the view that it should lay the color's nature completely bare. Yet as expectations rise about our experience's power to reveal, so too do doubts about the intelligibility of the corresponding feature. Can we really make sense of an *objective* property that is, in Evans's phrase, "exactly as we experience redness to be"?[54]

Here is a way of making the worry sharper. Redness and greenness are fundamentally different, perhaps fundamentally opposed. If color experience is revealing, this ought to be reflected somehow in our different experiences of them. But the only relevant experiential difference would seem to be in qualitative feel. And what can qualitative feel possibly tell us about the nature of an intrinsic, categorical property of external objects, a property that is "there anyway", regardless of the impression it makes on human observers? No wonder Evans complains, against other-than-dispositional conceptions of redness, that "what one conceives when one conceives that objects which appear red to us are in addition really red . . . is quite opaque".[55]

The challenge is to think what redness could be, that the right way to conceive it is in terms of experiences of such-and-such a qualitative type. Dispositionalists and physicalists have ready replies. If redness were a disposition to produce red-type experiences, then *clearly*, those experiences would be peculiarly apt to redness, and an invaluable guide to its nature. And although red-type experiences would *not* be peculiarly apt if redness were microphysical, neither would we expect them to be; science, not color experience, would be our guide to the nature of redness. It is only on the naive account that a certain type of experience is *needed* for knowledge of what redness is (science isn't going to help) at the same as it is *prohibited* (redness being objective). This seeming paradox is our final topic.

## XI

Near the beginning of "Values and Secondary Qualities", McDowell declares that the colors are not "adequately conceivable except in terms of how their possessors would look".[56] By an "adequate" conception of X, let's take him to mean a conception whereby one knows what X is. Then McDowell's claim is that whoever does not conceive redness in terms of how it makes things look does not know what redness is—or, as we might put it, that it is *epistemically essential* to red to make things look that way. But is this true of red?

With surprising regularity, paranormal perception buffs report the existence of "color-touchers", or individuals capable of detecting color by tactile means.

[54] Evans 1985*a*, 272.

[55] Ibid. 273.

[56] McDowell 1985, 113.

Whether the reports are true doesn't matter;[57] the issue for us is the conceptual one of whether someone who accessed color by touch alone could still be said to conceive them adequately.[58] Take for instance the subjects discussed in "Seeing Color with the Fingers", a story in the June 1964 issue of *Life* on "dermo-optical perception":

> Yellow, they said, felt slippery, soft and lightweight. Blue, while not so slippery as yellow, was smoother and the hand could move more freely over it. Red was sticky and clinging. Green was stickier than red but not so coarse. Indigo was very sticky but harder than red or green. Orange was hard and rough, and inhibited movement . . . Black was very inhibiting and clinging, almost gluey, while white was quite smooth, though coarser than yellow.[59]

How shall we describe these people? They have a *way of thinking* about yellowness—an *idea* of it if you like—but that is all. What yellowness *is* they do not know.[60]

The epistemic essentiality claim seems right, then. But McDowell puts a *construal* on the claim that I want to raise a question about. Properties that are "not adequately conceivable except in terms of certain subjective states", he says, are "subjective themselves in a sense that that characterization defines".[61] The question is whether "that characterization" defines a sense of "subjective" at all. To call a property "subjective" is to comment in an ontological vein about what it is. But to say that it is not adequately *conceived* except (e.g.) in terms of how it makes things look is to applaud certain ways of *thinking* of the property. Unless standards of adequate conception are dictated *by the property and it alone*, no ontological conclusions follow.

Now, it may seem *obvious* that the property sets the standards. All that we mean by an adequate idea of X, recall, is one marking its possessor as knowledgeable about what X is. And surely, the standards for knowing what a thing is flow from *what it is*! To have a name for this view, let's call it *absolutism* about knowing what. *Anti-absolutism* holds that standards for knowing what are sensitive to (so far unspecified) additional factors. Some examples will help us to decide which view is closer to the truth.

[57] They are not true. See Gardner 1966.

[58] "By touch alone" because I propose to ignore the fact that most reputed color-touchers, including those about to be discussed, have had normal color vision.

[59] Rosenfeld 1964, courtesy of David Hills's paranormally good memory. Duplessis 1975 contains a discussion of color-touching among the blind. See Cytowic 1989 and 1993 for the related, and apparently genuine, phenomenon of synesthesia, or cross-modal perception. See also Churchland 1979: ch. 2, for a nicely elaborated fable about temperature-seeing.

[60] The remainder of the paper is greatly indebted to Crimmins 1989. Note that I speak of "knowing what X is" rather than "having the concept of X". These are different. Astrophysicists have the concept of dark matter, but they don't know what satisfies it, that is, what dark matter is. And one needn't have the concept of a Platonic solid to know what the Platonic solids are.

[61] McDowell 1985, 113. Though compare this, from the same paper: "I have written of what property-ascriptions are *understood* to be true in virtue of, rather than what they *are* true in virtue of" (p. 112; my emphasis).

Imagine a person who because of some sort of agnosia is unable to recognize presented squares: not by touch or by sight or in any other way. Asked whether she knows what a square is, we'd be hard put to say she did. On the one hand, she does have an idea of squareness; she may even know that squares are four-sided regular polygons. On the other, here she is with a square in her hands and she can't ascertain its shape!

Now switch to the property of being a milliagon, here defined as a million-sided regular polygon. With respect to milliagons, all of us are in the position of our agnosic friend, the position of not being able to recognize them either by sight or by touch. By parity of reasoning, shouldn't we suspect ourselves of not knowing what a milliagon is? Yet we do not; somehow, to know what a milliagon is, the ability to recognize one is not required. To the absolutist, this can only mean that as *n* decreases, the property of *n*-sided regular polygonhood puts stronger and stronger demands on those would seek to know it. But the truth is surely that since *most* people can recognize squares perceptually but not milliagons, the ability is required in the one case but not the other.[62]

Listening to the summer weather report, you may hear, in addition to tomorrow's *temperature*, the expected *heat index*—a function of temperature and humidity that is supposed to predict how hot the day will *feel*. ~~Poor~~ Henry ignores this figure; for him, every day is a 100% humidity day, because Henry is unable to perspire. The humidity normally being quite a bit less than 100%, Henry spends much of the summer feeling considerably hotter than we do; on a day when we might be out enjoying the breeze, Henry will be huddled next to the air conditioner.

Of course, Henry appreciates, in a sense, that it's merely warm out there, not hot. Even so, given how he feels on such days, one wants to say: poor guy, he doesn't know what it is for a day to be (merely) warm. How will the absolutist explain this? She must say that Henry's idea is objectively wrong; the *right and true* way to feel when it's warm out is the way that *we* feel. This is wildly implausible, however. Perspiration functions to drop our skin temperature *below* that of the surrounding air, so if anyone is appreciating the temperature "as it really is", it's Henry.[63] (Except for the historical accident of our greater

[62] Consider in this connection Oliver Sacks's "twins": "A box of matches on their table fell, and discharged its contents on the floor. '111' they both cried simultaneously . . . I counted the matches—it took me some time—and there were 111. 'How could you count the matches so quickly?' I asked. 'We didn't count,' they said. 'We see the 111' " (Sacks 1990, 199). Knowing what 111 is would be a different and more demanding thing if more of us had this ability. Mark Crimmins gives a related example: "Some objects reflect light primarily in the infra-red spectrum—outside the band of visible light. Suppose we are given one. This object has a certain color-like property, call it *infra-mauve*, of reflecting such-and-such frequencies of light to such-and-such degrees. We cannot recognize visually when an object is infra-mauve, but clearly we know what property it *is*. And isn't our situation with respect to infra-mauve just like [the sightless person's] with regard to red?" (Crimmins 1989).

[63] This is based on a remark of Kripke's at the 1989 Color Symposium.

numbers, it would have been wetskins like ourselves that were under suspicion of experiencing the temperature incorrectly.)

Absolutists say that standards of adequate conception are dictated by the (nature of the) thing conceived. But if so, then the following clearly possible thing should *not* be possible: knowing what $X_1$ is without knowing what $X_2$ is, even though $X_1$ and $X_2$ are identical.[64] All that most people know about sodium chloride (e.g.) is that it is some sort of chemical; what *salt* is, however, they *do* know. And while few have led such sheltered lives as to be ignorant of what cold is, ignorance of low random kinetic energy abounds.[65] (Note that it could have been the other way around.[66]) Since different things count as knowing what X is depending on how it is picked out, responsibility for the line between adequate ideas of X and inadequate ones does not lie with X alone.

So much is to challenge an influential *argument* for the visual nature of the colors. Now let me raise a doubt about the argument's conclusion. The color-touchers, let's imagine, evolved in isolation from those accessing color in other ways; they found the tactile mode of color perception as natural and inevitable as we do the visual one. Bananas and canaries they called "silft" ("slippery, soft and lightweight"), ripe strawberries were "styngy" ("sticky and clinging"), and so on. To know what silftness and so on *were*, one had to know how they made things *feel*. Only later, when contact was made with the likes of us, was it realized that sliftness, to fix on that example, was none other than yellowness. Whereupon some of the more philosophical color-touchers reasoned as follows: since silftness is tactile by nature, yellowness has a tactile nature as well.

TR silftness

Now, the fact that this conclusion is drawn in a counterfactual world doesn't make it any more tolerable. If yellowness *could* have been tactile by nature, then, given the modal fixity of natures, it *is* tactile by nature; and as we know, it is no such thing. This turns the color-touchers' *modus ponens* into a *modus tollens*; since yellowness is not of a tactile nature, neither is silftness. But of course, two can play this game. By whatever authority *we* are able to vouch for the non-tactile nature of yellowness, the color-touchers affirm that silftness is not of a visual nature. It follows that yellowness isn't visual either.

## XII

Yellowness is *supposed* to be conceived in terms of how it makes things look. But the connections between the way a thing is supposed to be conceived, and the way it is, are complicated. Even the obvious-looking principle that "one is not obliged to conceive X as F unless it *is* F" may be doubted. And the stronger one that substitutes "unless X is F by nature" is definitely mistaken.

[64] And indeed known to be identical. See Hintikka 1962, 149 n. 9, for an evasion of the point.

[65] Using "cold" in the sense of low temperature.

[66] See Crimmins 1989.

But our story has a lacuna exactly here. If the ways we are supposed to think of things need not reflect anything metaphysically important about them, why on earth are we supposed to think of them in these ways?[67] FN:67

Start with a reason going back in essence to Frege. Agreement, testimony, dispute—all of these depend on words being used in reference to the same items. The Frege point is that that is not enough. Take the case where you say, "Aristotle is amazing", speaking of the philosopher, and I hear "Aristotle is amazing", understanding our friend Aristotle Sundog Greenglass. Obviously, ~~we have failed to communicate~~ I have misunderstood you; and this remains so even if, unbeknownst to anyone, the philosopher's private researches into generation and corruption were so successful that the two Aristotles are in fact the same. The point applies on the property side as well; if you say, "Zemly is all charged up", meaning that she is excited, and I hear you as talking about Zemly's electrical condition, then the exchange is not saved if future science reveals that to be excited *is* to be electrically charged.

Along with speaking of the same items, then, communicators aim to think of these items in related ways. (Of course—communication would not be worth the effort unless it had guessable effects on the participants' states of mind.) But how better to arrange for this result than by indexing standards of conception to the pieces of public language that they exchange? So denoted, the Morning Star is to be understood in terms of its morningish appearance, and the Evening Star in terms of how it appears in the evening; the road from Thebes to Athens asks to be envisaged as traveled Athens-ward, the one from Athens to Thebes as traveled the other way. So denoted, salt is a condiment, the Sun is the preeminent celestial body, and sound is something to be heard; sodium chloride, meanwhile, is a chemical, Sol is one star among many, and compression waves are particles in motion.

Now, that it aids communication if different words carry with them different standards of right ideation was *supposed* to be a *non*-metaphysical reason for playing favorites. But any sort of favoritism among ideas stands as a temptation to the metaphysician. What is the relation between the Athens–Thebes road and the Thebes–Athens road? Naively, identity. But when one thinks of these roads in the ways their names prescribe, certain subtle "differences" emerge; one of the roads runs uphill, the other down, one offers better views, and so on. And now Leibniz's Law appears to show that the roads are not identical after all. Or consider the following. Science identifies sound with compression waves—a type of motion. But motion is perceived visually or by touch, while sound is perceived through the ears. So the wave theory of sound is wrong. Both arguments have had their advocates.

[67] What may be the deepest reason I don't feel ready to discuss; we valorize certain styles of conception as part of an ongoing project of grooming ourselves to respond similarly to new cases. (See Pettit 1990 for a congenial account of rule following, especially pp. 16 ff.)

Still harder not to read metaphysical meaning into is a second sort of favoritism among ideas. Imagine that the color-touchers' verdicts agree with ours until chemists devise a substance which, although "yellow to the fingers", is blue to the eye. Who would, or should, win the ensuing argument seems clear. (I assume that the story is not filled out in prejudicial ways.) The visual perspective on color is *privileged*; judgments framed from it are, other things equal, dialectically weightier than those framed from other perspectives.[68] But this privilege surely testifies to some sort of *special rapport* between color and vision. And now it begins to seem that the colors *of their nature* favor vision over the other senses.

Everything here depends on what we make of the bruited special rapport. The first interpretation that comes to mind is simply that vision is a superior *detector* of color; those who look make fewer mistakes than those who touch. But so what if it is? After all, vision is also a superior detector of faces; animals do a remarkable job of recognizing each other by smell; and coastlines are best judged from the air. This hardly suggests that faces are of a visual nature, or that animals and coastlines are olfactory and aerial.[69] The most that follows from a perspective's greater statistical reliability is that if reliability is your goal, that is the perspective to adopt.

Yet such a reply, although correct within its limits, is superficial. This comes out when we press the question of *why* the colors are most reliably accessed by sight. With coastlines, the reasons are clear: due to their great size they are best viewed from afar; because they are (more or less) planar, the ideal viewing angle is from above or below; because air is transparent and rock is opaque, the view from above is better. Facts about coastlines thus explain *why* the aerial perspective should track the truth about them especially closely. But we have cited no facts about the naive colors that would give the visual perspective a truth-tracking advantage over that of the color-touchers. And in fact it is consistent with our fantasy that the color-touchers, unbothered by variation in lighting conditions, make *fewer* mistakes than ourselves. Somehow, though, this does not seem to rob the visual perspective of its dialectical advantage. Were a brute conflict to arise in which neither judgment could be written off to ambient lighting, sensory fatigue, or what have you, the visual perspective would prevail.

[68] Not that this privilege could never be lost—a point urged on me by Philip Kitcher and Paul Churchland. Similar privilege accrues to the first-person perspective on sensations as against the third-person, and to the interpretive perspective on intentional states as against that afforded by physics plus (alleged) bridge laws. Here too metaphysical conclusions have been thought to follow, e.g., by Davidson in his attack on psychophysical identities.

[69] Another form of the argument is this. The visual idea of yellowness is in closer rapport with it than the tactile one, while with silftness it is the other way around; since identicals stand in the same relations, yellowness can't be silftness. But all the supposed difference in relations comes to is that *attributions* of yellowness and silftness respond differently to the same evidence. Of course—they are different attributions! The properties attributed can still be one and the same. They can even be *believed* to be the same, albeit with probability less than one. Admittedly, if visual and tactile evidence were to push in opposite directions, and push very hard, the identity-belief would come under pressure. But that recalcitrant experience *would* force us to distinguish "two things" doesn't force us to distinguish them *now*. (A good thing too, or few identity claims would be left standing.)

So color's special rapport with vision runs deeper than statistics. As deep as metaphysics, perhaps? I think we can explain the added depth in another way.[70] Let the color-touchers' idea track yellowness as accurately as you like, this remains a *de facto* connection which both sides stand ready to sacrifice in the interests of protecting the property's *de jure* connection to the visual idea. Silftness, meanwhile, is signified *de facto* by the visual idea, *de jure* by the tactile one; both sides would surrender the first connection to protect the second.[71] So if, contrary to what we suppose, the properties are distinct, the visual idea goes with yellowness and the tactile one with silftness. There are ways of putting this that lend it a metaphysical air; for instance, "the price of cognitive access to *yellowness* is to think of it visually or else in a way that agrees with the visual idea", or, the logical next step, "yellowness is *empirically* visual even if not noumenally so". But the point is just that an idea cannot denote yellowness without denoting the same property as our yellowness-idea, which is visual. And this is no more than a truism.

## XIII

Remember our paradox: to know what yellowness is, one must know how it makes things look; yet if yellowness is objective, the impression it makes on human observers should not be *relevant* to what it is. The paradoxical answer is that there can be perfectly objective properties such that to know what they are, one has to understand them in subjectivity-involving ways.[72] I take these properties to include, in addition to the colors, qualitative properties like feeling itchy and suffering pain; perceptual properties such as that of being shaped like *so*; and normative properties like rationality and goodness. To illustrate with the last example, it may well be that no one who appreciates what goodness is can remain indifferent to it. But if so, the reason is not that goodness exerts an irresistible magnetic pull; it's that you have to care about goodness to qualify as appreciating what it is.

## REFERENCES

Aristotle (1986). *De Anima* (New York: Penguin).

Armstrong, D. (1968). *A Materialist Theory of the Mind*. London: Routledge.

[70] Compare Wittgenstein's criteria/symptoms distinction in 1953, §354.

[71] So we are using an intensionalized version of Wittgenstein's distinction; different things can be criterial for X and Y even though X *is* Y.

[72] Someone might say that the objectivist has won a hollow victory; for even if an objective property makes F-attributions true, the property's status *as* truth-maker is owing to subjective factors. This is something I am still pondering, but I agree to this extent: objectivist metaphysics is one thing, objective discourse another.

Baxandall, M. (1985). *Patterns of Intention*. New Haven: Yale University Press.

Block, N. (ed.) (1980). *Readings in the Philosophy of Psychology*, i. Cambridge, Mass.: MIT Press.

—— (1990). "Can the Mind Change the World?". In G. Boolos (ed.), *Meaning and Method: Essays in Honor of Hilary Putnam*, Cambridge: Cambridge University Press, 137–70.

Boër, S., and Lycan, W. (1986). *Knowing Who*. Cambridge, Mass.: MIT Press.

Boghossian, P., and Velleman, D. (1989). "Colour as a Secondary Quality". *Mind* 98: 81–103.

—— —— (1991). "Physicalist Theories of Color". *Philosophical Review* 100: 67–106.

Boyd, R. (1988). "How to Be a Moral Realist". In Sayre-McCord (1988), 181–228.

Campbell, J. (1993). "A Simple View of Colour". In Haldane & Wright (1993), 257–68.

Campbell, K. (1969). "Colours". In R. Brown and C. D. Rollins (eds.), *Contemporary Philosophy in Australia*, London: Allen & Unwin, pp. 132–57.

Churchland, P. (1979). *Scientific Realism and the Plasticity of Mind*. Cambridge: Cambridge University Press.

Cottingham, J., Stoothoff, R., and D. Murdoch, (eds.) (1984). *Philosophical Writings of Descartes*. Cambridge: Cambridge University Press.

Crimmins, M. (1989). "Having: Ideas and Having the Concept". *Mind & Language* 4: 280–94.

Curley, E. (1972). "Locke, Boyle, and the Distinction between Primary and Secondary Qualities". *Philosophical Review* 81: 438–64.

Cytowic, R. (1989). *Synesthesia: A Union of the Senses*. New York: Springer-Verlag.

—— (1993). *The Man Who Tasted Shapes*. New York: Putnam.

Dennett, D. (1978). *Brainstorms*. Montgomery: Bradford Books.

—— (1991). *Consciousness Explained*. Boston: Little, Brown & Company.

Duplessis, Y. (1975). *The Paranormal Perception of Color*. New York: Parasychology Foundation.

Evans, G. (1985*a*). "Things Without the Mind". In *Collected Papers*, Oxford: Clarendon Press, 249–90.

—— (1985*b*). *The Varieties of Reference*. Oxford: Oxford University Press.

Field, H. (1972). "Tarski's Theory of Truth". *Journal of Philosophy* 69: 347–75.

● Q1 Gardner, M. (1966). "Dermo-optical Perception: A Peek Down the Nose". *Science* ●vol. 151 pp 654–657 ~~(eds.)~~

Haldane, J., and Wright, C. 1993. *Reality, Representation, and Projection*. Oxford: Oxford University Press.

Hardin, C. (1984). "Are 'Scientific' Objects Coloured?" *Mind* 93: 491–500.

—— (1988). *Color for Philosophers*. Indianapolis: Hackett.

Hintikka, J. (1962). *Knowledge and Belief*. Ithaca, NY: Cornell University Press.

Jackson, F., and Pargetter, R. (1987). "An Objectivist's Guide to Subjectivism about Color". *Internationale Revue de Philosophie*, 127–41.

—— and Pettit, P. (1990). "Program Explanation: A General Perspective". *Analysis* 50: 107–17.

—— Pargetter, R., and Prior, E. (1982). "Functionalism and Type–Type Identity vol. 41 Theories". *Philosophical Studies* 42: 209–25.

Johnston, M. (1989). "Dispositional Theories of Value". *Proceedings of the Aristotelian Society*, supp. vol. 63: 139–74.

——(1992). "How to Speak of the Colors". *Philosophical Studies* 68: 221–63.

——(1993). "Objectivity Refigured". In Haldane and Wright (1993), pp. 85–130.

Kripke, S. (1980). *Naming and Necessity*. Cambridge, Mass.: Harvard University Press.

Lewis, D. (1966). "An Argument for the Identity Theory". *Journal of Philosophy* 63: 17–25; repr. in Lewis (1983) pp. 99–107.

——(1980). "Review of Putnam". In Block (1980), pp. 232–3.

——(1983). *Philosophical Papers*, i. Oxford: Oxford University Press.

——(1986). *Philosophical Papers*, ii. Oxford: Oxford University Press.

——(1989). "Dispositional Theories of Value". *Proceedings of the Aristotelian Society*, supp. vol. 63: 113–37.

Mackie, J. (1976). *Problems from Locke*. Oxford: Clarendon Press.

——(1977). *Ethics: Inventing Right and Wrong*. New York: Penguin.

Malcolm, N. (1968). "The Conceivability of Mechanism". *Philosophical Review* 87: 45–72.

McDowell, J. (1983). "Aesthetic Value, Objectivity and the Fabric of the World". In E. Schaper (ed.), *Pleasure, Preference, and Value*, Cambridge: Cambridge University Press, pp. 1–16.

——(1985). "Values and Secondary Qualities". In T. Honderich (ed.), *Morality and Objectivity*, London: Routledge & Kegan Paul; ●pp. 110–129 repr. in Sayre-McCord (1988), 166–180 as reprinted. ● Q2

McGinn, C. (1983). *The Subjective View*. New Yo: Oxford University Press.

Millikan, R. (1994). "On Unclear and Indistinct Ideas". *Philosophical Perspectives* 8: 75–100.

● Q3 Nassau, K. (1980). "The Causes of Color". *Scientific American*, ●Vol. 243 1980 pp. 124–154.

Pargetter, R., and Prior, E. (1982). "The Dispositional and the Categorical". *Pacific Philosophical Quarterly* 63: 366–70.

Peacocke, C. (1989). "Perceptual Content". In J. Almog, J. Perry, and H. Wettstein (eds.), *Themes from Kaplan*, New York: Oxford University Press, pp. 297–329.

Pettit, P. (1990). "The Reality of Rule-Following". *Mind* 99: 1–21.

——(1991). "Realism and Response-Dependence". *Mind* 100: 587–626.

Prior, E. (1985). *Dispositions*. Aberdeen: Aberdeen University Press.

——Pargetter, R., and Jackson, F. (1982). "Three Theses about Dispositions". *American Philosophical Quarterly* 9: 251–7.

Putnam, H. (1983). *Realism and Reason*. Cambridge: Cambridge University Press.

Railton, P. (1986). "Moral Realism". *Philosophical Review* 95: 163–207.

Reid, T. (1969). *Essays on the Intellectual Powers of Man*. Cambridge, Mass.: MIT Press.

Rosenfeld, A. (1964). "Seeing Color with the Fingers". *Life*, 102–13.

Russell, B. (1912). *Problems of Philosophy*. Oxford: Oxford University Press.

Sacks, O. (1990). *The Man who Mistook his Wife for a Hat*. New York: Harper.

Sayre-McCord, G., (ed.) (1988). *Essays on Moral Realism*. Ithaca, NY: Cornell University Press.

Shoemaker, S, (1986). "Review of McGinn, *The Subjective View*". *Journal of Philosophy* 83: 407–13.

——(1990). "Qualities and Qualia: What's in the Mind?". *Philosophy and Phenomenological Research* 1 (supplement): 109–31.

Shope, R. (1978). "The Conditional Fallacy". *Journal of Philosophy* 75: 397–413.

Smith, M. (1989). “Dispositional Theories of Value”. *Proceedings of the Aristotelian Society*, supp. vol. 63: 89–111.
Wittgenstein, L. (1953). *Philosophical Investigations*. Oxford: Blackwell.
Wright, C. (1992). *Truth & Objectivity*. Cambridge, Mass.: Harvard University Press.
Yablo, S. (1992). “Mental Causation”. *Philosophical Review* 101: 245–80; Ch. 8 above.

## Queries in Chapter 9

Q1. Author correction is not clear. correct info, except for (eds.) -- I can't speak to the formatting

Q2. Author correction is not clear. info is correct -- I can't speak to the formatting

Q3. Author correction is not clear. info is correct -- I can't speak to the formatting

# 10

# Wide Causation

## 1. INTRODUCTION

Are physical events subject to mental influence? Even to raise the question suggests what the answer had better be. Deny mental causation and you are denying that anyone ever *does* anything: answer a question or anything else.[1] Tongues may wag and arms may wave about, but there is no action unless these things occur at the bidding of appropriate mental states. Nor is action the only casualty if mental states are physically inert. Smirking, beaming, moping about, shivering in anticipation, raising a skeptical eyebrow, favoring a tender limb—these are just an inkling of the human phenomena making no sense in a world where thoughts and feelings keep causally to themselves.

Of course, to say that mental states had *better* be physically influential does not begin to explain how such a thing is possible. And the fact is that bafflement about the *how* of mental causation has been growing, to the point that doubts are now creeping in about the *whether*. A good many philosophers seem ready to give in to these doubts and accede to some form of epiphenomenalism: here, the view that mental phenomena exert no causal influence over the course of physical events. A good many others "resist" epiphenomenalism by maneuvers so subtle that it is mainly on their own ~~impassioned~~ testimony that they are not counted into the first camp. Still other philosophers would junk mental phenomena altogether rather than see them causally enfeebled.

This paper follows up on "Mental Causation" (Yablo 1992*a*; ch. 8 above); it takes back the suggestion there that "wide content" events may be inefficacious since "their 'narrow' counterparts seem ordinarily to be more commensurate . . . with their supposed effects". I owe thanks to Louise Antony, Mark Crimmins, Frank Jackson, Joe Levine, Philip Pettit, and Sydney Shoemaker for written comments, Georges Rey for an eye-opening conversation, and especially Krista Lawlor for discussions in 1993 by which time she had hit independently on the same basic idea. Too late I discovered Jerry Fodor's recantation in *The Elm and the Expert* (1994). Sections 7 to 19 owe a lot to Jackson and Pettit 1988. Section 20 however owes a lot to his pp. 40 ff. Research was done at the National Humanities Center with support from a University of Michigan Humanities Fellowship.

[1] I use "mental causation" for mental causation of physical effects.

transpose circled sentence with sentence before it.

All of this adds up to what has been described as an outbreak of *epiphobia*. (Epiphobia $=_{\mathrm{df}}$ the fear that one is turning into an epiphenomenalist.[2]) Even allowing for the strange logic of thought disorders, it has to be said that this one is asserting itself at ~~rather~~ a surprising historical moment. Epiphenomenalism was supposed to be somebody *else's* problem: somebody long dead, or at any rate hopelessly out of touch with recent materialist developments like multiple realization and supervenience. Why epiphobia now?

## 2. A STORY

Time was when epiphobics had a genuine threat to point to: the gaping divide dualists had postulated between the mental realm, said to be lacking in kinematical qualities, and the thoroughly kinematical physical realm. Not even Descartes claimed to understand how causal relations were supposed to reach across this divide,[3] and his critics (notably Gassendi) found the notion positively incoherent:

> you must explain to us how this "directing" of movement can occur without some effort—and therefore motion—on your part. How can there be effort directed against anything, or motion set up in it, unless there is mutual contact between what moves and what is moved?[4]

Even at the time, however, such worries were ~~easily~~ shrugged off by philosophers who, while agreeing with Descartes that the mind was something apart, had their own ideas about its particular nature. (Gassendi is a case in point: "I will grant you [that you are really distinct from your body], but will not therefore grant that you are incorporeal . . .".[5]) Centuries of subsequent squabbling about the intelligibility of cross-category interaction never quite succeeded in breaking this stalemate. No argument from the ~~gaping~~ ontic divide between mind and body could get a grip, simply because no one felt sure of what the divide's mental side looked like.

Then the brainstorm hit that the mind's precise characteristics might not *matter*; trouble for mental causation could be conjured out of physical assumptions alone.[6] For the issue is not ~~Never mind~~ whether mental causation is beyond *understanding* (that depends on the natures of the relata, hence in particular on the nature of

[2] Fodor 1989. [3] See Bedau 1986. [4] CSM II, 237.

[5] CSM II, 237. Cf. Lucretius: this "reasoning proves the nature of the mind and spirit to be corporeal. For when it is seen to hurl the limbs forward, to snatch the body out of sleep, to alter the face, and to govern and steer the entire man—and we see that none of these is possible without touch, nor touch without body—you must surely admit that the mind and spirit are constituted with a corporeal nature" (Long and Sedley 1987, 67).

[6] Along, of course, with the assumption that mental phenomena aren't physical. This recalls another crucial stimulus to contemporary epiphobia, the Putnam/Fodor multiple realization argument. See their papers in Block 1980.

mind) it is enough for the epiphenomenalist if it is beyond *belief.* And that mental causation is beyond belief can be maintained just on the strength of the physical realm's well-attested autonomy and self-sufficiency. A strategy like this was employed by C. D. Broad in his "argument from energy":

> I will to move my arm, and it moves. If the volition has anything to do with causing the movement we might expect energy to flow from my mind to my body. Thus the energy of my body ought to receive a measurable increase, not accounted for by the food that I eat and the oxygen that I breathe. But no such physically unaccountable increases of bodily energy are found,[7]

and his "argument from the structure of the nervous system":

> . . . the nervous processes involved in deliberate action do not differ in kind from those involved in reflex action; they differ only in degree of complexity. . . . So it is unreasonable to suppose that the mind has any more to do with causing deliberate actions than it has to do with causing reflex actions.[8]

But it was Norman Malcolm in "The Conceivability of Mechanism" (1968) who first grasped the new genre's full potential. Assuming a physical theory rich enough to "provide *sufficient* causal explanations of behavior",

> the movements of the man on the ladder would be *completely* accounted for in terms of electrical, chemical, and mechanical processes in his body. This would surely imply that his desire or intention to retrieve his hat had nothing to do with his movement up the ladder.[9]

The most important stimulus to contemporary epiphobia is this argument of Malcolm's. Because it sees would-be mental causes as preempted by underlying physical states, we can call it the *argument from below*. Later it will be set out in more detail, but the essence is simply this: with each physical effect causally guaranteed by its physical antecedents, what is there left for its mental antecedents to do?

Although the argument as stated targets mental causes, the underlying logic applies to *all* nonphysical states. If an effect is causally inevitable given preexisting physical conditions, then the effect's biological, geological, economic, etc. antecedents are just as much out of a job as its mental ones. Whether because of concern about the sweepingness of this result or for some other reason, attention has been shifting to a second and in some ways more discriminating argument, the *argument from within*.

The target this time is *intentional* mental states: states like belief and desire individuated in terms of truth or satisfaction conditions. If Putnam is right that truth conditions can vary between internally indiscernible agents (e.g., me and my doppelganger on Twin Earth), then intentional states are *extrinsic*, or not

[7] Broad 1925, 104. I should stress that he is not impressed by either argument.

[8] Broad 1925, 100.

[9] Watson 1982, p. 133.

FN:10 wholly a matter of what goes on within the thinker's skin.[10] Add to this that it is intrinsic states that determine causal powers—

> you can change [extrinsic states], remove them, or imagine them to be different in various respects, without ever changing the causal powers of the object or person that is in this extrinsic condition—

and you see the problem:

> how can extrinsic facts about *A*, depending as they do on factors that are spatially and temporally remote from *A*, help explain *A*'s current behavior? Surely what explains, causally explains, *A*'s raising her arm or pushing a button are intrinsic facts about *A*.[11] FN:11

Any behavior that beliefs and desires might *seem* to generate must really be due to some intrinsic surrogate: syntactic states, perhaps, or narrowly contentful attitude-analogues, or even brain states.[12] FN:12 Intentional causes are thus displaced by factors internal to the agent, which gives the argument its name.

## 3. THE CONNECTION

The key point for us is that mental causation faces two separate threats: "from below" and "from within". With so much effort gone into distinguishing these threats in recent years,[13] FN:13 no one seems to have noticed that they are *connected*, and in a way that makes them more formidable as a package than taken separately. Any decent response to BELOW (as I'll call it) will have to make use of the principle of

*proportionality*: causes must be proportional to their effects.

But, WITHIN can claim to be little more than an application of the very same principle! If that is right, then we are damned if we do (accept proportionality) and damned if we don't; either way, one of the two arguments goes through. How proportionality is supposed to play this double role is the topic of the next few sections, but in general terms the idea is this.

Start with the physical states that BELOW casts as preempters. Seen in the light of proportionality, these appear to be overloaded with unneeded microstructural

[10] Putnam 1975. Despite our perfect intrinsic similarity, my doppelganger on Twin Earth wants *twater*, the colorless drinkable stuff in his environ~~s~~ while it is *water* that I desire. ment,

[11] Dretske 1993, 187, with inessential deletions.

[12] See various papers in Woodfield 1982, especially McGinn's; Loar 1985; Stich 1978, 1980, 1983; Fodor 1980*a, b*, 1987, 1991*a*, 1991*c*; Dretske 1988 and 1993; and various papers in McLaughlin 1991, especially those by Kim and Horgan. Here is Kim's version of the argument: "semantical properties [are] relational, or extrinsic, whereas we expect causative properties involved in behavior production to be nonrelational, intrinsic properties of the organism. If inner states are implicated in behavior causation, it seems that all the causal work is done by their "syntactic" properties, leaving their semantic properties causally idle. . . . *How can extrinsic, relational properties be causally efficacious in behavior production*?" (1991, 55).

[13] See in particular Kim 1991, Horgan 1991.

detail. (Had my pain been implemented in a different microphysical way, the effect would in all likelihood still have occurred.) As for the mental states they are said to preempt, these are simply the result of stripping some of the unneeded detail away. But then to call the mental state an unneeded excrescence gets matters exactly backwards. You might as well say that since my screaming "wake up right now!!" in my cat's ear sufficed to wake him, my screaming in his ear as such made no contribution; it was only along for the ride.

All right so far. But now WITHIN chimes in that intentional "causes" are *also* overloaded with unneeded detail, not microstructural this time but *extrinsic*. Regardless of whether my desire had been for water or twater, as long as I stayed intrinsically the same, the behavioral results would not have been any different. The question is, why should this excess extrinsic detail be any less offensive to proportionality than the unneeded microstructural detail of the last paragraph?

This is a question I hope to answer. After various preliminaries and softening up exercises in the next few sections, the argument will unfold in three stages. First, WITHIN is *not* an application of the same principle used to defeat BELOW. Second, BELOW falls to a principle that is *tolerant* of intentional causation, albeit intentional causation of an interestingly unexpected sort.[14] Third, WITHIN relies on an enormously *stronger* principle that undermines just about any intuitive causal relation you care to mention. Details will be given in due course; for now proportionality is left at an intuitive level so as to give WITHIN the best possible run for its money.

## 4. THE ARGUMENT FROM BELOW

According to *closure*, each physical outcome $E$ is causally guaranteed by some prior physical $C$.[15] *Dualism* says that no mental $C^*$ is identical to any physical $C$. *Exclusion* says that if $E$ is causally guaranteed by $C$, then no $C^*$ distinct from $C$ is causally relevant to $E$. These three assumptions granted, no physical effect owes anything to its mental antecedents. How can it, with underlying physical states already ensuring that the effect is going to occur?[16]

All of the assumptions could be quarreled with, but closure and dualism can be considered the price of admission to the debate. Someone who denies dualism

[14] There is more on BELOW in Yablo 1992*a* (ch. 8 above); there it is called the exclusion argument.

[15] Or at least sufficient for $E$'s objective probability.

[16] $E$ in this paper is always a token event. But $C$, $C^*$, and so on can be either tokens or types. Words like "state", "phenomenon", "antecedent", and "event" are meant to share in this ambiguity and as far as grammar permits they will be used indifferently either way. (Although see note 46.) I appreciate that some people are scandalized by type/token laxity and I apologize to each and every one of them. The alternative was to run essentially the same argument twice over. See Yablo 1992*a*, *b* for a more careful treatment.

(e.g.) thinks that mental states *are* physical states and so is not interested in any supposed threat from below. It makes sense then to focus on exclusion, which has in any case an obvious problem. Look again at what is being claimed:

> for every "form of nonidentity" $R$ (every irreflexive relation) and every $R$-related pair $C$ and $C^*$, if $C$ is causally sufficient for an effect, then $C^*$ is causally irrelevant to it.

No doubt there are *some* irreflexive relations $R$ whose relata do compete for causal influence as the principle says. But for many $R$s this competition arises only sometimes, and for others it never arises. $R =$ causation is a case in point; taken at its word, the exclusion principle predicts that $E$ owes nothing to the causal intermediaries by which $C$ brings $E$ about! This shows that the exclusion principle is overdrawn. But is it overdrawn in a way that bears on the causal relevance of my pain? How plausible is it really that my pain serves as a causal intermediary between its physical basis $C$ and my grimace?

Never mind that this would require my pain to be literally an *effect* of $C$, whereas pain intuitively stands in a *closer* than causal relation to its physical basis.[17] The relation pain bears to $C$ *is*, as the word "basis" attests, often thought to be causal-*like*; it is considered a dependency relation of some sort.[18] And that ought to be just as good. The real difficulty is still to come. Much as we might like the idea of our thoughts and feelings functioning as intermediaries, how exactly are they supposed to be slotted in? If there were gaps in the physical event-chains linking brain states to behaviors, then (who knows?) mental states might perhaps find work plugging them. This would violate the exclusion principle but only in the way that intermediaries do generally. To foist my pain on a process that is complete and self-sufficient without it, though, goes against what seems *right* in exclusion: a thing can do causal work only when causal work is there to be done.

[17] John Searle says that mental states are "caused by and realized in" physical states of the brain (1983, ch. 10). At times he even seems to suggest that they are caused by and *identical to* brain states:

> if brain processes cause consciousness, then it seems to many people that there must be two different things, brain processes as causes, and conscious states as effects, and this seems to imply dualism. This . . . mistake derives from a flawed conception of causation. (Searle 1995, 60)

Passages like this notwithstanding, Searle *agrees* that there are "two different things": "the sheer qualitative feel of pain is a very different feature of the brain from the pattern of neuron firings that cause the pain" (ibid. 63). His view is thus type dualism; mental types are caused by, realized in, *and distinct from*, physical ones. Searle sometimes presents (this version of) type dualism as a solution to the mental causation problem; for many people it is where the problem starts.

[18] Part of the reason that supervenience theories of mind met with such a euphoric response was supervenience's claim to

> belong to that class of relations, including causation, that . . . represent ways in which objects, properties, facts, events, and the like enter into *dependency* relationships with one another. (Kim 1993, 54)

Hence the disillusionment when it sunk in that the standard covariational definitions of supervenience failed to capture any such dependency, and the subsequent insistence that "any physicalist who believes in the reality of the mental must accept pervasive psychophysical property covariance . . . *plus* the claim that a dependency relation underlies this covariance" (Kim 1993, 169).

## 5. DEPENDENCE

A lot of people seem to think that the best way of getting mental states in on the causal act is to make them strongly enough dependent on physical states. (Supervenient dependence is particularly recommended in this regard.) But a dependent is ontologically *posterior* to what it depends on, and so all the dependency hypothesis achieves is to cast my pain as a lagging indicator of the fact that a process causally sufficient for the effect is already under way. T. H. Huxley saw this implication already in the last century and did not flinch from it:

> all states of consciousness . . . are immediately caused by molecular changes of the brain-substance . . . our mental conditions are simply the symbols in consciousness of the changes which take place automatically in the organism . . . the feeling we call volition is not the cause of a voluntary act, but the symbol of that state of the brain which is the immediate cause of that act.[19]

Those of us who do flinch from Huxley's conclusion have our work cut out for us. If mental states do not *depend* on "molecular changes of the brain-substance", how *are* they connected to brain activity? What alternative picture of mental/physical relations is available?

Of course, one clear alternative is mapped out by the identity theory. There is no question on this theory of pain's *depending* on (brain state) $C$, for $C$ is *already* a state of pain. Another thing there is no question of on this theory is $C$'s beating pain to the causal punch. It is only in the matter of truth value that the theory disappoints. Identicals necessitate one another, but any state specific enough to necessitate pain (a condition we assume $C$ to meet) is *too* specific to be necessitated by it in return. $C$ is thus one of a number of brain states $C_i$ each necessitating pain asymmetrically.[20]

No surprises so far. The surprise is that an essentially similar picture, in which (certain) brain states are *already* states of pain, continues to be available even if the identity theory is rejected. An analogy shows how this can be. Just as pain is not identical to any of the brain states $C_i$ that necessitate it, red is not identical to any of the more precise shades $R_i$ (scarlet, crimson, etc.) that necessitate *it*. Yet there is no question of redness *depending* on scarlet, for to be scarlet is *already* to be red. Scarlet is, as we say, a *way* of being red, or, in an older terminology, a *determinate* of redness. Why shouldn't the $C_i$s likewise be determinates of pain?[21]

[19] "Animal Automatism" in Huxley 1911, 244; the essay dates from 1874.

[20] I have run this as an argument against type identity, but it is effective against token identity as well; see Yablo 1992*a*; ch. 8 above. Kripke in *Naming and Necessity* takes a similar position.

[21] Admittedly, the pain/$C_i$. red/scarlet analogy isn't perfect. This doesn't concern me, *unless* the disanalogies are such as to make pain more causally competitive with $C_i$ than colors are with their shades. As far as I can see, all that "$Y$ is a determinate of $X$" *needs* to mean in this paper is that $Y$

At last we have hit on a relation that brain states plausibly bear to mental ones,[22] *and* that makes nonsense of the causal competition idea. Imagine a pigeon Sophie trained to peck at red shapes. No one would call the triangle's *redness* irrelevant to her pecking on the grounds that the effect was already provided for by its specific *shade* of red.[23] Nor would anyone think that my screaming as such was irrelevant since my screaming "wake up!!" was sufficient. Examples like these confirm what seems obvious anyway: determinates do not preempt their determinables.[24] Understand pain as a determinable of the $C_i$s, and preemption should not be possible in this case either.

## 6. DETERMINABLES AND CAUSATION

The argument from below rests everything on a certain principle: a sufficient cause drains whatever it bears $R$ to of causal relevance. But the principle is not true when $R =$ the determinate/determinable relation. Since this is a relation in which physical and mental states plausibly stand, my pain *can* (for all anyway that BELOW has to say about it) be relevant to effects for which my brain state suffices.

To stop here, though, leaves the impression of a power-sharing arrangement between pain and brain state—an arrangement, indeed, favoring the brain state, since it, after all, *suffices* for an effect to which the pain claims only some unspecified relevance.

One could try to counter this impression by enlarging on what has already been said, viz. that to be in pain is *part of what it is* to be in such-and-such a brain state. When one state is included in another, any influence that the first has on subsequent events is included in the influence had by the second. Brain state and pain thus share power in a more literal sense than the one intended: not by dividing it up between themselves, in the way that books share space on

necessitates $X$ (not because it has a metaphysically infallible way of bringing $X$ about but) because $X$ is immanent in or included in $Y$. This is all it takes to kill the appearance of causal competition. To illustrate with a deliberately far-fetched example, suppose that physical states turned out to be *conjunctions* with mental states as conjuncts. Conjunctions are not in any traditional sense determinates of their conjuncts, but so what? They do determine them in the sense just explained, and that is enough; *P & Q* can no more preempt *P* than scarlet can preempt redness.

[22] The determinate/determinable story is meant to apply to tokens as well as types; it is not just pain as such but the particular pain I am experiencing right now that can be had in a number of physical ways. Pain stands to its physical determinates in the relation that red bears to scarlet; my particular pain stands to *its* physical determinates in the relation that something's *turning* red bears to its turning scarlet. See Yablo 1992*a*, (ch. 8 above), section 6, for more on token determinates and determinables.

caps

[23] ~~Not least because,~~ for all that has been said so far, Sophie is shade-blind and can't tell crimson from any other sort of redness.

[24] I am not saying that redness *inherits* causal relevance from scarlet; I am just denying that scarlet can *deprive* redness of causal relevance.

a shelf with other books, but by possessing it in common, in the way that an encyclopedia shares ~~shelf~~ space with the volumes making it up.

And yet built into this account of how the two states *share* power appears to be a concession that the brain state has *more* power. (Just as the encyclopedia fills more space.) This greater power shows up quantitatively in the fact that my brain state bears the most powerful form of causation—causal sufficiency—to *more effects* than my pain. And it shows up qualitatively in the fact that each of these extra events (e.g., say, my grimace) is *more the effect* of my brain state than of my pain. Because again, it is the brain state that stands to the grimace in the most powerful form of the causal relation there is.

I say that two distinct notions of "effectiveness" are being run together here, in a way we need the principle of proportionality to help us sort out. There is no denying that my brain state has the quantitative advantage mentioned. But sufficing for *more* effects is one thing, greater license to claim them as *your* effects, another. And proportionality says that my brain state may well be in a *worse* position to cause some of these additional effects than is my pain.

How we confused ourselves was by thinking of sufficiency and relevance as unequally powerful forms of causation, when in truth they are not forms of causation at all. *X* can be *relevant* to *Y* despite omitting factors crucially important to *Y*'s occurrence (my addressing the cat was relevant to its waking) and *sufficient* for *Y* despite incorporating any number of irrelevant extras (its waking was causally guaranteed by my shrieking in its pointier ear at a prime number of decibels a message with the semantic content that it should immediately wake up). But *X* does not *cause Y* unless it is *proportional* to it, in a sense that ~~at least~~ FN:25 implies some degree of freedom from these excesses.[25]

If causation is subject to a proportionality constraint, what does that say about my brain state's claim to be more the cause of my grimace than its mental competitor? Arguably it is the brain state, weighed down with superfluous microphysical detail, that suffers in the comparison. After all, I would still have grimaced even if my pain had occurred in a different microphysical way. Whereas the issue of how I would have behaved had the brain state occurred in the pain's absence cannot even be raised, because the brain state includes the pain.

## 7. THE ARGUMENT FROM WITHIN

This is where WITHIN sees its opening. I desire water and extend my hand. But of course Twin Me, who desires not water but twater, would have done the *same* FN:26 in my circumstances[26]—as indeed would *anyone* intrinsically just like me, even

[25] Details are given in section 15.

[26] There is a considerable tradition of attempting to answer WITHIN by denying this sameness; my Twin, unlike me, would have been reaching for *twater*. (See the first few papers in Pettit and

a Swampelganger Me with no intentional states whatever. So intentional states, like brain states, are overloaded with unneeded detail. The only difference is that this time the unneeded detail is "without" rather than "below".

If beliefs, desires, and the like do not cause behavior, what does? The only remaining candidates would seem to be intrinsic states of some sort: syntactical in nature, or neural, or narrow analogues of the attitudes. But we know from the Twin Earth examples that states like these do not *of themselves* represent the world as being any particular way. (What they *can* perhaps claim is association with a staggering array of different truth conditions, which depending on the causal/historical context in which they are imagined to be embedded; see section 11. But context aside, the intrinsic counterpart of my belief that water is wet no more concerns $H_2O$ than XYZ, and no more these than a pattern of electrical signals emanating from the walls of some brain-ready vat.) And now we see the real threat posed by WITHIN: the part of our mental life with the strongest intuitive claim to influence behavior—the part *representing* the circumstances which that behavior seeks to change, and the outcomes it seeks to bring about—may have to take a back seat to states with limited or nonexistent representational powers.

## 8. A NOMIC ANALOGUE

Notice a way in which this reasoning stops curiously short. Fixated as we become on the causally excessive aspects of *intentional* states, and determined to find relief in intrinsic surrogates, it never occurs to us to ask whether the intrinsic surrogates might not be excessive in their own way.[27] I want to sneak up on this question by switching temporarily (until section 13) to a *nomic* version of the argument.

Imagine that we are asked to find the cause of someone's receiving a speeding ticket near a police radar unit; in a familiar jargon, we are asked to solve "*X* caused her to be ticketed" for *X*. Bearing in mind proportionality's call for an *X* that is *enough* for the effect without being *too much*, we quickly see that there are two opposite ways of bungling the task, illustrated by

(1) her driving through the radar caused her to be ticketed, and
(2) her speeding through the radar sober caused her to be ticketed

respectively. Her driving through the radar was not enough, since she had to be driving over the speed limit, while her speeding through the radar sober was too

McDowell 1986, and for criticism Fodor 1991*c*.) I agree that there is *something* my Twin does that is different from what I do, and vice versa. But I would hate to pin the case against WITHIN on this, for there is something *else* we do, viz. simply reaching out, that is the same in both cases. I want to argue that WITHIN is wrong even about the behaviors that my Twin and I have in common.

[27] Yablo 1992*a* (ch. 8 above) is strangely complacent about this; see the "first remark" in that paper's note 51.

much, since her sobriety had nothing to do with it. The true cause will be an event that lies somewhere between the two, presumably her speeding through the radar *per se*.

Now let the task be to solve "*X*s ceteris paribus conduct electricity" for *X*—to find a nomic ancestor rather than a causal one. Again there seem to be two roughly opposite ways of going wrong. This time let our examples be

(3) matter c.p. conducts electricity, and
(4) pennies c.p. conduct electricity.

Because lots of matter *doesn't* conduct electricity, including some paradigmatic enough not to be scared off by the ceteris paribus clause, (3) has an *under*specific antecedent, making it an *over*generalization. (4) has the opposite problem; copper conducts electricity regardless, so (4) is an *under*generalization with an *over*specific antecedent.

This suggests that laws too observe a kind of proportionality constraint. For it to be a law that *A*s are c.p. *B*s, *A* should be determinate enough to make (other things equal) for *B*, but *that's all*; there should be no piling on of nomically irrelevant detail. Otherwise we run the risk of breaking unitary generalizations up into a large number of pointlessly different variants: "pennies conduct electricity", "copper foil conducts electricity", "the bottoms of Revere Ware pans conduct electricity", and so on.

Isn't nomically irrelevant detail just what we are getting, though, in intentional generalizations like "people who want water c.p. go ahead and drink"? Any behavior issuing from *my* intentional states issues equally from the very different intentional states of my other-worldly Twins. Set against the intrinsic properties they and I share, that it is *water* I want looks like precisely the sort of nomic irrelevancy that proportionality warns against. Ignoring this warning amounts to turning our back on a great mass of unitary causal generalizations: namely, all those entailed by the fact that doppelgangers behave *identically* despite believing and desiring different things.

That was the promised nomic analogue of WITHIN. No one could object to the principle behind it; carving up unitary generalizations is a bad thing. But if it is bad when the generalizations are over Twins, then it is bad whatever they are over. At a minimum, then, the argument is too quick. Nothing can be concluded until we consider what *other* generalizations might be on the chopping block; and whether it was the Twin generalizations that put them there; and which should be sacrificed if we are forced to choose.

## 9. MISSED GENERALIZATIONS

A fact that tends to get lost in all the excitement about our Twins is that *we have no Twins*. Neither here on Earth nor anywhere in darkest space can

molecule-for-molecule duplicates of flesh-and-blood human beings be found. As an immediate consequence, the forgone generalizations of the last paragraph (which say in effect that anyone intrinsically just like SY is in his circumstances going to do just as he does) are generalizations over things *all but one of which fail to exist*. This may not make the generalizations any less true, but neither does
FN:28 it recommend them as crashingly important.[28] Still less does it recommend them as crashingly *more* important than the generalizations we forgo if we insist on intrinsic typing—especially since these latter range over things a great many of which *do* exist.

I am about to drink some water, and something tells me I'm not the only one. Is there anything about the members of this group to set us apart from the
^population general ~~run of other~~s? The tempting reply is that *we* are the ones who *want* some
FN:29 water.[29] Pretending for argument's sake that soda, coffee, whiskey, and the rest are yet to be developed, so that water is the one and only drinkable, we can put substance ^
the relevant law like this:

(5) people who want water c.p. have a drink.

But now notice something important about the world's water-wanters. Once we get beyond their shared extrinsic property of being in a state with waterish satisfaction conditions, they are an exceedingly miscellaneous bunch. Unprincipled disjunctions aside, any intrinsic feature they possess in common is likely to be
FN:30 shared as well by a good many *non*-water-wanters.[30] If we insist on intrinsic typing nevertheless, the unitary generalization (5) breaks up into a jillion variations on the theme of

(6) people intrinsically just like SY c.p. have a drink.

And why should decoupling me from my Twins, who after all don't exist, be thought worse than decoupling me from my drinking buddies, who after all do?

Stop right there, you say—"generalizations entailed by the fact that doppelgangers behave identically" wasn't supposed to mean generalizations *limited* to

[28] If the campaign to purge geology of extrinsic notions has never taken off, the reason is that intrinsically-as-though-sedimentary rocks tend to be, well, sedimentary. No one cares about the counterfactual generality thus gained.

[29] Of course, it matters too that this desire is not outweighed by other desires, that water is available, that its whereabouts are known, and so on; I'll take all that for granted here.

[30] "Oh? Who's to say they don't all have the same sentence of mentalese in their desire box?" I have two responses. First, Fodor has promoted mentalese as providing a non-Fregean explanation of cognitive significance phenomena. This application falls apart if the relation between singular propositional contents and mentalese encodings is not one–many. (Do not say that the relation is one–many just when cognitive significance phenomena force it to be. This suggestion pulls in two directions at once, because a given attitude will engender lots of behaviors only *some* of which care how exactly the attitude is encoded.) Second, the argument was supposed to show that narrow taxonomy was better, because less generalization-killing, than broad. Now it looks as though narrow taxonomy *might* be less generalization-killing than broad; it is if a kind of narrow taxonomy can be made out that kills fewer generalizations. Who could argue with that?

doppelgangers, but rather generalizations *subsuming* doppelgangers; the plea in other words was not on behalf of (6) but something more like

(7) people in intrinsic state *F* c.p. have a drink,

where *F* is some limited shareable *aspect* of SY's total intrinsic state. So, contrary to the last paragraph, the generalizations we forgo by typing intentionally, and incur by typing intrinsically, have *plenty* of real-world instances. Add to this that (7) improves on (5) in extending to these instances' counterfactual doppelgangers, and the verdict is clear.

Some such line of response is the intrinsicalist's best bet. But it overlooks one thing: only (6) can be described as a generalization *entailed* by the fact that FN:31 doppelgangers behave identically in the same circumstances.[31] What we have in (7) is the *form* of a generalization that the intrinsicalist *hopes* for. Generalizations like this may well exist, but only if it proves possible to pare my total intrinsic state down to a part that, while specific enough to make c.p. for drinking, is not *so* specific as to be peculiar to myself. And the fact that doppelgangers behave identically cannot itself decide this issue—at least, not by any argument that we have yet seen.

TR

## 10. BRACKETING

To think of (7) as a *casualty* of intentional typing is premature; all that it represents so far is a lost opportunity. And yet, it is possible to wonder how there ~~can~~ could *fail* to be interesting (7)-type generalizations. Aren't these guaranteed, more or less, by the existence of (5)-type intentional generalizations, together with the fact that how people behave in a given situation depends only on what they are like intrinsically? If it is true, for instance, that

(5) people who want water c.p. have a drink,

then given the irrelevance to this behavior of their purely extrinsic features, it should also be true that

(8) people who [want water] c.p. have a drink—

where [wanting water] is wanting water with its purely extrinsic aspects bracketed away.[32] FN:32 This amounts in fact to a *recipe* for intrinsicalizing intentional

[31] Actually, not even (6) is *entailed* by this fact, since I have intrinsic duplicates in a huge (unlimited, in fact) range of external circumstances. Twin Me may be in circumstances like mine, but he is very much the exception.

[32] This argument is already a bit of a stretch, for a reason hinted at in the last note: water-wanters may well find themselves in circumstances more favorable to drinking behavior than the common run of [water-wanters]. But let me not distract attention from the more serious worry raised in the text.

generalizations like (5). Simply substitute for each offending *attitude* the corresponding *battitude*, that is, its image under the operation of bracketing.[33]

Sounds promising, but why stop there? If the recipe works at all, it gives a way of intrinsicalizing nonintentional generalizations as well: substitute for each offending *G* the corresponding [*G*]. I have heard, for example, that people from large families are by and large gregarious. But gregariousness in a given context depends on intrinsic features alone; a gregarious person's intrinsic duplicates are not going to be taciturn and withdrawn. Apparently, then, there has got to be an *intrinsic* property of [being from a large family]—the intrinsic "core" of being from a large family—that *also* makes c.p. for gregariousness. Again, the poor must share an intrinsic property of [poverty] that accounts for their feelings of not having enough food in their stomachs. And now the fallacy must be plain. The most that follows from the irrelevance of the purely extrinsic is that each water-wanter has *some intrinsic feature or other* that leads c.p. to drinking. The further conclusion that they *share* an intrinsic feature that leads to drinking is just wishful thinking.

## 11. BATTITUDES

How wishful it is can be seen by looking at the two main *theories* of the battitudes. One gives us states that are shared but not sufficiently specific, the other, states that are specific but not shared.

The simpler of the two theories says that you share my [belief that *p*] iff some possible doppelganger of yours believes that *p*.[34] (Similarly for desire and the other attitudes.) Twin Me on Twin Earth [wants water], for instance, since he has a doppelganger, myself, who wants water. Doppelgangers of other terrestrial water-wanters [want water] too, not only on Twin Earth but on all planets, be they actual or hypothetical.

But it is not just doppelgangers elsewhere of terrestrial water-wanters who [want water]. Doppelgangers here of extraterrestrial water-wanters [want water] too. And now it becomes hard to think who does *not* [want water] on this theory. For let Dino be a person wanting essentially any old thing.[35] And let Twin Dino be Dino's doppelganger in a world where the-thing-that-manifests-itself-in-the-way-that-the-object-of-Dino's-desire-*actually*-manifests-itself is *water*. Twin Dino wants water in *that* world, so Dino [wants water] in this one. Battitudes

[33] "Bracketing" makes it sound as though battitudes were stripped-down attitudes. This is true in the case of "thick" attitudes (section 16). But I want to leave the door open to "thin" attitudes too subjectively impoverished for bracketing so understood to yield anything worthwhile.

[34] That is, some possible doppelganger of yours has a belief with the singular propositional content that *p*. See Walker 1990 and Stich 1991. I am indebted in this section to Stalnaker 1989, 1990, 1991 and Brown 1993.

[35] This is sloppy but not, I think, in a way that matters.

as explained by the first theory are thus wildly *under*specific, turning "people [wanting water] c.p. have a drink" into a gross *over*generalization along the lines of "matter conducts electricity".

Why does the theory deliver such coarse-grained results? Reformulate it ~~like so~~ as follows and the reasons jump out: you and I share a battitude iff *there are* worldly contexts, *not necessarily identical*, in which your doppelganger judges the same proposition as mine. Each of the highlighted phrases makes for a separate kind of trouble. "Not necessarily identical" leaves the door open to *tailoring* the two contexts so as to offset bona fide battitudinal differences. Perhaps Dino is a martini fiend who would sooner chew tinfoil than take in a drop of any other liquid, but the fact that he (or rather, his doppelganger) *would* want water in a world where it was water that lay behind martini-appearances suffices to make him a [water-wanter] like me.

All right; we need to drop the "not necessarily identical" and require that the same proposition be judged *in the same context*. That we are free to choose this context at will (due to the existential quantifier "there are") means that a problem remains. Battitudes that are *capable* of latching onto different propositions are absolutely distinct.[36] But a *capability* is not the sort of thing that every context can be relied on to register. Sargon's [longing to visit the Morning Star] was quite a different battitude from his [longing to visit the Evening Star], even if he lived out his days in a setting where their distinctness did not manifest itself in different propositional outcomes desired.

Where are we? Not only should battitudinalizers be compared in the same context, that context should be allowed to vary arbitrarily. Both of these modifications together give us the second main theory of the battitudes.[37] For someone to share my [belief that $p$], their doppelganger in $w$ should believe (not the proposition $p$ that I in fact believe, as on the first theory but) the proposition $p(w)$ that my doppelganger in $w$ believes.[38] And this should hold not for a single world $w$ (as on the first theory) but all of them.[39] Put another way, battitudes are individuated by the functions they induce from worldly contexts to the singular propositions that get judged in those contexts.

Not just anyone is going to share my [belief that $p$] on the new approach. They are going to have to grasp or conceive $p$ at least *somewhat* as I do, lest the difference induce a different proposition believed in some faraway world. But how much similarity of conception are we talking about here? All it takes for a

[36] They are certainly distinct in the worlds where they exercise this capability; and if there, then everywhere, for duplicates are battitudinally indiscernible.

[37] See the first two chapters of White 1991 (one of which dates back to 1982) and Fodor 1987.

[38] This papers over a real problem: namely, how to decide *which* of the propositions believed by my doppelganger in $w$ gets to count as $p(w)$—$p(w)$ being the proposition your doppelganger had better also believe in $w$ if she wants to share my [belief that $p$].

[39] Or as many as makes sense; one doesn't *have* doppelgangers in every world. I'm going to ignore this problem.

thinker *not* to share my [belief that *p*], remember, is that there be *something* in their take on reality, no matter how little connected to *p*, that in some world *w*, however distant or contrived, swings the propositional content of their belief away from that of my doppelganger in *w*. It is natural to wonder whether there is *any* difference in [attitude] that could not be exploited to achieve this result in a suitably wacky world.

Here is why. What my [belief that *p*] is about in a world *w* depends on what it is in covariational thrall to there. But on anybody's account, the covariational channels through which content flows are shaped and sustained by various sorts of external props: paradigms, measuring devices, experts, and the like. No doubt there are worlds in which all available props converge on the same external referent; all the instruments agree, as it were. But there will also be worlds in which switching the prop puts the thinker en rapport with a *different* referent. Any change in [attitude] with even the *potential* to shift my allegiances as between props thus engenders an *actual* change in the function from contexts to attitudes that constitutes my [belief that *p*]. And it is hard to think how a change in [attitude] could lack this potential—how I could "change my mind" without *in any circumstances whatever* tipping the balance in favor of deference to a different class of paradigms, measuring devices, experts, or what have you. Variation in ~~any~~ one [attitude] therefore entails variation in all of them.[40] FN:40

This problem for the second theory of the battitudes can be called *subjective meltdown*. Because what we are seeing is that to share my [belief that *p*], you must share my total subjective outlook—or, what comes to the same, my [belief that *p*] *is* my total subjective outlook.[41] FN:41 ~~If there is anything to~~ From subjective meltdown ~~at all,~~ we see that the second theory essentially just inverts the difficulties we found with the

[40] Here is Fodor:

> what I use to manipulate the correlation between my *elm* thoughts and elms is not an instrument but a botanist. To do *that* sort of thing, I must be able to pursue policies with respect to another person's mind as well as my own. And also with respect to the causal relations between our minds. I am relying on its being reliable that elms will cause the botanist to have *elm* thoughts; which in turn will cause him to utter elm reports; which in turn will cause me to believe that it is an elm I have to do with. Setting things up so that all this *is* reliable requires that I be very clever, that I know a lot (for example, I have to know which experts I can trust) and that I be prepared to pay what a botanist's services cost. But it is likely to be worth the trouble. (1994, 36)

Fodor says that I have to know which experts I can trust. Had my *elm* thoughts been under the control of different experts, they might have been correlated in the content-determining way with a different kind of tree. But since who I trust about *elm* is a function of collateral [attitudes], so is *elm*'s meaning in my neurolect.

[41] Compare Block 1991's arguments that narrow content is holistic and Fodor's response in the same volume. Fodor thinks that Block has confused two questions: namely, "what fixes the propositional content of mentalese token 'X' in a given context?" (answer: nomic relations with the outside world, regardless of relations with collateral mental items) and "what in my mental life helps to sustain 'X' in the relevant nomic relations?" (answer: relations with collateral mental items among other things):

> the N-relation ["the nomological relation N such that your "X" tokens refers to Xs . . . iff they bear N to Xs"] is . . . *robust*; many theories . . . might succeed in sustaining the N-relation between

first; it delivers *over*specific battitudes, turning "[water-wanters] c.p. drink" into an *under*generalization along the lines of "pennies conduct electricity".

## 12. BATTITUDES AS OVERCOMMITTAL ANYWAY

Now for the "real" reason not to take it for granted that proportionality backs the battitudes over the attitudes. This is a reason that continues to apply even if subjective meltdown is somehow avoided—even if battitudes are the separately identifiable cognitively revealing items their proponents have wanted them to be.

Imagine that to each of my attitudes *A* corresponds a distinct subjective state $[A]_{SY}$ that sums up what I within the privacy of my own head to be in *A*.[42] So, [desire for water]$_{SY}$ is what I do internally to *desire water* (as opposed to what I do to desire fried green tomatoes or to believe that okra is slimy). The state thus picked out *might* be a hankering after the odorless, tasteless, transparent, river-filling stuff that etc., etc. But it might equally well be some sort of syntactical and/or neural state.

FN:42

The point in either case is the same. The appeal to these states can exempt the intrinsicalist from charges of fracturing the intentional generalization (5) only on a certain condition: all or almost all water-wanters must be in the state of [wanting water]$_S$ for some value of *S*. And this is just not plausible. How a person's water-desire is neurally implemented, the precise mentalese orthography involved, the fine detail of the water's internal mode of presentation, all of these may be expected to vary enormously without much effect (ceteris paribus) on the desirer's probability of drinking.

## 13. BACK TO WITHIN

equalize font size

A case can thus be made that WITHIN's nomic analogue is guilty of double dealing. After much handwringing about intentional states' overspecificity relative to this

'dogs' and dogs in this world, 'dogs' and twin-dogs in Twin-world, 'dogs' and things-just-like-our-dogs-except-for-the-ears in Cousin-world . . . and so forth. So, many different belief systems might implement the narrow content DOG. Or, if this is not right, Block needs an argument to show that it isn't. And, so far, I don't see that he's got one. (Fodor 1991*a*, 266)

The argument I would propose on Block's behalf is that while different belief systems will indeed implement the same N-relations in *some* worlds, the "and so forth" is unwarranted. Because the "and so forth" says in effect that differences in surrounding theory are *necessarily* (across all possible worlds) incapable of bearing on what "dog" is N-related to. And it is not clear how both of the following can be true together: first, surrounding theory helps to *sustain* "dog" in its N-relations, but second, tweaking surrounding theory not only does not but *cannot* affect those N-relations. (Again, a change of N-relations in *any* world entails a change of narrow content *here*.)

[42] "Subjective" in the sense of "intrinsic to the subject". This is to allow for syntactical and/or neural battitudes. Note that we might want to relativize to other parameters as well; the same person

or that intrinsic surrogate, that the surrogate states are similarly overspecific relative to their intentional originals is completely overlooked. The question is how much of this transfers over to WITHIN itself, which you'll recall goes as follows: According to the proportionality principle, causes should not be overloaded with unneeded detail—detail in whose absence the effect would still have occurred. But unneeded detail is exactly what we are getting when my desire for water is nominated as the cause of my hand going out to the cup. Had it been *twater* I wanted rather than water, then, holding my [desire] fixed, my hand would still have gone out.

Now, it might well be asked why (in the absence of information about how I *came* by my altered desire) this counterfactual should strike us as correct rather than merely baffling. But our problem is much more basic. Assume that my [desire] does screen off my desire in the way described. This can't itself put my [desire] in the driver's seat, for my desire might well return the favor. Even if it is true, in other words, that

(9) had my desire been different, then provided my [desire] had been the same, my hand would still have gone out,

it might *also* be the case that

(10) had my [desire] been different, then provided my desire had been the same, my hand would still have gone out.[43]

FN:43

And since (9) and (10) are absolutely symmetrical, any causal advantage the one might seem to confer on my [desire] is nullified by the other.

## 14. SYMMETRY

At least, my [desire]'s advantage is nullified if (10) is *true*. The intrinsicalist will say that it is not. Don't we have a million Frege-inspired examples to show that tiny differences in the way a proposition is presented can have enormous behavioral ramifications? Whereas if different propositions are presented in the *same* way (as in the Twin Earth examples), the same behavior results. The clear lesson of these examples is that behavior is driven less by *what* one believes/desires—by the

may judge the same proposition in different intrinsic ways at different times, or even at the same time via different mental representations.

[43] Compare Ruth Millikan: "Jerry Fodor has been considerably exercised (as he likes to say) by the (undoubted) fact that, knowing only that it is true *of* the girl next door that John wants to meet her, we cannot predict that John will exhibit next-door-directed behavior. For John may believe that this girl whom he wishes to meet languishes in Latvia . . . But a very straightforward (though extremely fallible) surmise still follows immediately from the fact that John desires to meet Jane . . . and from this fact *alone*. Namely, eventually John *will* meet Jane (say, after he gets back from Latvia)" (1993, 69).

propositional content of one's attitude—than by how that content is grasped. And if so then the very last thing we would expect is that switching the [desire] behind my desire, as in (10), will leave my behavior in place.

Sorry, but the *clear* lesson of the Frege and Twin Earth examples is only this. If we distinguish "what I believe/desire" from "how I believe/desire it" as factors in my extending my hand, then adjusting the how-factor alone *can* affect my behavior while adjusting the what-factor alone *cannot.* And this is compatible with (10), as an example brings out.

Whenever Isaac spots his bubbe in a photograph, he grins in recognition. Distinguish two factors in the grin on his face right now: *what* the photograph depicts (its subject or subjects, in this case my mother), and *how* it depicts (how intrinsically speaking the colors are arrayed). These two factors interact in something very like the way under discussion. Adjusting the how-factor alone *can* affect Isaac's behavior—had the photograph been much fuzzier, Isaac would have been baffled by it—while adjusting the what-factor alone cannot—leave the colors alone, and regardless of subject, Isaac grins. Shouldn't we then conclude that Isaac's behavior is controlled more by the picture's intrinsic colorproperties # than by its extrinsic, representational, ones? And if it is controlled more by the color properties, then the very last thing we would expect is that a *differently* colored picture of his bubbe would still have lead Isaac to grin.[44] FN:44

And yet, this is *precisely* what we would expect. Isaac is a boy capable of tracking his bubbe through a huge variety of photographic images, and the image at issue here is not anything special or strange but the one his bubbe *would* have given rise to if the actual image were for some reason ruled out. Why Isaac should suddenly lose sight of his bubbe in the alternative-image world nearest to this one is hard to understand. Harping on the fact that a change in intrinsic color properties is necessary and, if suitably dramatic, sufficient for a change in Isaac's reaction only drives the problem home; why should there be a *dramatic* change in color properties in the *nearest* alternative-image world to actuality?

Here is what the Frege and Twin Earth cases may indeed show: if you want to stop me from extending my hand, mucking with my [water-desire] alone can do it, whereas mucking with my water-desire alone cannot. But this is fully compatible with saying that many or most ways of mucking with my [desire] leave my behavior in place, provided that I keep on wanting water. And it is supremely compatible with the notion that I would still have extended my hand if I had wanted water in the way involving the least possible departure from actuality.[45] FN:45

[44] Cf. Walton's claim that to see a person's photograph is, or can be, to see the person (Walton 1984). His concern is *whether* Isaac sees bubbe; mine is *why* he sees, or seems to see, her. his

[45] To say it a little more clearly: From (i) no change in behavior without a change in battitude, that is, a purely attitudinal change won't do it, and (ii) a purely battitudinal change *will* do it, it does not follow that (iii) attitudes cannot screen battitudes off. (ii) is irrelevant to (iii), since whether my desire screens my [desire] off turns on the results of holding my desire fixed while varying my

So what, in other words, if a desire for water conceived as the-stuff-I-once-saw-through-an-electron-microscope, or as whatever-she's-drinking, would not have set my hand in motion? The screening-off issue concerns not *these* modes of presentation but the one(s) I *would* have enjoyed had I not conceived of water in the way that I in fact do. (Given the richness and multifacetedness of my actual conception of water, it would seem bizarre for there to be no closer alternative to my actual conception of water than one robbing my desire of its motive power.) So what if a sufficiently perverse mentalese encoding would have cut my desire off from its behavioral effects, as long as the closest alternative encoding(s) are not perverse?

## 15. PROPORTIONALITY

Generalizing madly, let us assert the following: any intrinsic state rich and complex enough to count as what-I-do-internally-to-judge-that-*p* is bound to exceed in some respects the causal requirements of any particular bit of behavior. If that is right, then the intrinsic causes that WITHIN favors are as open to charges of disproportionality as the extrinsic, intentional, causes that it rejects. Either the charges stand up in both cases—in which case *nothing* causes behavior—or they stand up in *neither* case. I say that they stand up in neither case. But then there must something wrong with WITHIN's understanding of proportionality.

What could it be? Proportionality has been kept at an intuitive level until now, mainly in order not to rain prematurely on WITHIN's claim to be relying on the same principle used in the response to BELOW. Suppose we look at that response again, this time with an eye to what it has in mind by proportionality:

> my brain state cannot expose my pain as causally irrelevant to my grimace, because it is a determinate of my pain; my pain, however, can knock my brain state out of contention for the role of cause, by screening it off and so exhibiting it as not required for, and hence out of proportion with, my grimace.

Working backwards, my brain state is not proportional to my grimace because it is not required; and it is not required because my pain—one of its determinables, note—screens it off. Here are the definitions right way around:[46] FN:46

(11) $C_1$ *screens* $C_2$ *off from* $E$ iff, had $C_1$ occurred without $C_2$, $E$ would still have occurred.

[desire]—not the other way around as in (ii). And between (i) and (iii) there is a palpable gap; what (i) says *can* happen through pure variation in [desire], (iii) says *would* happen *were* there pure variation in [desire].

[46] I have framed these definitions, and most of the subsequent discussion, with token causation in mind. For the application to types, think of $C$, with or without subscripts, as qualifying an implicitly given token cause $X$, and substitute "$X$ has $C$" for "$C$ occurs". $E$ remains a token effect.

(12) *C is required for E* iff none of its determinables screens it off, and *C is enough for E* iff it screens off all of its determinates.[47]

(13) *C is proportional to E* iff it is both required by and enough for *E*.[48]

To complete the response we should explain how my pain, having knocked my brain state out of contention for the role of cause, might come to occupy that role itself. If it were to screen off its *other* determinates (other than the brain state, that is), then by (12), my pain would be *enough* for the grimace. If it escaped a similar fate at the hands of its determinables, then by (12) again, it would be *required*. Both results together would by (13) make my pain *proportional* to the grimace and to that extent its cause.

## 16. THICK AND THIN

Here is what proportionality means in the response to BELOW: you are proportional iff you screen off your determinates and you avoid being screened off by your determinables. The question is whether WITHIN can get by on the same interpretation. Does the fact that attitude *A* is screened off from a behavioral effect by battitude [*A*] knock *A* out of proportion with that effect *in the sense of proportionality just laid down*?

That depends on how we resolve an unremarked ambiguity in talk of attitudes like *A*. That *A* is extrinsic is agreed (remember Putnam and Twin Earth). But an extrinsic state need not be extrinsic through and through; it can have intrinsic parts or aspects. This is obvious in the case of rigged-up examples like *being spherical and P*, where *P* is a property as extrinsic as you like. But there are plenty of ordinary examples as well. Being a horse (stamp, crater, . . .) involves a horsy history *together with* a horsy intrinsic character. Even that paradigm of extrinsicness, the property of being five miles from a burning barn, is not altogether free of intrinsic content. To be five miles from anything, you need spatial boundaries, and it seems an intrinsic property of a thing that, along some dimensions at least, it finally peters out.

Now, from the Twin Earth examples we know that *A* is extrinsic in respect of its truth conditions or singular propositional content.[49] But this does not prevent it from being intrinsic in other respects. One possibility is that *A* *includes* the thinker's internal contribution to the fact that such-and-such is

[47] I used to say that *C* was enough for *E* iff no determinate of *C* was *required* for *E* (Yablo 1992*a*, b). This was a weaker reading of enoughness, since determinates of *C* had only to be screened off by some determinable or other, not necessarily by *C* itself. I now prefer the definition in the text.

[48] Yablo 1992*a* (ch. 8 above) and 1992*b* put two further conditions on proportionality which can be ignored here. *E* is *contingent* on *C* iff it would not have occurred if *C* had not occurred; and *C* is *adequate* for *E* iff had *C* not occurred, *E* would have occurred if it had.

[49] These are different, but not in a way that matters here.

the truth-conditional content she judges; that is, *A* might be a *determinate* of [*A*] = its image under the bracketing operation. Attitudes like this, which have their corresponding battitudes as determinables, will be called *thick*. Another possibility is that *A* is (relatively) noncommittal about the thinker's internal contribution; it is *not* a determinate of [*A*]. Attitudes like this will be called *thin*.[50]

How does the thick/thin distinction affect [*A*]'s ability to knock *A* out of proportion with behavioral effects? Simple—thick *A* has [*A*] as a determinable, and (11)–(13) say that *A* had better not be screened off by any of its determinables if it wants to come out proportional to *E*. *A* is not proportional to *E*, then, if it is screened off by [*A*]. (Compare: my screaming "wake up!!" in my cat's ear is not proportional to its waking up if it is screened off by my screaming in the cat's ear as such.) Whether screening off in fact occurs depends on the details of the case—on whether *E* would still have occurred had the thinker judged a different proposition by way of the same battitude. But if the factors responsible for the switch in proposition are far enough removed from the causal scene, then *E* is probably not going to be affected.

About *thick* attitudes, then, WITHIN has a point; they really *are* in danger of being knocked out of proportion with typical behavioral effects by their intrinsic counterparts. But if you have been following me this far, you will see that *thin* attitudes are in no comparable danger. This is because thin *A* has no intrinsic determinables worth speaking of—certainly not [*A*], for *A* does not determine [*A*][51]—and it takes a determinable of *A* to expose it as not required for the effect. Thin attitudes have nothing to fear from WITHIN.

## 17. SUPERPROPORTIONALITY

If a *determinable* of *C* screens it off, then *C* is not required for *E*. But, why the restriction to determinables? What is so special about them that only they have the power to break *C*'s causal connection with *E*? This is crucial because extending equivalent veto power to *non*-determinables would bring thin *A* under the same proportionality pressure as thick.[52] And if thin *A* loses its advantage over thick, then not much remains of our defense of wide causation.

[50] "Attitude" in this paper has generally meant "thin attitude", and it will continue to do so. Note that on the "thick" reading it would be (trivially) false to say that my desire for water is more widely shared than my [desire for water].

[51] Nor does [*A*] determine it; the two are just incomparable, like the property of being a photo of Isaac's bubbe and the property of being a photo with such-and-such intrinsic color features.

[52] Demonstrably so, since [*S*] screens off both if either. *Proof*: Because thin *S*'s differences from thick lie entirely within [*S*], [*S*] occurs without the one in the same worlds as it occurs without the other. But then *E* inhabits the nearest world containing [*S*] without thick *S* iff it inhabits the nearest world containing [*S*] without thin *S*.

So again, what is so special about *C*'s determinables? And while we're at it, what is so special about its determinates that *C* need only screen *them* off to be proportional to *E*?

Nothing, you might say. The fact that *C* is screened off at all shows that, other things holding fixed, the effect would still have occurred without it. And even a single state not screened off by *C* shows that *C* cannot itself supply all of the effect's causal needs. What is the point of a proportionality condition if not to show "causes" like this the door? ~~Never mind~~ the ineffectual (12) and (13); Instead of
let's have

(14) *C* is *super*required for *E* iff *nothing* screens it off, and *super*enough for *E* iff there is *nothing* it fails to screen off[53]

and

(15) *C* is *super*proportional to *E* iff *C* is superrequired and superenough for *E*.

Bertrand Russell seems to have been in the grip of some such idea in "On the Notion of Cause". Because here is what he argued, or provided the materials for arguing:

> *C* cannot cause a strictly later event *E* except via some causal intermediary *D*. But then *C* is not superenough for *E*, since it would not have been followed by *E* but for *D*'s assistance.[54] (Nor is it superrequired, since given *D* it makes no difference to *E* whether *C* occurs or not.) So there can be no temporal gap between cause and effect. Can we at least say that *C begins* earlier than *E*, ending at (or after) the time at which *E* begins? No, for the parts of *C* occurring prior to *E* would have to be written out as not superrequired.[55] The only true causation is simultaneous causation.[56]

So much for the old canard about the future being causally beholden to the past. By the time Russell is done, the universe has disintegrated into a loose succession of moments, each sponsoring feverish causal activity on a rigidly intramural basis.

[53] "Nothing" here means "no state or event actually in existence". The mere fact that there *could* have been a state or event that, had it existed, *would* have screened *C* off does not prevent *C* from being superrequired.

[54] "[T]here must be some finite lapse of time . . . between cause and effect. This, however, at once raises insuperable difficulties. However short we make the interval . . . something may happen during this interval which prevents the expected result. In order to be sure of the expected result, we must know that there is nothing in the environment to interfere with it. But this means that the supposed cause is not, by itself, adequate to insure the effect" (Russell 1917, 136–7).

[55] "[I]f the cause is a process involving change within itself, we shall require . . . causal relations between its earlier and later parts; moreover it would seem that only the later parts can be relevant to the effect . . . Thus we shall be led to diminish the duration of the cause without limit, and however much we may diminish it, there will still remain an earlier part, which might be altered without altering the effect, so that the true cause . . . will not have been reached" (Russell 1917, 135).

[56] That is, cause and effect must *begin* at the same time. Similar reasoning suggests that they must end at the same time as well.

Now, Russell intended his argument as a reductio of the whole notion of cause. But it works better as a reductio of (13)'s overheated conception of proportionality. The real lesson of Russell's argument is that to insist that causes screen off subsequent events, while not being screened off by them in return, imposes an absurd degree of intimacy on causal relations. This perhaps explains why no one has ever tried to deduce epiphenomenalism from the fact that mental states are screened off by the causal chains they extend towards behavior. If this sort of screening off were truly disqualifying, epiphenomenalism would be the least of our problems; essentially *everything* would be robbed of its intuitive causal powers.

No one imagines it makes beliefs and desires epiphenomenal to be screened off by events subsequent to themselves. But many *do* seem to think it makes them epiphenomenal that they are screened off by associated [beliefs] and [desires]. This is interesting because it seems to me that to count *this* sort of screening off disqualifying *also* imposes a disastrous degree of intimacy on causal relations.[57] The difference is that now the intimacy is of a modal nature rather than a temporal one. Instead of being forced to exist at the same times, *C* and *E* are forced to occur at the same or similar worlds.

## 18. DEDICATED PSEUDOCAUSES

Why a modal intimacy this time? Because it is primarily in modal respects that attitudes differ from their corresponding battitudes. As far as *this* world is concerned, my desire for water and my [desire for water] are just alike. They occur at the same time and, Putnam's slogan that "meanings ain't in the head" notwithstanding, in the same place. (He might as well have said that pennies ain't in the pocket, since events within the pocket do not suffice to make them *pennies*.) To the extent that content is categorical, they can even be said to have the same content or contents. *All* of their categorical properties are shared, or near enough not to matter. Where my desire and my [desire] differ is in which of these properties they have essentially, or, what comes to the same, in their counterfactual careers. The desire persists into worlds where it is *water* that I want, even water grasped in a different intrinsic way; the [desire] persists into worlds where it is *thusly* that I want, even if the thing thusly wanted is not water.

Using the term *coincident* for items that are categorically alike but hypothetically different, we can put the claim like this. Applied to events occurring at different times, superproportionality imposed an undue degree of temporal intimacy; applied to coincident events (events occurring at the same time but in different worlds) it imposes an undue degree of modal intimacy.[58]

[57] The "proportionality" principle laid down in the last paragraph of Yablo 1987 amounted to (15) restricted to coincidents; I hereby withdraw it.

[58] Except of course when one of the coincidents is a determinate of the other; then we are back to simple proportionality. The relation between (token) determination and coincidence is this. *D* is

Epiphenomenalism is the least of our problems either way, because too much intimacy of either sort makes an absolute hash of the causal order.

A few sections back we saw how coincident would-be causes can screen one another off; in the terms of (14), each exposes the other as not superrequired for the effect. That was just the tip of the iceberg, however. Another scenario involves three candidate causes, all coincident, with the first screening off the second, the second screening off the third, and the third screening off the first. (Had the miller girl guessed the little man's name without guessing "Rumpelstiltskin"—his name was "Ralph"—or guessed "Rumpelstiltskin" without guessing his deepest secret—he had a still *deeper* secret—or guessed his deepest secret without guessing his name—"Rumpelstiltskin" was not *his* name but that of his invisible friend—he would *still* have stamped himself into the ground.[59]) Again, none of the candidate causes is superproportional with the effect. How often does this sort of situation arise?

Here are some crude statistics to suggest what the superproportionalist is up against. If $C_1, \ldots, C_n$ are coincident events each up for the role of causing $E$, then $C_i$ causes $E$, according to superproportionality, only if

for all $C_j$, $E$ would still have occurred had $C_i$ occurred in $C_j$'s absence, and

for all $C_j$, $E$ would not have occurred had $C_j$ occurred in $C_i$'s absence.

Call the scenario where *none* of the $C_i$s passes this test—where each has its candidacy destroyed by some other—*collective self-destruction*. What we are after is an estimate of its probability. As a basis for calculation let's say that between the hypothesis that $E$ *would* have occurred had $C_j$ occurred without $C_i$, and the hypothesis that it *wouldn't* have occurred, there is nothing to choose; one candidate cause is a priori as likely to screen another off as not to do so. (This is debatable, but never mind; any other estimate only increases the chances of collective self-destruction.) Then the probability of $C_i$'s escaping elimination at the hands of $C_i$ is 1/4—for there is half a chance of its being screened off by $C_j$ and half a chance of its failing to screen $C_j$ off. Assuming that these probabilities are relevantly independent,[60] we can reason as follows:

the chance of $C_i$ escaping elimination by $C_j = 1/4$, so

the chance of $C_i$ escaping elimination altogether $= (1/4)^{n-1}$, so

a determinate of $C$ iff (i) $C$ inhabits every world that $D$ does, and (ii) wherever both exist, they are coincident. Details can be found in Yablo 1987, 1992*a* (ch. 8 above), and 1992*b*.

[59] To avoid any appearance of scope confusion, the claim is this. Where $C_1 =$ her guessing the little man's name, $C_2 =$ her guessing "Rumpelstiltskin", and $C_3 =$ her guessing his deepest secret, had $C_1$ occurred without $C_2$, or $C_2$ without $C_3$, or $C_3$ without $C_1$, the effect $E$ would still have occurred. That $C_1$ *can* occur without $C_2$ (etc.) shows that we have not one event here but three, albeit three *coincident* events (see Yablo 1987 and 1992*b*).

[60] Assuming, that is, that (i) elimination at the hands of one candidate cause is independent of elimination at the hands of another, and that (ii) one candidate cause's being eliminated is independent of another's being eliminated. (ii) is not strictly true, since the hypothesis that $C_i$ is

the chance of $C_i$ being eliminated $= 1 - (1/4)^{n-1}$, so

the chance of each $C_i$ being eliminated $= (1 - (1/4)^{n-1})^n$.

This is not a negligible figure, even for small values of *n*. With two candidate causes, self-destruction is 56% likely; with three it is 82% likely; with four it is 94% likely; and with five it is 98% likely. With six candidate causes there is only FN:61 one chance in a hundred that some $C_i$ will ~~stave off~~ elimination.[61] escape

It is true that the "right" candidate cause could beat the odds. But think what "right" has to mean here. A $C_i$ which occurred in the very same worlds as $E$ would not be in any danger. But any departure from this ideal is potentially a departure from superproportionality. For $E$ to occur without benefit of $C_i$ in even a single world $w$ opens $C_i$ up to charges of not being superrequired for $E$. (What it would take to make the charges stick is a $C_j$ such that $w =$ the closest world to actuality in which $C_j$ occurs in $C_i$'s absence.) Likewise a single world in which $C_i$ occurs without $E$ opens $C_i$ up to charges of not being superenough for $E$. (Here we would need a $C_j$ such that $w =$ the closest world in which $C_i$ occurs j in $C_i$'s absence.) Superproportionality comes perilously close to the demand that causes be unconditionally necessary and sufficient for their effects—as close as FN:62 the pool of candidate causes permits.[62]

Pressurizing causes to exist in the same worlds as their effects is a bad idea. That $E$ is not likely to *have* an antecedent quite this modally attuned to it is only part of the problem. Even if such an antecedent were found, call it $C_e$, we would be hard put to regard it as $E$'s cause. After all, this would be an event with existence-conditions roughly as follows: $E$'s causal needs are somehow or other met. Surely it is not $E$'s causal needs being met that does the causing, it's the whatever-it-is that in fact meets them.

Imagine, though, that we stifle our doubts and accept $C_e$ as cause; then our troubles are just begun. An event so closely identified with $E$ is in a poor position to cause *other* effects, especially if causation requires the high degree of

eliminated *raises* the chances that it was eliminated by $C_j$, which *lowers* the chances that $C_i$ eliminates $C_j$ in return. (If $C_j$ eliminated $C_i$ by screening it off, then $C_i$ cannot eliminate $C_j$ by failing to be screened off by it, and vice versa.) The formula in the text is close enough to the truth not to matter.

[61] If the power of elimination is reserved to $C_i$'s determinates and determinables, chances of self-destruction are *zero* until the number of candidate causes hits four. And self-destruction will always be rare, because of the following fact. Using < for the is-a-determinable-of relation, and letting a *zigzag* be a sequence of $C_i$s such that $C_1 < C_2 > C_3 < C_4 > \ldots$, a set of candidate causes self-destructs only if each of its members is connectable by a zigzag to a circular zigzag of cardinality four or more.

[62] Ordinary proportionality raises similar problems (Yablo 1992*b*, section 11), but not on anything like the same scale. Technically this is because the chances of finding a $C_j$ screening $C_i$ off (a $C_j$ that $C_i$ fails to screen off) are greatly reduced if we require $C_j$ to exist in all (only) the worlds that $C_i$ exists in. Intuitively it is because a determina~~te~~ of $C_i$ that screens it off (a determina~~ble~~ te of $C_i$ that it fails to screen off) is prima facie a *better* candidate than $C_i$ for the role of cause. ble Superproportionality allows $C_i$ to be killed off by its causal inferiors; proportionality keeps $C_i$ alive until something better turns up.

modal attunement now being contemplated. (Do not suppose that it will cause these other effects via $E$. Ordinary events like $E$ have long since fallen out of superproportion with their supposed progeny.) And how $C_e$ is supposed to be FN:63 provided with a superproportional cause of its own is anybody's guess.[63] Any comfort that superproportionality might seem to lend epiphenomenalism is thus a sideshow compared to its real ~~project~~ upshot. The world we have now is a richly connected cosmos, run through with multiply branching causal chains. Given the right sort of ammunition (the right pool of candidate causes), superproportionality would lay waste to this arrangement, leaving behind a great disorderly mass of effects FN:64 each tracing back to an unmoved mover dedicated precisely to it.[64]

too small?

## 19. BELOW vs. WITHIN

About one thing WITHIN is right: intentional causes incorporate unneeded detail. But *all* intuitive causes, intentional or not, are like this. Do we really want to deny that the miller girl's guessing "Rumpelstiltskin" caused the little man to stamp himself into the ground, on the basis that so long as she had guessed *his name* (whatever it happened to be) the result would have been the same? If so, then we are well on the way to a world of dedicated pseudocauses consisting in *whatever it takes* for a given effect to eventuate.

Now this, coming on the heels of our objection to brain states as incorporating unneeded microphysical detail, may seem to raise double dealing to new heights. Intentional causes can do it, but neural ones cannot—is that it? But there is an objective difference here. $C$ incorporates unneeded detail iff it incorporates detail that (as I keep on saying) the effect could have done without, *suitable other things holding fixed*. Focus with me for a moment on these "suitable other things", the ones in whose continued presence the effect would still have occurred. Are they *included* in $C$, or do they lie outside it?

If the suitable other things are included in $C$, then they *remain* when the unneeded detail is stripped away. Stripping that detail away therefore yields a *determinable* of $C$ that would still have been succeeded by the effect even in $C$'s absence. This is how it is with my brain state and my pain.

Now suppose that the suitable other things are *not* included in $C$, as when $C$ is a thin attitude and what gets held fixed is $[C]$. Then the result of stripping

[63] This would be an event with the existence-conditions that *something* has happened given which *something* has happened given which $E$'s causal needs are met.

[64] Compare Yablo 1992*b*, section 11. There may be room for a screening-off-type condition stronger than proportionality but weaker than superproportionality. Several people have suggested the following: $C$ causes $E$ only if $C$ is not screened off *asymmetrically*, that is, by anything that it does not screen off in return. This is helpful when the candidate causes number exactly two, but when $n = 3$ or more, it becomes possible for each $C_i$ to asymmetrically screen off one of its fellows while being asymmetrically screened off by another. The Rumpelstiltskin example is (or can be made ~~to be~~) a case in point.

into

the unneeded detail away (unneeded extrinsic detail in the case of interest) is too impoverished to do meaningful causal work.[65] *This* sort of unneeded detail will have to be tolerated, because there is nothing to cover for it in its absence.

The point not to lose sight of is that there is no hope of evading the difficulty by attempting to *compensate* the cause somehow for its extrinsic losses. This can only push the bulge elsewhere, because apart from the whatever-it-takes pseudocauses rejected above, *all* causes, even purely intrinsic ones, contain an element of the unneeded. Tradeoffs are unavoidable; we buy relief from one sort of unneeded detail by taking on detail of another sort. When the tradeoffs balance out, we can attribute the effect to a relatively extrinsic cause or a relatively intrinsic one as we choose.[66] When a modicum of extrinsic detail buys up an abundance of intrinsic, we have wide causation pure and simple.

## 20. INNOCENCE

If we could but recover our pre-Fregean intentional innocence, it would seem incredible that the desire leading me to reach just now for water had much more to its content than this: I get water.[67] What normally and primarily drives behavior is outwardly directed attitudes, not how those attitudes happen to be encoded in people's heads.[68]

[65] Assuming that stripping it away yields an entity at all—there are questions here about the limits of determinability.

[66] Note that as long as neither determines the other, $C_i$ and $C_j$ can *both* be proportional to a given effect. (Which is good; see the discussion in Yablo 1992*b*, section 12, of "world-driven" and "effect-driven" causes.) Superproportionality, by contrast, leaves at most one candidate cause standing. (Suppose for contradiction that $C_i$ and $C_j$ are both superproportional with $E$. Either $C_i$ screens $C_j$ off from $E$, or it doesn't. In the first case, $C_j$ is not superrequired; in the second, $C_i$ is not superenough.)

[67] This is a play on the last paragraph of Davidson's "On Saying That": "Since Frege, philosophers have become hardened to the idea that content-sentences in talk about propositional attitudes may strangely refer to such entities as intensions, propositions, sentences, utterances, and inscriptions. . . . If we could recover our pre-Fregean semantic innocence, I think it would seem to us plainly incredible that the words 'The Earth moves,' uttered after the words 'Galileo said that,' mean anything different, or refer to anything else, than is their wont when they come in other environments" (1984, 108).

[68] If I said that this followed from the fact that attitudes screen battitudes off, I would be guilty of the very thing I've been warning against: drawing an asymmetrical conclusion from symmetrical premises. (Battitudes screen attitudes off as well.) I can't appeal to proportionality considerations either, for there is nothing in the definition of proportionality—as opposed to superproportionality—to prevent two candidate causes' *both* being proportional to an effect, provided that they are incomparable with respect to determination. (See the second last note, and Yablo 1992*a* (Ch. 8 above), section 12.) Then why give the nod to the attitudes? Partly for shock value; partly because of skepticism about the battitudes; partly because of the rationality argument to follow; and partly because of a hard-to-defend intuitive feeling that that is the way the tradeoffs play out—on the whole and for the most part, you can buy more intrinsic detail with a fixed amount of extrinsic (truth-conditional) detail than the other way around.

And a good thing too. Because think what life would be like if the same truth-conditional contents, variously grasped, induced a comparable variety of behaviors. Frustrating, that's what. The more behaviors a fixed set of attitudes issues in, the harder it becomes for these behaviors to converge on desired results.

FN:69 How is it that people are so good at getting what they want?[69] Three generalizations go a long way towards accounting for this. First, people have a tendency to do the subjectively reasonable thing, as defined by their [desires] and [beliefs]. (Decision theory is not a complete descriptive failure.) Second, the subjectively reasonable thing is quite often the *objectively* reasonable thing, in a sense defined by the agent's desires and beliefs. (Lois Lane snubbing Clark Kent, whom she *de re* adores, is the exception that proves the rule; it piques our interest because it doesn't usually happen.) Third, the objectively reasonable action is quite often the objectively *right* action, in a sense given by the agent's desires and the world. (Facts relevant to the success of our behavior are generally known to us.)

Imagine that we were the sort of creature that was liable to be driven hither and thither by variation in [desire] and [belief], even with all relevant desires and beliefs held fixed. Then the second of the three generalizations would be undermined. According to it, the subjectively reasonable action (the one that typically gets performed, remember) tends to be the objectively reasonable action. But how is it possible for these actions to remain the same when the one is changing with each shift in [attitude] and the other is staying put? Sensitivity to pure variation in [attitude] hurts our chances of doing the objectively reasonable thing, and hence of doing the objectively right thing, and hence of obtaining FN:70 satisfaction.[70]

## 21. CONCLUSION

Nourished from earliest days on a one-sided diet of Frege ~~examples~~ cases, and impressed by the vast causal difference a slight shift in subjective conception can make, philosophers have assumed that the richer intentional states are in subjective detail, the better ~~adapted~~ suited they are to the causation of behavior.

But (as one might have guessed from the fact that it took a Frege to think them up) Frege examples are *special*. What ordinarily happens is that the agent

[69] I assume that "getting what you want" at least involves the referential content of your desire coming true.

[70] "But sensitivity to pure variation in [attitude] can work to our advantage. Julie's [desire to be with Jekyll] and [desire *not* to be with Hyde] combine to put her in Jekyll's company in his high-functioning periods while keeping her out of his clutches when he goes into monster mode. Hasn't she gotten what she wanted?" Apparently so. This kind of case needs further discussion. (Thanks here to Mark Crimmins.)

could have grasped her proposition in any *number* of ways at no cost to the ensuing behavior. This and related oversights lead the standard view to reverse the true state of affairs. The richer an intentional state is in subjective detail, the more proportionality argues for *rejecting* it in favor of its subjective core.[71] Better equipped for causal duty are subjectively *impoverished* attitudes. These are safe from WITHIN, and, stressing as they do the external situation grasped over subjective nuances, more commensurate with typical behavioral effects. Normally I reach for water because I want *water*, never mind the phenomenology.[72]

## REFERENCES

Anderson, C. A., and Owens, J. (1990). *Propositional Attitudes: The Role of Content in Logic, Language, and Mind*. Stanford, Calif.: CSLI.

Bedau, M. (1986). "Cartesian Interaction". *Midwest Studies in Philosophy* 10: 483–502.

____(1991). "What Narrow Content is Not". In Loewer and Rey (1991), 33–64.

Broad, C. D. (1925). *Mind and its Place in Nature*. London: Routledge & Kegan Paul.

Brown, C. (1993). "Belief States and Narrow Content". *Mind & Language* 8: 343–65.

Cottingham, R. Stoothoff, and D. Murdoch CSM (eds.) (1984). *The Philosophical Writings of Descartes*, ii. Cambridge: Cambridge University Press.

Davidson, D. (1984). *Inquiries into Truth and Interpretation*. Oxford: Oxford University Press.

Dretske, F. (1988). *Explaining Behavior: Reasons in a World of Causes*. Cambridge, Mass.: MIT Press.

____(1993). "The Nature of Thought". *Philosophical Studies* 70: 185–99.

Fodor, J. (1980*a*). "Methodological Solipsism Considered as a Research Strategy in Cognitive Psychology". *Behavioral and Brain Sciences* 3: 63–98; repr. in Fodor (1981) pp. 225–53.

____(1980*b*). "Methodological Solipsism: Reply to Commentators". *Behavioral and Brain Sciences* 3: 99–109.

____(1981). *Representations*. Cambridge, Mass.: MIT Press.

____(1982). "Cognitive Science and the Twin-Earth Problem". *Notre Dame Journal of Formal Logic* 23: 98–119.

____(1987). *Psychosemantics*. Cambridge, Mass.: MIT Press.

____(1986). "Individualism and Supervenience". *Proceedings of the Aristotelian Society*, supp. vol. 60: 235–62.

____(1989). "Making Mind Matter More". *Philosophical Topics* 17: 59–79.

____(1991*a*). "Replies to Critics". In Loewer and Rey (1991), pp. 255–319.

____(1991*b*). *A Theory of Content and Other Essays*. Cambridge, Mass.: MIT Press.

____(1991*c*). "A Modal Argument for Narrow Content". *Journal of Philosophy* 88: 5–26.

____(1994). *The Elm and the Expert: Mentalese and Its Semantics*. Cambridge, Mass.: MIT Press.

[71] Using "subjective", as above, for "intrinsic to the subject".

[72] And the orthography, and the neurology.

Horgan, T. (1991). “Actions, Reasons, and the Explanatory Role of Content”. In McLaughlin (1991), pp. 73–101.

Huxley, T. H. (1911). *Method and Results*. New York: Appleton.

Jackson, F., and Pettit, P. (1988). “Functionalism and Broad Content”. *Mind* 97: 381–400.

Kim, J. (1991). “Dretske on How Reasons Explain Behavior”. In McLaughlin (1991)
● Q1 ●pp. 57–77; repr. in Kim (1993), pp. 285–308.

—— (1993). *Supervenience and Mind*. Cambridge: Cambridge University Press.

Kripke (1980) *Naming and Necessily*, ~~pp.~~ Oxford: Blackwell.

Loar, B. (1985). “Social Content and Psychological Content”. In R Grimm and D Merrill, *Contents of Thought*, Tucson: University of Arizona Press, pp. 99–110.

Loewer, B., and Rey, G. eds. (1991). *Meaning in Mind: Fodor and his Critics*. Oxford: Blackwell.

Long, A. A., and Sedley, D. N. (1987). *The Hellenistic Philosophers*, Cambridge: Cambridge University Press.

Malcolm, N. (1968). “The Conceivability of Mechanism”. *Philosophical Review* 77: 45–72; repr. in Watson (1982), pp 127–49.

McGinn, C. “The Structure of Content” in *Thought and Object* (1982) ed. Woodfield, A. pp 207–58.

McLaughlin, B. (1991). *Dretske and his Critics*. Oxford: Blackwell.

Millikan, R. (1993). *White Queen Psychology and Other Essays for Alice*. Cambridge, Mass.: MIT Press.

Pettit, P., and McDowell, J. (1986). *Subject, Thought, and Context*. Oxford: Oxford University Press.

Putnam, H. (1975). “The Meaning of ‘Meaning’ ”. In *Mind, Language, and Reality*, Cambridge: Cambridge University Press, pp. 215–72.

Russell, B. (1917). *Mysticism and Logic*. London: Allen & Unwin.

Searle, J. (1983). *Intentionality*. Cambridge: Cambridge University Press.

—— (1995). “The Mystery of Consciousness”. *New York Review of Books* 42, 17: 60–6.

Stalnaker, R. (1989). “On What’s In the Head”. *Philosophical Perspectives* 3: 287–316.

—— (1990). “Narrow Content”. In Anderson and Owens (1990), pp. 131–45.

—— (1991). “How to Do Semantics for the Language of Thought”. In Loewer and Rey (1991), pp. 229–37.

Stich, S. (1978). “Autonomous Psychology and the Belief-Desire Thesis”. *The Monist* 61: 573–91.

—— (1980). “Paying the Price for Methodological Solipsism”. *Behavioral and Brain Sciences* 3: 97–8.

—— (1983). *From Folk Psychology to Cognitive Science: The Case Against Belief.* Cambridge, Mass.: MIT Press.

● Q2 —— (1991). “Narrow Content Meets Fat Syntax”. In Loewer and Rey (1991), 239–54.●
Walker, V. (1990). “In Defense of a Different Taxonomy: Reply to Owens”. *Philosophical Review* 99: 425–31.

Walton, K. (1984). “Transparent Pictures: On the Nature of Photographic Realism”. *Critical Inquiry* 11: 246–77.

Watson, G. (1982). *Free Will.* Oxford: Oxford University Press.

White, S. (1991). *The Unity of the Self.* Cambridge, Mass.: MIT Press.

Woodfield, A. (1982). *Thought and Object*. Oxford: Oxford University Press.

Yablo, S. (1987). "Identity, Essence, and Indiscernibility". *Journal of Philosophy* 84: 293–314.

____ (1992*a*). "Mental Causation". *Philosophical Review* 101: 245–80; Ch. 8 above.

____ (1992*b*). "Cause and Essence". *Synthese* 93: 403–49.

## Queries in Chapter 10

Q1. Author correction is not clear.

Q2. [✓ 239–54] or 259–354? Author has provided two separate page number sets

# 11

# Causal Relevance: Mental, Moral, and Epistemic

~~An astonishing number of philosophers hold~~ Why would anyone think that mental states ~~are~~ were poorly positioned to cause behavior? Timothy Williamson in recent work distinguishes two sources of doubt about the causal prospects of *wide* mental states in particular.[1]

One is that causal attributions are supposed to have "an appropriate generality" (p. 81).[2] The idea goes back at least to Hume, who asserts in Section XV of Book I of the *Treatise*, "Rules by which to judge of causes and effects", that

> where several different objects produce the same effect, it must be by means of some quality, which we discover to be common amongst them. For as like effects imply like causes, we must always ascribe the causation to the circumstances, wherein we discover the resemblance.

Enter now the Twin Earth examples. You do the same thing—drink—whether it is water ($H_2O$) you desire or twater (XYZ). This seems to be just what Hume was talking about. Different "objects" (water-desire and twater-desire) produce the same effect (drinking). The causation must be ascribed to some quality that is common amongst them. That common element would seem to be something intrinsic: a narrow-content state, or a piece of brain-writing, or ~~perhaps even a brain state.~~ the agent's overall brain state.

A second source of doubt about wide mental causes is this. Causes operate via causal mechanisms; so they should be where the mechanisms are, specifically,

This paper was written for a conference on Tim Williamson's *Knowledge and Its Limits* held at University College, London, and presented subsequently at the 2001 Sofia conference in Veracruz, with comments by Louise Antony. Thanks to Louise, Alex Byrne, Tim Crane, Elizabeth Fricker, Paul Horwich, Michael Martin, David Papineau, Ralph Wedgwood, and Tim Williamson for extremely helpful comments and criticism. I am indebted, finally, to Braun 1995.

[1] Williamson 1998 and 2000. The focus here will be on propositional attitude states, counting perception as a propositional attitude. These can be wide either because of the content involved (realizing that *water is refreshing*) or the attitude taken toward that content (*realizing* that water is refreshing). It is width due to the attitude taken that primarily interests Williamson, but for our purposes the cases can be lumped together.

[2] Williamson references are to the book unless otherwise indicated.

where they are set into motion. (This idea goes back to Hume too, in his emphasis on contiguity.) Since behavior issues from mechanisms internal to the agent ("the causing of my present action is here and now" (p. 65)), its causes should be internal as well.

> [N]arrow conditions must play a privileged role in the causal explanation of action. If a causal explanation of action cites a broad mental condition, an underlying narrow condition must do the real work. We can isolate that narrow condition by subtracting from the broad mental condition the environmental accretions that make it broad . . . (p. 65)

It seems to strengthen the case against wide causation that the same conclusion is reached from two almost opposite perspectives. One argument *zooms in* on the causal transaction, on the theory that whatever sets the machinery into motion can be seen in its entirety from close up. The other argument *pans out* so that the transaction appears against the background of other transactions of the same sort. That extrinsic factors can be varied indefinitely at no cost to the effect exposes them as irrelevant hangers-on.

FN:3 All of this is to remind us that defending wide causation is a non-trivial task.[3] Mistakes will have to be found in both of the above arguments: *generality* and *locality*, I will call them. And they will have to be distinct and independent mistakes, since the arguments take such very different approaches. Williamson appears to do this. His complaint about *generality* is that it rests on a false assumption.[4] FN:4

• Q1 good

> •The narrow explanation is *not* always the more general. Take my seeing the ball on a certain occasion and the internal state of my head on that occasion. It may be that the internal state is as sufficient for the catch, other things equal. But it is nowhere near as necessary. I could have seen the ball in any number of ways, intrinsically speaking, without that compromising my ability to make the catch. I could have been watching it through either eye, for instance, or with my head cocked at any of a ~~large~~ number of angles. If the generality of an explanation is a matter of how well the factors cited correlate with the effect, then the narrow explanation is in this case the less general.

Against *locality* he says that

> The ~~most the~~ argument attempts to shift causal responsibility from ~~shows is that other things equal, the real cause is not~~ my knowing, or widely believing that *p*, to ~~, but~~ the internal core thereof: the "organismic contribution" as Dennett used to say. But ~~Other things are far from being equal, however, for~~ states of knowing (or widely believing) do not *have* internal cores to speak of. Knowing *would* have an internal core if it were "composite" = the conjunction of an internal fact with an external one. But knowing is "prime". If you insist on splitting knowledge states up between internal and external components, it is a huge disjunction of such conjunctions. The answer to "how do you have to be *internally*, to know there is

[3] Other recent defenders include Gibbons 2001, Jackson and Pettit 1988, and Yablo 1997.

[4] What follows is my account of what I take to be ~~Tim~~ Williamson's position. Similarly for the response to locality.

water nearby?" is "no particular way; what you need is to be in the right kind of harmony with your environment". If you tried to whittle knowledge states down to their environment-independent internal core, you would be whittling them down to nothing.

FN:5

As promised, each argument runs into a distinctive sort of trouble. *Generality* fails because wide mental states oftentimes *correlate* better. *Locality* fails because wide mental states are typically *prime*.[5] ~~This tidy arrangement proves unstable, however, for the~~ response to locality is incomplete.

But it seem to me that Williamson's

The primeness point shows that

~~If Williamson is right, then~~ a certain popular *strategy* for finding intrinsic surrogate states does not work; one cannot carve off the extrinsic parts and expect to have anything useful left. But why should the *real* causes of behavior have to be reachable by performing some kind of bracketing operation on the *alleged* causes? Maybe the alleged causes are wrong through and through. To get to the real causes, you have to drop them entirely and look somewhere else. It is clear there are *some* intrinsic surrogates, for there are brain states. Why not assign causal responsibility to them?

Now, Williamson does have something to say against the causal ambitions of brain states:

> If one cites a sufficient condition for the condition to be explained . . . the purported explanation can nevertheless fail because the condition to be explained would still have obtained in the same way even if the cited condition had not obtained. . . . Many features of [the brain state] will be quite irrelevant to the obtaining of [the effect]. They will concern physical events that form no part of the causal chain between the agent's initial mental state and the final performance of the action. The agent would have performed the action anyway, even if those features had been different. (pp. 81–2)[6]

FN:6

FN:7

Brain states are faulted here on the score of generality. They may be sufficient, but they are nowhere near as necessary, for their alleged behavioral effects.[7] Switching to (partly extrinsic) mental causes buys us a lot of necessity at the cost of not much sufficiency.

Isn't this a good and convincing reply? It is a good reply to something. But remember, the worry was about locality: no action at a distance. How are *correlational* considerations supposed to affect the *metaphysical* thought that extrinsic states bring in factors too far away to make a causal difference?

ital

Of course, Williamson does not suppose for a moment that every correlation is causal. He says, "The high correlations between prime mental conditions and conditions on subsequent action constitute *defeasible* evidence for the causal effectiveness of the prime conditions" (p. 88; emphasis added). But the route from correlation to causation is represented as pretty short: "Higher correlations

[5] Primeness enters in a limited way into the first response too, in that a prime state is touted as better correlated with the effect.

• Q2

[6] I have substituted brain states for ~~Tim's~~• actual target in this passage because the same considerations seem to apply.

Williamson's

[7] Williamson favors a probabilistic reading of sufficiency and necessity.

constituting a genuinely rival explanation would be needed to defeat that evidence" (p. 88).

The problem is that some high correlations are non-causal for reasons other than the one Williamson mentions, viz. that a higher correlation exists constituting a genuinely rival explanation.[8] One can, after all, get a perfect correlation by taking the disjunction of all conditions sufficient for the effect. It is hard to do better than perfect, yet the correlation isn't causal. So there have got to be other defeaters than trumping by rival correlations "constituting a genuinely rival explanation".

Someone worried about the locality problem will say: "Here is my opening. I suggest that your knowledge correlation is non-causal for one of these other reasons. It is non-causal because it is too extrinsic; it incorporates elements too far away to make a causal difference."

This is not such an unmotivated thing to say. A little generality is, causally speaking, a good thing: being hit by a bus, Williamson notes, is a better candidate for cause of death than being hit by a *red* bus. But there are limits. The pursuit of greater and greater generality eventually takes one *away* from the cause. Williamson acknowledges this in places:

> Some purported explanations achieve spurious generality by using disjunctive concepts. For example, if someone was crying because she was bereaved, it does not improve the explanation to say that she was crying because she was bereaved or chopping onions. But ordinary mental concepts of prime conditions (such as the concept of seeing) are not disjunctive. (p. 83)

I agree that ordinary mental concepts are not disjunctive. But how do we know they do not possess some *other* feature that defeats their claim to feature in genuine explanations?

Not all "spurious generality" has its source in disjunctiveness. If you are looking for a property of liquids that correlates well with the property of unplugging clogged drains, it would be hard to improve on *plumber-recommendedness*. Plumbers keep track of the substances currently getting stuck in drains, and the types of solvent available, and they recommend accordingly. Plumber-recommendedness certainly scores higher than any *chemical* property that might be mentioned; for there are various chemical preparations that work about equally well.

And yet plumber-recommendedness does no causal work. The various chemical properties do it, notwithstanding their poorer correlation with the effect. The objection to plumber-recommendedness is not that it is disjunctive but that it brings in factors too far away to influence goings-on in the drain. That could be how it is with knowledge and the brain state; knowledge correlates better, the brain state does the work.

[8] Williamson certainly knows this. The perfect correlation example about to be given is his, and he comments about it that "we are willing to sacrifice some degree of correlation" for the sake of unified (non-disjunctive) correlata (p. 89).

So far, then, we lack an answer to the metaphysical charge that wide mental states bring in factors too far away to bear on the causation happening here and now. I see only three lines of response with any chance of success. The first is *denial*: "Who says the factors making a state wide are too far away?" The second is *dismissal*: "Let me tell you my theory of causal relevance; *it* says wide mental states can be causes." The third is *diagnosis*: "Here is why the locality argument *seems* right." The rest of the paper looks at examples of each strategy.

## DENIAL

Williamson sometimes suggests that the factors making a mental state wide are *not* too far away, or at least won't *continue* to be as the time of action draws closer.

> One is thirsty; how likely is one to be drinking soon? Likely enough, if one sees water. Much less likely, if what one sees is a mirage . . . Concepts of broad mental conditions give us a better understanding of connections between present states and actions in the non-immediate future, because the connections involve interactions with the environment. FN:9 (p. 75)[9]

Having water in view lets me advance towards the water, renewing my perceptual link to it at various points along the way. It is not just that my perceptual state is at once wide and effective; it is effective *because* it is wide. Advancing on the water, I draw causally on the factors that make my state one of seeing water, as opposed to merely seeming to see it or seeing watery stuff. For I draw causally on the water itself, and on my continuing perceptual rapport with it.

rom I *said* that I draw causally on the factors that make my state wide. But, truth be told, it is really only some of them I draw on. That it is *water* I am seeing is owing in part to the stuff's being $H_2O$ as opposed to XYZ. But as far as my behavior is concerned, it might as well have been XYZ. That it is *seeing* I am up to depends on the process by which my visual experience is renewed. That process must be not only reliable but "of the right type" (no deviant causal chains). As far as my behavior is concerned, though, deviant causal chains would be fine, as long as they made for a replenishable supply of good information on the same topic.

So it does not seem that *all* the factors making the broad state broad make themselves felt en route to the action.

A second example is this. Knowing that the mine contains gold, you dig until you find it. Your belief that it contained gold does not explain your persistence as well as your knowing does. For your knowing involves inter alia that your belief

[9] "In deliberating, one assesses alternative courses of action in light of one's beliefs and desires, decides which is best, and forms the intention to pursue it. . . . How and whether one puts the intention into effect depend on one's interaction with the environment in the intervening period" (Williamson 1998, 396).

is based on true considerations and not false ones. (This is the no false lemmas condition on knowledge.) Suppose your evidence had been that there was gold behind a Maltese Cross carved into the cave wall. You would have looked there first, and abandoned the search on not finding anything. Your persistence was due in part to your not concentrating all your hopes on that particular spot. Here again, extrinsic aspects of your knowledge are not too remote to make a causal difference.

But again, that one such aspect plays a causal role doesn't mean that all do. You would (also) not have known that there was gold in the mine if some misleading testimony given in Carson City—testimony you should have been aware of but weren't—had not been refuted by court records in Reno—with you again unaware of the fact. The court records in Reno play no role in your continued digging here. But they are a factor in your knowing. This seems to bear out the localist complaint that your knowing incorporates factors too far away to affect the course of events.

## DISMISSAL

This is the strategy where we attempt an analysis of causal relevance, and show that broad mental conditions are relevant in the terms of that analysis. The locality argument is not refuted; it is not even mentioned; it is overruled by a higher court. We try to establish by other means the conclusion that locality was supposed to threaten.

I had Williamson claiming that states of mind are likelier causes of behavior FN:10 than brain states because they are better correlated with behavior.[10] But his words can be taken another way. He starts out by saying that a "purported explanation can . . . fail because the condition to be explained would still have obtained in the same way even if the cited condition had not obtained" (p. 81). (~~Psst . . .~~ note the counterfactual.) This applies in particular to action explanations that pin the blame on lower-level features of the agent's brain: "The agent would have performed the action anyway, even if those features had been different" (p. 82). (There's that counterfactual again.)

caps

To judge by these passages, the brain state is rejected not for correlational FN:11 reasons but counterfactual ones.[11] This could help, because counterfactuals, more anyway than correlation coefficients, appear to have a direct causal significance. Almost everyone's first thought about causal relevance is

(CT)

A property P of $x$ is causally relevant to effect $y$ iff $y$ would not have ensued, had $x$ occurred without P.

[10] I said that this was not decisive, for too much correlation can be a *bad* thing causally speaking.

[11] I doubt that this is Williamson's idea, not least because he told me it wasn't. Still, it would serve his interests if it worked.

Call that the *counterfactual theory of causal relevance*. What does it say about the cases of concern to us here?

Had I not seen that water, I would not now be drinking. So, according to the counterfactual theory, my seeing the water was causally relevant. It cannot be said, in most cases, that I would not be drinking had I not earlier been in a rather precise neural condition. So my precise neural condition is not causally relevant.

There is only one problem with this, which is that the counterfactual theory is wrong. This is clear from the debate about the causal relevance of *moral* properties.

Gilbert Harman said roughly this: there is no knowledge of moral properties unless moral beliefs are caused by such properties. When you look at particular cases, though, it turns out to be the underlying natural properties that do the causing. There is no need to appeal to the cruelty of setting cats on fire to explain our belief that those (cat-burning) children are being cruel. We came *into* the situation believing that torturing cats was cruel. Our belief that the children are being cruel is sufficiently explained by the non-moral fact that (cat-torturing) is what they are doing.

Nicholas Sturgeon asked why it should make moral properties *irrelevant* that non-moral properties are *sufficient*. After all, the moral properties might supervene on, or be otherwise bound up with, the non-moral ones. Cat-burning, to stick with that example, is necessarily cruel, or close enough for present purposes. But then, had the children not been behaving cruelly, they would not have been burning a cat, whence (unless Plan B was to *pretend* to burn a cat) their behavior would not have induced the belief that the children were being cruel. There would have been no belief, then, had there not been cruelty. How then can the cruelty be considered irrelevant to the belief?

Because Sturgeon argues from counterfactual dependence to causal relevance, responses to Sturgeon are a good source of counterexamples to (CT). Consider the case of Donald (due to Judy Thomson). To relieve his boredom at a particularly dreadful talk,

> [Donald] suddenly shouted Boo! at the speaker. In consequence, there was a loud Boo! sound on the tape recording of the speech. Now if there are moral facts at all, there is such a moral fact as that Donald's shouting Boo! was rude. But that fact was surely epiphenomenal relative to the . . . presence of a Boo! sound on the tape: the fact of Donald's shouting's having been rude surely plays no role at all in explaining the fact of the Boo! sound . . . [And yet] the case passes the counterfactual test for operativeness . . . For we may suppose that if Donald's shouting . . . hadn't been rude, it wouldn't have been [a case of] his shouting Boo at the speaker in mid-speech . . . in that case there would have been no shout at all during the speech . . . and . . . therefore no Boo sound on the tape.[12]

That is pretty convincing, I think. But it doesn't offer any guidance about how to fix the theory; and so we turn to a couple of structural problems. The first is

[12] Harman and Thomson 1996: 81.

that an effect that depends counterfactually on P depends all too often on $P \vee Q$ as well, even if Q is quite irrelevant. The second is similar, except that now it is $P \wedge Q$ that inherits "relevance" from P.

### Parasitic Disjunction

Had Donald's Boo! not been loud, there would have been no Boo sound on the tape. So there is no Boo sound on the tape in the nearest world *w* where it fails to be loud. Assuming that the Boo! would not have been regretted the next day had it not been so loud, *w* is also the nearest world where the Boo! fails to be *loud or regretted the next day*; whence there would (also) not have been a Boo sound on the tape had the Boo! lacked the disjunctive property of being loud or regretted the next day. Its property of being loud or regretted the next day therefore helped it (says the counterfactual theory) to register on the tape.

### Parasitic Conjunction

There is no Boo sound on the tape in the nearest world *w* where Donald's Boo! fails to be loud. Assuming that Donald would have delivered this quieter Boo! from the same position at the back of the room, *w* is also the nearest world where the Boo! fails to be both loud and a long way from the microphone. Since this is by hypothesis a world where there is no Boo sound on the tape, the tape recorder would not have picked it up had the Boo! lacked the conjunctive property of being *loud and far from the microphone*. Its property of being loud and far from the microphone therefore helped it (says the counterfactual theory) to register on the tape.

What is going on in these cases? One reason *loud or regretted* seems irrelevant is that there is a stronger property *loud* on which the effect still depends. But that cannot be all that is wrong, for the Boo! was also over ten decibels. It is true of *over ten decibels* too that there is a stronger property *loud* on which the effect still depends. And that the Boo! was over ten decibels seems highly relevant to the tape recorder's picking it up.

Is the problem that *loud or regretted* is not as natural as *loud*? No, because ~~that doesn't bother us~~ some loss of naturalness is tolerated when the weaker property is obtained by chipping away aspects of the stronger one that the effect would not have missed. It does not seem at all irrelevant to its registering on the tape that the Boo! was *loud enough given its distance from the microphone*.

The problem with *loud or regretted* is that it exhibits both of these failings at once. Passing from *loud* or *loud or regretted*, we suffer a decline in naturalness with no compensating gain on the score of attunement with the effect. (The effect would not have occurred had the Boo! been loud or regretted but not loud.) *Loud or regretted* is not merely weak but *egregiously, pointlessly*, weak.

The conjunctive property of being *loud and far from the microphone* offends in the opposite way. Passing from *loud* to it, we pile on irrelevancies—loud is

better proportioned to the effect than the property that replaces it—with no
FN:13 compensating gain on the score of naturalness.[13] This is a property that is not merely strong but *egregiously* strong.

Now let's try to make these notions a bit precise. Suppose that we are trying to identify the causally relevant properties of some object or event *x*, and that Q− and Q+ are weaker and stronger properties of *x*; Q+ necessitates Q− but not conversely. Egregiousness is defined in two steps:

put "still" above "not"!!

TR

*Def* Q− is $\left\{\begin{matrix}\text{better}\\ \text{worse}\end{matrix}\right\}$ proportioned to *y* than Q+ iff *y* would $\left\{\begin{matrix}\text{not}\\ \text{still}\end{matrix}\right\}$ have occurred, had x possessed Q− but not Q+.

ital

*Def* ~~A property~~ P ~~of *x*~~ is egregiously $\left\{\begin{matrix}\text{weak}\\ \text{strong}\end{matrix}\right\}$ (relative to effect *y*) iff some $\left\{\begin{matrix}\text{more natural stronger}\\ \text{as natural weaker}\end{matrix}\right\}$ property of *x* is better proportioned to *y* than P is.[14]
FN:14

Consider now the *proportionality theory of causal relevance*:

(PT)

A property P of *x* is causally relevant to effect *y* iff

FN:15 (a) had *x* lacked P, *y* would not have occurred[15]
(b) P is not egregiously weak or strong.

*Loud or regretted* meets condition (a) but not (b). That *loud* is a more natural stronger property better proportioned to the effect means that *loud or regretted* is egregiously weak. Similarly, *loud and far from the microphone* meets (a) but not (b). It is egregiously strong, since *loud* is a no less natural weaker property better proportioned to the effect.

How does my seeing the water fare on this theory? The worry would be that it
FN:16 is egregiously strong. That is, my seeing the water is screened off[16] by a weaker

[13] If a gain in naturalness is required to compensate for excessive strength, why not also to compensate for excessive weakness? (All we ask of the worse-proportioned weaker property is that it not be *less* natural than the better-proportioned stronger one.) There is a reason for this. An excessively strong property has aspects on which the effect does not depend, while an excessively weak one merely fails to include material on which the effect does depend. Including material that wouldn't have been missed is much more destructive of overall causal relevance than omitting material that would have been missed. (With causal sufficiency it is the other way around; omitting material that would have been missed is "worse" than including material in whose absence the effect would still have occurred.)

[14] This definition puts naturalness ahead of proportionality in the following sense: no amount of the latter can compensate for even a small loss of the former. A different definition would allow proportionality to be traded off—at a large discount, I assume—against naturalness. (This note was prompted by an example of Alex Byrne's.)

[15] Counterfactual theories of causal notions are subject to a standard objection: namely, the counterfactual fails if another cause would have taken over in the actual cause's absence. I take preemption worries seriously, but this not the place to discuss them. See Yablo 2002.

[16] One property of *x* screens off another iff *x* could have possessed the first property without the second, and had it done so, the effect would still have occurred.

property that is no less natural. The obvious choice is my seeming to see the water and its really being there. (I assume this is no less natural than my seeing the water, or at least that the objector could consistently maintain that it is.) Would the drinking still have occurred, had I only seemed to see water which was in fact there?

It is hard to feel sure about this; one wants to know *why* the seeming to see wasn't real seeing. Had I veridically hallucinated the whole time, my progress towards the water would not have been much affected. Had I veridically hallucinated for a brief moment between periods of non-veridical hallucination, it would have been another story. It does seem clear, though, that the drinking *might well* not have occurred, had I only seemed to see water which was however there. The seeing is not egregiously strong unless I *would* have had the drink had I merely seemed to see what was really there. That I might not have had it shows that the would-claim is false. So the seeing is not egregiously strong.

Or consider my knowing that you are home; the effect is my knocking and knocking for ten minutes. The worry would again be that ~~it~~ the knowing is egregiously strong, that is, screened off by a weaker property that is no less natural. The obvious choice is my truly and rationally believing that you are home. Would I still have knocked for ten minutes, had I correctly and with reason believed that you were home, without knowing you were home?

Once again, it depends on why the knowledge was missing. But it does seem that I *might* not have knocked so long; for I might have failed to know because of overlooking the "Not Home" sign posted on your door. Unless I somehow continued to overlook it for ten whole minutes, the effect (knocking for ten minutes) would not have occurred. And that is all ~~Tim~~ Williamson needs. The knowing is not egregiously strong unless I *would* have kept on knocking for ten minutes had I failed to know while still rationally and truly believing. That I might have stopped knocking shows that it is not the case that I would have kept on. So the knowing is not egregiously strong.

Now the rude Donald example. Thomson asks what properties of the *shouting* are relevant, but that raises a question we would rather avoid: would the effect still have occurred had the shouting lacked a property that is essential to it (loudness, say)? Better to ask about properties of the *shouter*. Donald has at a particular time the property of behaving rudely. How relevant is his possession of that property to the sound on the tape?

Unlike my knowing that you were at home, which we worried might be egregiously strong, Donald's rudeness is under suspicion of being egregiously weak, that is, of being weaker than some more natural property that it fails to screen off. A stronger property would be an action-type $\varphi$ such that necessarily, all $\varphi$ing is rude. Suppose for argument's sake that $\varphi$ing is shouting Boo!, or shouting Boo! in a crowded room. Boo!-shouters are all doing roughly the same thing, while (as any Grade 3 teacher will tell you) there is no end to the forms that rudeness can take. Shouting Boo!, even shouting it in a crowded room, is a whole lot more natural than the property of putting on some sort of rude performance or other.

One further thing is required for Donald's rudeness to be irrelevant because egregiously weak. Donald's property of shouting Boo! in a crowded room must be better proportioned to the effect—to the Boo sound on the tape—than his property of behaving rudely is. Would there still have been a Boo sound on the tape, had Donald been rude otherwise than by shouting Boo! in a crowded room? Well, consider some nearby alternatives. Would there still have been a Boo sound on the tape had Donald in a crowded room rudely shouted something *other* than Boo!? Or had he rudely *whispered* Boo! in a crowded room? Or had he rudely shouted Boo! in a room containing just him and the speaker? On the face of it, there would *not* have been a Boo sound in these scenarios. So on the face of it, the rudeness is egregiously weak $=_{\text{df}}$ up against a stronger, more natural, property that is better proportioned to the effect.[17]

FN:17

## DIAGNOSIS

A reason has been given for rejecting the locality argument: its conclusion is ~~is~~ at odds with a plausible-seeming analysis of causal relevance. But we have yet to find an actual mistake in the argument. Its central claim, that there are ~~extrinsic~~ factors in knowing too far away to affect the outcome, seems true enough. A factor in your knowing there is gold in the mine is that some misleading testimony given on the subject in Carson City was refuted by court records in Reno, all without your knowledge. The court records in Reno play no role in your continued digging here. But they are a factor in your knowing. And so there are factors in your knowing that make no difference to the event your knowing is said to cause. Doesn't this show that your knowing was more than the effect needed and so not its cause?

[17] There are other examples where the case for egregious weakness/strength is harder to make out. Sometimes a property escapes being egregiously weak only by incorporating "elsewhere" material irrelevant to the effect, and/or escapes being egregiously strong only omitting "elsewhere" material relevant to the effect. So, it might be that *loud-or-regretted and far from the mike*, although not egregiously weak or strong, would be egregiously weak if not for the *far*, and egregiously strong if not for the *regretted*. This suggests a refinement of the (PT) account. P has *egregiously weak aspects* iff some better-proportioned and no less natural P− is egregiously weak. P has *egregiously strong aspects* iff some better-proportioned and no less natural P+ is egregiously strong. A property is causally relevant iff (a) the effect depends on it, (b) it is not egregiously weak or strong, and (c) it has no egregiously weak or strong aspects, *Loud-or-regretted and far from the mike* violates (c) twice over. It has egregiously strong aspects, since *loud-or-regretted* is (i) better proportioned to the effect, (ii) no less natural, and (iii) egregiously weak. It has egregiously strong aspects, since *loud and far from the mike* is (i) better-proportioned to the effect, (ii) no less natural, and (iii) egregiously strong. This might be the right account for *behaving rudely* as well. It has egregiously strong aspects because for some suitable action-type $\varphi$, *rudely $\varphi$ing* is (i) better proportioned to the effect, (ii) no less natural, and (iii) egregiously strong since screened off by just $\phi$ing.

EXTRA #

wf make it look like the other phi's

That depends. “More than the effect needed” might mean that the knowing has elements without which the effect would still have occurred. In that sense, the knowing is more than the effect needed. In that sense, though, *every* cause is more than the effect needs. You can always find irrelevant aspects to take out. It is just that after a while, the gains in relevance are more than offset by the loss of naturalness. Concern about this puts a natural brake on the process of whittling away irrelevancies: a process that, left unchecked, would lead to every effect being blamed on the fact that things occurred sufficient for an effect just like that one.

To say that a proposed cause involved “more than the effect needed” might, on the other hand, mean that it involves irrelevancies that can be whittled away *with no loss of naturalness*: it might in other words mean that the proposed cause is *egregiously strong*. If that is what is meant, though, then it just does not follow, from the fact that your digging was insensitive to goings-on in Carson City and Reno, that your state of knowing the mine to contain gold was more than the effect needed. It might be that, although there are Carson-City-indifferent states that screen the knowing off from the effect, none of them is as natural as the knowing. It might be that the strongest *natural* weakening of the knowing is too weak to do the job itself.

Now we can see where the locality argument goes wrong. The problem does not lie with the assumption that behavior has intrinsic causes. Maybe it does and maybe it doesn’t; the defender of wide causation need not take a stand on this. Nor should we be bothered by the internalist’s claim that wide “causes” are liable to contain extrinsic irrelevancies. They *are* liable to contain extrinsic irrelevancies. The mistake is to run these thoughts too closely together, maintaining that they contain extrinsic irrelevancies *because* they make irrelevant additions to internalist-style intrinsic causes. Wide causes contain irrelevancies because all causes do, including narrow causes if such there be. The internalist’s larger mistake is to forget that proportionality is not pursued at all costs but traded off against naturalness. It seems clear ~~hardly open to doubt~~ that wide mental conditions effect *one* attractive such tradeoff—one local maximum of the relevant utility function. Internalists are welcome to search for a second local maximum more to their liking.

## REFERENCES

Braun, D. (1995) “Causally Relevant Properties”. *Philosophical Perspectives* 9: 447–75.

Jackson, F., and Pettit, P. (1988). “Functionalism and Broad Content”. *Mind* 97: 381–400.

Gibbons, J. (2001). “Knowledge in Action”. *Philosophy and Phenomenological Research* 62: 579–600.

Harman, G., and Thomson, J. J. (1996). *Moral Realism and Moral Objectivity.* Oxford: Blackwell.

Williamson, T. (1998). "The Broadness of the Mental: Some Logical Considerations". *Philosophical Perspectives* 12: pp. 389–410

____ (2000) *Knowledge and Its Limits*. New York: Oxford University Press.

Yablo, S. (1992). "Cause and Essence". *Synthese* 93: 403–49.

____ (1997). "Wide Causation". *Philosophical Perspectives* 11: 251–81; ch. 10 above.

____ (2002). "De Facto Dependence". *Journal of Philosophy* 99: 130–48.

## Queries in Chapter 11

Q1. We have captured this as Prose. Please check and confirm. looks good

Q2. Author correction is not clear. put "Williamson" for "Tim"