Prof. Dr. André Fuhrmann

Institut für Philosophie
Goethe-Universität Frankfurt

Winter 2016-17

Seminar: Die Philosophie von David Lewis

Dieser Reader enthält:

Mad pain and Martian pain (1980)
Reduction of mind (1994)

Elusive knowledge (1996)

Naming the colours (1997)

Mill and Milquetoast (1989)
The punishment that leaves something to chance (1989)
Dispositional theories of value (1998)

Are we free to break the laws? (1981)

(with Stephanie Lewis) Holes (1970)
(with Stephanie Lewis) Casati and Varzi on holes (1996)
Many, but almost one (1993)

Der Reader ist nur für den Seminargebrauch bestimmt.
Mad Pain and Martian Pain

David Lewis

I

There might be a strange man who sometimes feels pain, just as we do, but whose pain differs greatly from ours in its causes and effects. Our pain is typically caused by cuts, burns, pressure, and the like; his is caused by moderate exercise on an empty stomach. Our pain is generally distracting; his turns his mind to mathematics, facilitating concentration on that but distracting him from anything else. Intense pain has no tendency whatever to cause him to groan or writhe, but does cause him to cross his legs and snap his fingers. He is not in the least motivated to prevent pain or to get rid of it. In short, he feels pain but his pain does not at all occupy the typical causal role of pain. He would doubtless seem to us to be some sort of madman, and that is what I shall call him, though of course the sort of madness I have imagined may bear little resemblance to the real thing.

I said there might be such a madman. I don’t know how to prove that something is possible, but my opinion that this is a possible case seems pretty firm. If I want a credible theory of mind, I need a theory that does not deny the possibility of mad pain. I needn’t mind conceding that perhaps the madman is not in pain in quite the same sense that the rest of us are, but there had better be some straightforward sense in which he and we are both in pain.

Also, there might be a Martian who sometimes feels pain, just as we do, but whose pain differs greatly from ours in its physical realization. His hydraulic mind contains nothing like our neurons. Rather, there are varying amounts of fluid in many inflatable cavities, and the inflation of any one of these cavities opens some valves and closes others. His mental plumbing pervades most of his body—in fact, all but the heat exchanger inside his head. When you pinch his skin you cause no firing of C-fibers—he has none—but rather, you cause the inflation of many smallish cavities in his feet. When these cavities are inflated, he is in pain. And the effects of his pain are fitting: his thought and activity are disrupted, he groans and writhes, he is strongly motivated to stop you from pinching him and to see to it that you never do again. In short, he feels pain but lacks the bodily states that either are pain or else accompany it in us.

This paper was presented at a conference on mind-body identity held at Rice University in April 1978. I am grateful to many friends, and especially to Patricia Kitcher, for valuable discussions of the topic.
There might be such a Martian; this opinion too seems pretty firm. A credible theory of mind had better not deny the possibility of Martian pain. I needn't mind conceding that perhaps the Martian is not in pain in quite the same sense that we Earthlings are, but there had better be some straightforward sense in which he and we are both in pain.

II

A credible theory of mind needs to make a place both for mad pain and for Martian pain. Prima facie, it seems hard for a materialist theory to pass this two-fold test. As philosophers, we would like to characterize pain a priori. (We might settle for less, but let's start by asking for all we want.) As materialists, we want to characterize pain as a physical phenomenon. We can speak of the place of pain in the causal network from stimuli to inner states to behavior. And we can speak of the physical processes that go on when there is pain and that take their place in that causal network. We seem to have no other resources but these. But the lesson of mad pain is that pain is associated only contingently with its causal role, while the lesson of Martian pain is that pain is connected only contingently with its physical realization. How can we characterize pain a priori in terms of causal role and physical realization, and yet respect both kinds of contingency?

A simple identity theory straightforwardly solves the problem of mad pain. It goes just as straightforwardly wrong about Martian pain. A simple behaviorism or functionalism goes the other way: right about the Martian, wrong about the madman. The theories that fail our two-fold test so decisively are altogether too simple. (Perhaps they are too simple ever to have had adherents.) It seems that a theory that can pass our test will have to be a mixed theory. It will have to be able to tell us that the madman and the Martian are both in pain, but for different reasons: the madman because he is in the right physical state, the Martian because he is in a state rightly situated in the causal network.

Certainly we can cook up a mixed theory. Here's an easy recipe: First, find a theory to take care of the common man and the madman, disregarding the Martian—presumably an identity theory. Second, find a theory to take care of the common man and the Martian, disregarding the madman—presumably some sort of behaviorism or functionalism. Then disjoin the two: say that to be in pain is to be in pain either according to the first theory or according to the second. Alternatively, claim ambiguity: say that to be in pain in one sense is to be in pain according to the first theory, to be in pain in another sense is to be in pain according to the second theory.

This strategy seems desperate. One wonders why we should have a disjunctive or ambiguous concept of pain, if common men who suffer pain are always in pain according to both disjuncts or both disambiguations. It detracts from the credibility of a theory that it posits a useless complexity in our concept of pain—useless in application to the common man, at least, and therefore useless almost always.

I don't object to the strategy of claiming ambiguity. As you'll see, I shall defend a version of it. But it's not plausible to cook up an ambiguity *ad hoc* to account for the compossibility of mad pain and Martian pain. It would be better to find a widespread sort of ambiguity, a sort we would believe in no matter what we thought about pain, and show that it will solve our problem. That is my plan.

III

A dozen years or so ago, D. M. Armstrong and I (independently) proposed a materialist theory of mind that joins claims of type-type psychophysical identity with a behaviorist or functionalist way of characterizing mental states such as pain.² I believe our theory passes the
twofold test. Positing no ambiguity without independent reason, it provides natural senses in which both madman and Martian are in pain. It wriggles through between Scylla and Charybdis.

Our view is that the concept of pain, or indeed of any other experience or mental state, is the concept of a state that occupies a certain causal role, a state with certain typical causes and effects. It is the concept of a state apt for being caused by certain stimuli and apt for causing certain behavior. Or, better, of a state apt for being caused in certain ways by stimuli plus other mental states and apt for combining with certain other mental states to jointly cause certain behavior. It is the concept of a member of a system of states that together more or less realize the pattern of causal generalizations set forth in commonsense psychology. (That system may be characterized as a whole and its members characterized afterward by reference to their place in it.)

If the concept of pain is the concept of a state that occupies a certain causal role, then whatever state does occupy that role is pain. If the state of having neurons hooked up in a certain way and firing in a certain pattern is the state properly apt for causing and being caused, as we materialists think, then that neural state is pain. But the concept of pain is not the concept of that neural state. ("The concept of . . ." is an intensional functor.) The concept of pain, unlike the concept of that neural state which in fact is pain, would have applied to some different state if the relevant causal relations had been different. Pain might have not been pain. The occupant of the role might have not occupied it. Some other state might have occupied it instead. Something that is not pain might have been pain.

This is not to say, of course, that it might have been that pain was not pain and nonpain was pain; that is, that it might have been that the occupant of the role did not occupy it and some nonoccupant did. Compare: "The winner might have lost" (true) versus "It might have been that the winner lost" (false). No wording is entirely unambiguous, but I trust my meaning is clear.

In short, the concept of pain as Armstrong and I understand it is a nonrigid concept. Likewise the word "pain" is a nonrigid designator. It is a contingent matter what state the concept and the word apply to. It depends on what causes what. The same goes for the rest of our concepts and ordinary names of mental states.

Some need hear no more. The notion that mental concepts and names are nonrigid, wherefore what is pain might not have been, seems to them just self-evidently false. I cannot tell why they think so. Bracketing my own theoretical commitments, I think I would have no opinion one way or the other. It's not that I don't care about shaping theory to respect naive opinion as well as can be, but in this case I have no naive opinion to respect. If I am not speaking to your condition, so be it.

If pain is identical to a certain neural state, the identity is contingent. Whether it holds is one of the things that varies from one possible world to another. But take care. I do not say that here we have two states, pain and some neural state, that are contingently identical, identical at this world but different at another. Since I'm serious about the identity, we have not two states but one. This one state, this neural state which is pain, is not contingently identical to itself. It does not differ from itself at any world. Nothing does. What's true is, rather, that the concept and name of pain contingently apply to some neural state at this world, but do not apply to it at another. Similarly, it is a contingent truth that Bruce is our cat, but it's wrong to say that Bruce and our cat are contingently identical. Our cat Bruce is necessarily self-identical. What is contingent is that the nonrigid concept of being our cat applies to Bruce rather than to some other cat, or none.

IV

Nonrigidity might begin at home. All
actualities are possibilities, so the variety of possibilities includes the variety of actualities. Though some possibilities are thoroughly otherworldly, others may be found on planets within range of our telescopes. One such planet is Mars.

If a nonrigid concept or name applies to different states in different possible cases, it should be no surprise if it also applies to different states in different actual cases. Nonrigidity is to logical space as other relativities are to ordinary space. If the word "pain" designates one state at our actual world and another at a possible world where our counterparts have a different internal structure, then also it may designate one state on Earth and another on Mars. Or, better, since Martians may come here and we may go to Mars, it may designate one state for Earthlings and another for Martians.

We may say that some state occupies a causal role for a population. We may say this whether the population is situated entirely at our actual world, or partly at our actual world and partly at other worlds, or entirely at other worlds. If the concept of pain is the concept of a state that occupies that role, then we may say that a state is pain for a population. Then we may say that a certain pattern of firing of neurons is pain for the population of actual Earthlings and another for Martians.

A state occupies a causal role for a population, and the concept of occupant of that role applies to it, if and only if, with few exceptions, whenever a member of that population is in that state, his being in that state has the sort of causes and effects given by the role.

The thing to say about Martian pain is that the Martian is in pain because he is in a state that occupies the causal role of pain for Martians, whereas we are in pain because we are in a state that occupies the role of pain for us.

V

Now, what of the madman? He is in pain, but he is not in a state that occupies the causal role of pain for him. He is in a state that occupies that role for most of us, but he is an exception. The causal role of a pattern of firing of neurons depends on one's circuit diagram, and he is hooked up wrong.

His state does not occupy the role of pain for a population comprising himself and his fellow madmen. But it does occupy that role for a more salient population—mankind at large. He is a man, albeit an exceptional one, and a member of that larger population.

We have allowed for exceptions. I spoke of the definitive syndrome of typical causes and effects. Armstrong spoke of a state apt for having certain causes and effects; that does not mean that it has them invariably. Again, I spoke of a system of states that comes near to realizing commonsense psychology. A state may therefore occupy a role for mankind even if it does not at all occupy that role for some mad minority of mankind.

The thing to say about mad pain is that the madman is in pain because he is in the state that occupies the causal role of pain for the population comprising all mankind. He is an exceptional member of that population. The state that occupies the role for the population does not occupy it for him.

VI

We may say that X is in pain simpliciter if and only if X is in the state that occupies the causal role of pain for the appropriate population. But what is the appropriate population? Perhaps (1) it should be us; after all, it's our concept and our word. On the other hand, if it's X we're talking about, perhaps (2) it should
be a population that $X$ himself belongs to, and (3) it should preferably be one in which $X$ is not exceptional. Either way, (4) an appropriate population should be a natural kind—a species, perhaps.

If $X$ is you or I—human and unexceptional—all four considerations pull together. The appropriate population consists of mankind as it actually is, extending into other worlds only to an extent that does not make the actual majority exceptional.

Since the four criteria agree in the case of the common man, which is the case we usually have in mind, there is no reason why we should have made up our minds about their relative importance in cases of conflict. It should be no surprise if ambiguity and uncertainty arise in such cases. Still, some cases do seem reasonably clear.

If $X$ is our Martian, we are inclined to say that he is in pain when the cavities in his feet are inflated; and so says the theory, provided that criterion (1) is outweighed by the other three, so that the appropriate population is taken to be the species of Martians to which $X$ belongs.

If $X$ is our madman, we are inclined to say that he is in pain when he is in the state that occupies the role of pain for the rest of us; and so says the theory, provided that criterion (3) is outweighed by the other three, so that the appropriate population is taken to be mankind.

We might also consider the case of a mad Martian, related to other Martians as the madman is to the rest of us. If $X$ is a mad Martian, I would be inclined to say that he is in pain when the cavities in his feet are inflated; and so says our theory, provided that criteria (2) and (4) together outweigh either (1) or (3) by itself.

Other cases are less clear-cut. Since the balance is less definitely in favor of one population or another, we may perceive the relativity to population by feeling genuinely undecided. Suppose the state that plays the role of pain for us plays instead the role of thirst for a certain small subpopulation of mankind, and vice versa. When one of them has the state that is pain for us and thirst for him, there may be genuine and irresolvable indecision about whether to call him pained or thirsty—that is, whether to think of him as a madman or as a Martian. Criterion (1) suggests calling his state pain and regarding him as an exception; criteria (2) and (3) suggest shifting to a subpopulation and calling his state thirst. Criterion (4) could go either way, since mankind and the exceptional subpopulation may both be natural kinds. (Perhaps it is relevant to ask whether membership in the subpopulation is hereditary.)

The interchange of pain and thirst parallels the traditional problem of inverted spectra. I have suggested that there is no determinate fact of the matter about whether the victim of interchange undergoes pain or thirst. I think this conclusion accords well with the fact that there seems to be no persuasive solution one way or the other to the old problem of inverted spectra. I would say that there is a good sense in which the alleged victim of inverted spectra sees red when he looks at grass: he is in a state that occupies the role of seeing red for mankind in general. And there is an equally good sense in which he sees green: he is in a state that occupies the role of seeing green for him, and for a small subpopulation of which he is an unexceptional member and which has some claim to be regarded as a natural kind.

You are right to say either, though not in the same breath. Need more be said?

To sum up. Armstrong and I claim to give a schema that, if filled in, would characterize pain and other states a priori. If the causal facts are right, then also we characterize pain as a physical phenomenon. By allowing for exceptional members of a population, we associate pain only contingently with its causal role. Therefore we do not deny the possibility of mad pain, provided there is not too
much of it. By allowing for variation from one population to another (actual or merely possible) we associate pain only contingently with its physical realization. Therefore we do not deny the possibility of Martian pain. If different ways of filling in the relativity to population may be said to yield different senses of the word "pain," then we plead ambiguity. The madman is in pain in one sense, or relative to one population. The Martian is in pain in another sense, or relative to another population. (So is the mad Martian.)

But we do not posit ambiguity ad hoc. The requisite flexibility is explained simply by supposing that we have not bothered to make up our minds about semantic niceties that would make no difference to any commonplace case. The ambiguity that arises in cases of inverted spectra and the like is simply one instance of a commonplace kind of ambiguity—a kind that may arise whenever we have tacit relativity and criteria of selection that sometimes fail to choose a definite relatum. It is the same kind of ambiguity that arises if someone speaks of relevant studies without making clear whether he means relevance to current affairs, to spiritual well-being, to understanding, or what.

VII

We have a place for commonplace mad pain, Martian pain, and even mad Martian pain. But one case remains problematic. What about pain in a being who is mad, alien, and unique? Have we made a place for that? It seems not. Since he is mad, we may suppose that his alleged state of pain does not occupy the proper causal role for him. Since he is alien, we may also suppose that it does not occupy the proper role for us. And since he is unique, it does not occupy the proper role for others of his species. What is left?

(One thing that might be left is the population consisting of him and his unactualized counterparts at other worlds. If he went mad as a result of some improbable accident, perhaps we can say that he is in pain because he is in the state that occupies the role for most of his alternative possible selves; the state that would have occupied the role for him if he had developed in a more probable way. To make the problem as hard as possible, I must suppose that this solution is unavailable. He did not narrowly escape being so constituted that his present state would have occupied the role of pain.)

I think we cannot and need not solve this problem. Our only recourse is to deny that the case is possible. To stipulate that the being in this example is in pain was illegitimate. That seems credible enough. Admittedly, I might have thought offhand that the case was possible. No wonder; it merely combines elements of other cases that are possible. But I am willing to change my mind. Unlike my opinions about the possibility of mad pain and Martian pain, my naive opinions about this case are not firm enough to carry much weight.

VIII

Finally, I would like to try to preempt an objection. I can hear it said that I have been strangely silent about the very center of my topic. What is it like to be the madman, the Martian, the mad Martian, the victim of interchange of pain and thirst, or the being who is mad, alien, and unique? What is the phenomenal character of his state? If it feels to him like pain, then it is pain, whatever its causal role or physical nature. If not, it isn't. It's that simple!

Yes. It would indeed be a mistake to consider whether a state is pain while ignoring what it is like to have it. Fortunately, I have not made that mistake. Indeed, it is an impossible mistake to make. It is like the impossible mistake of considering whether a number is composite while ignoring the question of what factors it has.
Pain is a feeling. Surely that is uncontroversial. To have pain and to feel pain are one and the same. For a state to be pain and for it to feel painful are likewise one and the same. A theory of what it is for a state to be pain is inescapably a theory of what it is like to be in that state, of how that state feels, of the phenomenal character of that state. Far from ignoring questions of how states feel in the odd cases we have been considering, I have been discussing nothing else! Only if you believe on independent grounds that considerations of causal role and physical realization have no bearing on whether a state is pain should you say that they have no bearing on how that state feels.

Notes


3. The closest we can come is to have something at one world with twin counterparts at another. See my “Counterpart Theory and Quantified Modal Logic,” *Journal of Philosophy* 65 (1968): 113-126. That possibility is irrelevant to the present case.

4. Occurrent pain, that is. Maybe a disposition that sometimes but not always causes occurrent pain might also be called “pain.”
Lewis, David: Reduction of Mind

I am a realist and a reductive materialist about mind. I hold that mental states are contingently identical to physical — in particular, neural — states. My position is very like the ‘Australian materialism’ of Place, Smart, and especially Armstrong. Like Smart and Armstrong, I am an ex-Rylean, and I retain some part of the Rylean legacy. In view of how the term is contested, I do not know whether I am a ‘functionalist’. (See functionalism; identity theories; physicalism; Ryle.)

Supervenience and Analysis

My reductionism about mind begins as part of an a priori reductionism about every-thing. This world, or any possible world, consists of things which instantiate fundamental properties and which, in pairs or triples or ..., instantiate fundamental relations. Few properties are fundamental: the property of being a club or a tub or a pub, for instance, is an unnatural gerrymander, a condition satisfied by miscellaneous things in miscellaneous ways. A fundamental, or ‘perfectly natural’, property is the extreme opposite. Its instances share exactly some aspect of their intrinsic nature. Likewise for relations (see Lewis, 1983a and 1986a, pp. 59–69). I hold, as an a priori principle, that every contingent truth must be made true, somehow, by the pattern of co-instantiation of fundamental properties and relations. The whole truth about the world, including the mental part of the world, supervenes on this pattern. If two possible worlds were exactly isomorphic in their patterns of co-instantiation of fundamental properties and relations, they would thereby be exactly alike simpliciter (Lewis, 1992, p. 218). (See supervenience.)

It is a task of physics to provide an inventory of all the fundamental properties and relations that occur in the world. (That’s because it is also a task of physics to discover the fundamental laws of nature, and only the fundamental properties and relations may appear in the fundamental laws; see Lewis, 1983a, pp. 365–70). We have no a priori guarantee of it, but we may reasonably think that present-day physics goes a long way toward a complete and correct inventory. Remember that the physical nature of ordinary matter under mild conditions is very well understood (Feinberg, 1966). And we may reasonably hope that future physics can finish the job in the same distinctive style. We may think, for instance, that mass and charge are among the fundamental properties; and that whatever fundamental properties remain as yet undiscovered are likewise instantiated by very small things that come in very large classes of exact duplicates. We may further think that the very same fundamental properties and relations, governed by the very same laws, occur in the living and the dead.
parts of the world, and in the sentient and
the insentient parts, and in the clever and
the stupid parts. In short: if we optimisti­
cally extrapolate the triumph of physics
hitherto, we may provisionally accept that
all fundamental properties and relations
that actually occur are physical. This is the
thesis of materialism.

(It was so named when the best physics
of the day was the physics of matter alone.
Now our best physics acknowledges other
bearers of fundamental properties: parts of
pervasive fields, parts of causally active
spacetime. But it would be pedantry to
change the name on that account, and
disown our intellectual ancestors. Or worse,

If materialism is true, as I believe it is,
then the a priori supervenience of every­
thing upon the pattern of coinstantiation of
fundamental properties and relations yields
an a posteriori supervenience of everything
upon the pattern of coinstantiation of fun­
damental physical properties and relations.
Materialist supervenience should be a con­
tingent matter. To make it so, we supply a
restriction that makes reference to actuality.
Thus: if two worlds were physically iso­
morphic, and if no fundamental properties
or relations alien to actuality occurred in
either world, then these worlds would be
exactly alike simpliciter. Disregarding alien
worlds, the whole truth supervenes upon
the physical truth. In particular, the whole
mental truth supervenes. So here we have
the common core of all materialist theories
of the mind (Lewis, 1983a, pp. 361–5).

A materialist who stops here has already
said enough to come under formidable
attack. An especially well-focused version of
the attack comes from Frank Jackson
(1982). Mary, confined in a room where all
she can see is black or white, studies the
physics of colour and colour vision and
colour experience (and any other physics
you might think relevant) until she knows
it all. Then she herself sees colour for the
first time, and at last she knows what it’s
like to see colour. What is this knowledge
that Mary has gained? It may seem that she
has eliminated some possibilities left open
by all her previous knowledge; she has dis­
tinguished the actual world from other pos­
sible worlds that are exactly like it in all
relevant physical respects. But if materialist
supervenience is true, this cannot be what
happened. (See CONSCIOUSNESS; QUALIA.)

Materialists have said many things about
what does happen in such a case. I myself,
following Nemirow (1990), call it a case of
know-how: Mary gains new imaginative
abilities (Lewis, 1990). Others have said
that Mary gains new relations of acquaint­
ance, or new means of mental representa­
tion: or that the change in her is just that
she has now seen colour. These suggestions
need not be taken as rival alternatives. And
much ink has been spent on the question
whether these various happenings could in
any sense be called the gaining of ‘new
knowledge’, ‘new belief’, or ‘new informa­
tion’. But for a materialist, the heart of the
matter is not what does happen but what
doesn’t: Mary does not distinguish the actual
world from other worlds that are its physi­
cal duplicates but not its duplicates simpliciter.

Imagine a grid of a million tiny spots –

pixels – each of which can be made light or
dark. When some are light and some are
dark, they form a picture, replete with inter­
esting intrinsic gestalt properties. The case
evokes reductionist comments. Yes, the
picture really does exist. Yes, it really does
have those gestalt properties. However the
picture and the properties reduce to the
arrangement of light and dark pixels. They
are nothing over and above the pixels. They
make nothing true that is not made true
already by the pixels. They could go
unmentioned in an inventory of what there
is without thereby rendering that inventory
incomplete. And so on.

Such comments seem to me obviously
right. The picture reduces to the pixels. And
that is because the picture supervenes on
the pixels: there could be no difference in the
picture and its properties without some dif­
derence in the arrangement of light and dark

413
pixels. Further, the supervenience is asymmetric: not just any difference in the pixels would matter to the gestalt properties of the picture. And it is supervenience of the large upon the small and many. In such a case, say I, supervenience is reduction. And the materialist supervenience of mind and all else upon the arrangement of atoms in the void – or whatever replaces atoms in the void in true physics – is another such case.

Yet thousands say that what’s good about stating materialism in terms of supervenience is that this avoids reductionism! There’s no hope of settling this disagreement by appeal to some uncontested definition of the term ‘reductionism’. Because the term is contested, and the aim of some contestants is to see to it that whatever position they may hold, ‘reductionism’ shall be the name for something else.

At any rate, materialist supervenience means that for anything mental, there are physical conditions that would be sufficient for its presence, and physical conditions that would be sufficient for its absence. (These conditions will include conditions saying that certain inventories are complete: an electron has only so-and-so quantum numbers, for instance, and it responds only to such-and-such forces. But it’s fair to call such a condition ‘physical’, since it answers a kind of question that physics does indeed address.) And no matter how the world may be, provided it is free of fundamental properties or relations alien to actuality, a condition of the one sort or the other will obtain. For all we know so far, the conditions associated with a given mental item might be complicated and miscellaneous – even infinitely complicated and miscellaneous. But so long as we limit ourselves just to the question of how this mental item can find a place in the world of fundamental physics, it is irrelevant how complicated and miscellaneous the conditions might be.

It may seem unsatisfactory that physical conditions should always settle whether the mental item is present or absent. For mightn’t that sometimes be a vague question with no determinate answer? A short reply to this objection from vagueness is that if it did show that the mental was irremediable to fundamental physics despite supervenience, it would likewise show that boiling was irremediable to fundamental physics – which is absurd. For it is a vague matter just where simmering leaves off and boiling begins.

A longer reply has three parts. (1) If the physical settles the mental insofar as anything does, we still have materialist supervenience. Part of what it means for two physically isomorphic worlds to be just alike mentally is that any mental indeterminacy in one is exactly matched by mental indeterminacy in the other. (2) Whenever it is a vague question whether some simplistic mental classification applies, it will be determinate that some more subtle classification applies. What’s determinate may be not that you do love him or that you don’t, but rather that you’re in a certain equivocal state of mind that defies easy description. (3) If all indeterminacy is a matter of semantic indecision (Lewis, 1986a, pp. 212–13), then there is no indeterminacy in the things themselves. How could we conjure up some irreducible mental item just by failing to decide exactly which reducible item we’re referring to?

It may seem that when supervenience guarantees that there are physical conditions sufficient for the presence or absence of a given mental item, the sufficiency is of the wrong sort. The implication is necessary but not a priori. You might want to say, for instance, that black-and-white Mary really did gain new knowledge when she first saw colour; although what she learned followed necessarily from all the physics she knew beforehand, she had remained ignorant because it didn’t follow a priori.

A short reply to this objection from necessity a posteriori is that if it did show that the mental was irremediable to fundamental physics, it would likewise show that boiling was irremediable to fundamental physics – which is absurd. For the identity between boiling and a certain process described in fundamental physical terms is necessary a posteriori if anything is.
A longer reply, following Jackson (1992), is founded upon the ‘two-dimensional’ analysis of necessity a posteriori put forward by Stalnaker (1978), Davies and Humberstone (1980), and Tichý (1983). Two-dimensionalism says that there is no such thing as a necessary a posteriori proposition. However, one single sentence θ may be associated in two different ways with two different propositions, one of them necessary and the other one contingent; and the contingent one can be known only a posteriori. Suppose we choose to adopt a conception of meaning under which our conventions of language sometimes fix meanings only as a function of matters of contingent fact – for example, a conception on which the meaning of ‘boils’ is left dependent on which physical phenomenon turns out to occupy the boiling-role. Then if we interpret a sentence θ using the meanings of its words as fixed in world W₁, we get proposition H₁; using the meanings as fixed in W₂, we get H₂; and so on. Call these the propositions horizontally expressed by θ at the various worlds; and let H be the proposition horizontally expressed by θ at the actual world. The proposition diagonally expressed by θ at W is true at W. So if we know D, we know that θ horizontally expresses some truth or other, but we may not know which truth. Sentence θ is necessary a posteriori iff H is necessary but D is knowable only a posteriori. Likewise, a proposition P necessarily implies that θ iff P implies H; but P a priori implies that θ iff P implies D. Our worry was that when θ was about the mind, and P was a premise made true by fundamental physics, P might imply that θ necessarily but not a priori. But if so, and if you think it matters, just take another proposition Q: let Q be true at exactly those worlds where θ horizontally expresses the same proposition H that it actually does. Q is true. Given the materialist supervenience of everything, Q as well as P is made true by fundamental physics. P and Q together imply a priori that θ. So the gap between physical premises and mental conclusion is closed. Anyone who wants to reopen it – for instance, in order to square materialist supervenience with Mary’s supposed ignorance – must somehow show that the two-dimensional analysis of necessity a posteriori is inadequate.)

If we limit ourselves to the question how mind finds a place in the world of physics, our work is done. Materialist supervenience offers a full answer. But if we expand our interests a little, we’ll see that among the supervenient features of the world, mind must be very exceptional. There are countless such features. In our little toy example of the picture and the pixels, the supervenient properties number 2 to the power: 2 to the millionth power. In the case of materialist supervenience, the number will be far greater. The infinite cardinal beth-3 is a conservative estimate. The vast majority of supervenient features of the world are given only by miscellaneously infinite disjunctions of infinitely complex physical conditions. Therefore they are beyond our power to detect, to name, or to think about one at a time. Mental features of the world, however, are not at all beyond our ken. Finite assemblies of particles – us – can track them. Therefore there must be some sort of simplicity to them. Maybe it will be a subtle sort of simplicity, visible only if you look in just the right way. (Think of the Mandelbrot set: its overwhelming complexity, its short and simple recipe.) But somehow it must be there. Revealing this simplicity is a job for conceptual analysis.

Arbiters of fashion proclaim that analysis is out of date. Yet without it, I see no possible way to establish that any feature of the world does or does not deserve a name drawn from our traditional mental vocabulary. We should repudiate not analysis itself, but only some simplistic goals for it. We should allow for semantic indecision: any interesting analysandum is likely to turn out vague and ambiguous. Often the best that any one analysis can do is to fall safely within the range of indecision. And we should allow for semantic satisficing: analysis may reveal what it would take to
deserve a name perfectly, but imperfect
deservers of the name may yet deserve it
well enough. (And sometimes the perfect
case may be impossible.) If so, there is
bound to be semantic indecision about how
well is well enough.

I offer not analyses, but a recipe for ana­
lyses. We have a very extensive shared
understanding of how we work mentally.
Think of it as a theory: FOLK PSYCHOLOGY.
It is common knowledge among us; but it is
tacit, as our grammatical knowledge is. We
can tell which particular predictions and
explanations conform to its principles, but
we cannot expound those principles system­
atically. (Pace Lewis, 1972, p. 256, eliciting
the general principles of folk psychology is
no mere matter of gathering platitudes.)
Folk psychology is a powerful instrument of
prediction. We are capable of all sorts of
behaviour that would seem bizarre and
unintelligible, and this is exactly the beha­
vour that folk psychology predicts, rightly,
will seldom occur. (But we take a special
interest in questions that lie beyond the pre­
dictive power of folk psychology: wherefore
ingrates may fairly complain of a lack of
Interesting predictions!) Folk psychology has
evolved through thousands of years of close
observation of one another. It is not the last
word in psychology, but we should be con­
fident that so far as it goes – and it does go
far – it is largely right.

Folk psychology concerns the causal rela­
tions of mental states, perceptual stimuli,
and behavioural responses. It says how
mental states, singly or in combination, are
apt for causing behaviour; and it says how
mental states are apt to change under the
impact of perceptual stimuli and other
mental states. Thus it associates with each
mental state a typical causal role. Now we
have our recipe for analyses. Suppose we’ve
managed to elicit all the tacitly known
general principles of folk psychology.
Whenever M is a folk-psychological name for a
mental state, folk psychology will say that
the state M typically occupies a certain
causal role: call this the M-role. Then we
analyse M as meaning ‘the state that typi­
cally occupies the M-role’. Folk psychology
implicitly defines the term M, and we have
only to make that definition explicit.

Since the causal roles of mental states
involve other mental states, we might fear
circularity. The remedy is due in its essen­
tials to Ramsey (1931a, pp. 212–236) and
Carnap (1963, pp. 958–66); see also Lewis
(1970, 1972). Suppose, for instance, that
folk psychology had only three names for
mental states: L, M, N. We associate with
this triplet of names a complex causal role
for a triplet of states, including causal rela­
tions within the triplet: call this the LMN­
role. Folk psychology says that the states L,
M, N jointly occupy the LMN-role. That
implies that M occupies the derivative role:
coming second in a triplet of states that
jointly occupy the LMN-role. Taking this as
our M-role, we proceed as before. Say that
the names L, M, N are Interdefined. The
defining of all three via the LMN-role is a
package deal.

We might fear circularity for another
reason. The causal roles of mental states
involve responses to perceptual stimuli. But
the relevant feature of the stimulus will
often be some secondary quality – for
instance, a colour. We cannot replace the
secondary quality with a specification of the
stimulus in purely physical terms, on pain
of going beyond what is known to folk psy­
chology. But if we analyse the secondary
quality in terms of the distinctive mental
states its presence is apt to evoke, we close a
definitional circle. So we should take inter­
definition further. Let folk psychology
include folk psychophysics. This will say, for
instance, that the pair of a certain colour
and the corresponding sensation jointly
occupy a complex causal role that consists
in part, but only in part, of the former being
apt to cause the latter. Now we have a de­
rivative role associated with the name of the
colour, and another associated with the
name of the sensation: the role of coming
first or coming second, respectively, in a
pair that jointly occupies this complex role.

We might worry also about the beha­
vour that mental states are apt for causing.
Often we describe behaviour in a mentally
loaded way: as action. To say that you
kicked the ball to your teammate is to describe your behavior. But such a description presupposes a great deal about how your behavior was meant to serve your desires according to your beliefs; and also about the presence of the ball and the playing surface and the other player, and about social facts that unite players into teams. More threat of circularity? More need for interdefinition? I don't know how such further interdefinition would work; and anyway, it would be well to call a halt before folk psychology expands into a folk theory of the entire Lebenswelt!

Describing the behavior in purely physical terms — the angle of the knee, the velocity of the foot — would get rid of those presuppositions. But, just as in the case of the stimuli, it would go beyond what is known to folk psychology. Further, these descriptions would never fit the behavior of space aliens not of humanoid shape; and yet we should not dismiss out of hand the speculation that folk psychology might apply to aliens as well as to ourselves.

Fortunately there is a third way to describe behavior. When you kicked the ball, your body moved in such a way that if you had been on a flat surface in Earth-normal gravity with a suitably placed ball in front of you and a suitably placed teammate some distance away, then the impact of your foot upon the ball would have propelled the ball onto a trajectory bringing it within the teammate's reach. That description is available to the folk. They wouldn't give it spontaneously, but they can recognize it as correct. It presupposes nothing about your mental states, not even that you have any; nothing about whether the ball and the playing field and the gravity and the teammate are really there; nothing about your humanoid shape, except that you have some sort of foot. It could just as well describe the behavior of a mindless mechanical contraption, in the shape of a space alien (with a foot), thrashing about in free fall.

(I don't say that we should really use these 'if-then' descriptions of behavior. Rather, my point is that their availability shows how to unload the presuppositions from our ordinary descriptions.)

If M means 'the state that typically occupies the M-role' and if that role is only imperfectly occupied, what are we to do? — Satisfice: let the name M go to a state that deserves it imperfectly. And if nothing comes anywhere near occupying the M-role? — Then the name M has no referent. The boundary between the cases is vague. To take an example from a different term-inducing theory, I suppose it to be indeterminate whether 'dephlogisticated air' refers to oxygen or to nothing. But folk psychology is in far better shape than the phlogiston theory, despite scare stories to the contrary. We can happily grant that there are no perfect deservers of folk-psychological names, but we shouldn't doubt that there are states that deserve those names well enough.

What to do if the M-role, or the LMN-role, turns out to be doubly occupied? I used to think (Lewis, 1970, 1972) that in this case too the name M had no referent. But now I think it might be better, sometimes or always, to say that the name turns out to be ambiguous in reference. That follows the lead of Field (1973); and it is consistent with, though not required by, the treatment of Carnap (1963). Note that we face the same choice with phrases like 'the moon of Mars'; and in that case too I'd now lean toward ambiguity of reference rather than lack of it.

My recipe for analyses, like Rylean analytic behaviorism, posits analytic truths that constrain the causal relations of mental states to behavior. (We have no necessary connections between distinct existences, of course; the necessity is verbal. The state itself could have failed to occupy its causal role, but would thereby have failed to deserve its mental name.) But the constraints are weak enough to be credible. Because the state that typically occupies a role need not occupy it invariably, and also because a state may deserve a name well enough in virtue of a role that it occupies imperfectly, we are safe from the behaviorist's bugbears. We have a
place for the resolute deceiver, disposed come what may to behave as if his mental states were other than they really are. We have a place for the total and incurable paralytic with a rich mental life and no behavioural dispositions whatever. We even have a place for a madman whose mental states are causally related to behaviour and stimuli and one another in a totally haywire fashion (Lewis, 1980). And yet not anything goes. At some point — and just where that point comes is a matter of semantic indecision — weird tales of mental states that habitually offend against the principles of folk psychology stop making sense; because at some point the offending states lose all claim to their folk-psychological names. To that extent, analytic behaviourism was right. To quote my closest ally in these matters, '... outward physical behaviour and tendencies to behave do in some way enter into our ordinary concept of mind. Whatever theory of mind is true, it has a debt to pay, and a peace to be made, with behaviourism' (Armstrong, 1968, p. 68).

When we describe mental state M as the occupant of the M-role, that is what Smart (1959) calls a topic-neutral description. It says nothing about what sort of state it is that occupies the role. It might be a non-physical or a physical state, and if it is physical it might be a state of neural activity in the brain, or a pattern of currents and charges on a silicon chip, or the jangling of an enormous assemblage of beer cans. What state occupies the M-role and thereby deserves the name M is an a posteriori matter. But if materialist supervenience is true, and every feature of the world supervenes upon fundamental physics, then the occupant of the role is some physical state or other — because there's nothing else for it to be. We know enough to rule out the chip and the cans, and to support the hypothesis that what occupies the role is some pattern of neural activity. When we know more, we shall know what pattern of neural activity it is. Then we shall have the premises of an argument for psychophysical identification: mental state M = the occupant of the M-role (by analysis),

physical state P = the occupant of the M-role (by science),

therefore M = P.

(See Lewis, 1966, 1972; and see Armstrong, 1968, for an independent and simultaneous presentation of the same position, with a much fuller discussion of what the definitive causal roles might be.)

That's how conceptual analysis can reveal the simple formula — or anyway, the much less than infinitely complicated formula — whereby, when we know enough, we can pick out a mental feature of the world from all the countless other features of the world that likewise supervene on fundamental physics.

The causal-role analyses would still hold even if materialist supervenience failed. They might even still yield psychophysical identifications. Even if we lived in a spook-infested world, it might be physical states that occupied the causal roles (in us, if not in the spooks) and thereby deserved the folk-psychological names. Or it might be non-physical states that occupied the roles. Then, if we knew enough parapsychology, we would have the premises of an argument for psycho-non-physical identification.

When our argument delivers an identification M = P, the identity is contingent. How so? — All identity is self-identity, and nothing could possibly have failed to be self-identical. But that is not required. It's contingent, and it can only be known a posteriori, which physical (or other) states occupy which causal roles. So if M means 'the occupant of the M-role' it's contingent which state is the referent of M: it's contingent whether some one state is the common referent of M and P: so it's contingent whether M = P is true.

Kripke (1972) vigorously intuits that some names for mental states, in particular 'pain', are rigid designators: that is, it's not contingent what their referents are. I myself
intuit no such thing, so the non-rigidity imputed by causal-role analyses troubles me not at all.

Here is an argument that ‘pain’ is not a rigid designator. Think of some occasion when you were in severe pain, unmistakable and unignorable. All will agree, except for some philosophers and faith healers, that there is a state that actually occupies the pain role (or near enough); that it is called ‘pain’; and that you were in it on that occasion. For now, I assume nothing about the nature of this state, or about how it deserves its name. Now consider an unactualized situation in which it is some different state that occupies the pain role in place of the actual occupant; and in which you were in that different state; and which is otherwise as much like the actual situation as possible. Can you distinguish the actual situation from this unactualized alternative? I say not, or not without laborious investigation. But if ‘pain’ is a rigid designator, then the alternative situation is one in which you were not in pain, so you could distinguish the two very easily. So ‘pain’ is not a rigid designator.

Philosophical arguments are never incontrovertible – well, hardly ever. Their purpose is to help expound a position, not to coerce agreement. In this case, the controverter might say that if the actual occupant of the pain role is not a physical state, but rather is a special sort of non-physical state, then indeed you can distinguish the two situations. He might join me in saying that this would not be so if the actual occupant of the role were a physical state – else neurophysiology would be easier than it is – and take this together with intuitions of rigidity to yield a reductio against materialism. Myself, I don’t see how the physical or non-physical nature of the actual occupant of the role has anything to do with whether the two situations can be distinguished. Talk of ‘phenomenal character’ and the like doesn’t help. Either it is loaded with question-begging philosophical doctrine, or else it just reiterates the undisputed fact that pain is a kind of experience.

(The controverter just imagined would agree with the discussion in Kripke, 1972, pp. 344–42. But I don’t mean to suggest that Kripke would agree with him. At any rate, the words I have put into his mouth are not Kripke’s.)

If there is variation across worlds with respect to which states occupy the folk-psychological roles and deserve the folk-psychological names (and if this variation doesn’t always require differences in the laws of nature, as presumably it doesn’t) then also there can be variations within a single world. For possibility obeys a principle of recombination: roughly, any possible kind of thing can coexist with any other (Lewis, 1986a, pp. 86–92). For all we know, there may be variation even within this world. Maybe there are space aliens, and maybe there will soon be artificial intelligences, in whom the folk-psychological roles are occupied (or near enough) by states very different from any states of a human nervous system. Presumably, at least some folk-psychological roles are occupied in at least some animals, and maybe there is variation across species. There might even be variation within humanity. It depends on the extent to which we are hard-wired, and on the extent of genetic variation in our wiring.

We should beware, however, of finding spurious variation by overlooking common descriptions. Imagine two mechanical calculators that are just alike in design. When they add columns of numbers, the amount carried goes into a register, and the register used for this purpose is selected by throwing a switch. Don’t say that the carry-seventeen role is occupied in one machine by a state of register A and in the other by a state of register B. Say instead that in both machines alike the role is occupied by a state of the register selected by the switch. (Equivalently, by a state of a part of the calculator large enough to include the switch and both registers.) If there is a kind of thinking that some of us do in the left side of the brain and others do in the right side, that might be a parallel case.

If M means ‘the occupant of the M-role’ and there is variation in what occupies the
M-role, then our psychophysical identities need to be restricted: not plain \( M = P \), but \( M \text{-in-K} = P \), where \( K \) is a kind within which \( P \) occupies the M-role. Human pain might be one thing, Martian pain might be something else (Lewis, 1980). As with contingency, which is variation across worlds, so likewise with variation in a single world: the variability in no way infects the identity relation, but rather concerns the reference of the mental name.

The threat of variation has led many to retreat from 'type-type' to 'token-token' identity. They will not say that \( M = P \), where \( M \) and \( P \) are names for a state that can be common to different things at different times – that is, for a property had by things at times. But they will say that \( m \equiv p \), where \( m \) and \( p \) are mental and physical names for a particular, unrepeatable event. Token-token identities are all very well, in their derivative way, but the flight from type-type identities was quite unnecessary. For our restricted identities, of the form \( M \text{-in-K} = P \), are still type-type.

But don't we at least have a choice? Couldn't our causal role analyses be recast in terms of the causal roles of tokens, and if they were, would they not then yield token-token identities? After all, the only way for a type to occupy a causal role is through the causes and effects of its tokens. The effects of pain are the effects of pain-events. I think, following Jackson, Pargetter, and Prior (1982), that this recasting of the analyses would not be easy. There are more causal relations than one. Besides causing, there is preventing. It too may figure in folk-psychological causal roles: for instance, pain tends to prevent undivided attention to anything else. Prevention cannot straightforwardly be treated as a causal relation of tokens, because the prevented tokens do not exist – not in this world, anyway. It is better taken as a relation of types.

If a retreat had been needed, a better retreat would have been to 'subtype-subtype' identity. Let \( MK \) name the conjunctive property of being in state \( M \) and being of kind \( K \); and likewise for \( PK \). Do we really want psychophysical identities of the form \( MK = PK \)? – close, but I think not quite right. For one thing, \( M \text{-in-K} \) is not the same thing as \( MK \). The former but not the latter can occur also in something that isn't of kind \( K \). For another thing, it is \( P \) itself, not \( PK \), that occupies the M-role in things of kind \( K \).

Non-rigidity means that \( M \) is different states in different possible cases; variation would mean that \( M \) was different states in different actual cases. But don't we think that there is one property of being in the state \( M \) – one property that is common to all, actual or possible, of whatever kind, who can truly be said to be in state \( M \)? – There is. It is the property such that, for any possible \( X \), \( X \) has it just in case \( X \) is in the state that occupies the M-role for \( X \)'s kind at \( X \)'s world. (In Lewis, 1970, I called it the 'diagonalized sense' of \( M \).) The gerund 'being in \( M \)' can be taken, at least on one good disambiguation, as a rigid designator of this property. However, this property is not the occupant of the M-role. It cannot occupy that or any other causal role because it is excessively disjunctive, and therefore no events are essentially havings of it (Lewis, 1986c). To admit it as causally efficacious would lead to absurd double-counting of causes. It would be like saying that the meat fried in Footscray cooked because it had the property of being either fried in Footscray or boiled in Bundoora – only worse, because the disjunction would be much longer and more miscellaneous.

Since the highly disjunctive property of being in \( M \) does not occupy the M-role, I say it cannot be the referent of \( M \). Many disagree. They would like it if \( M \) turned out to be a rigid designator of a property common to all who are in \( M \). So the property I call 'being in \( M \)', they call simply \( M \); and the property that I call \( M \), the occupant of the M-role, they call 'the realization of \( M \)'. They have made the wrong choice, since it is absurd to deny that \( M \) itself is causally efficacious. Still, their mistake is superficial. They have the right properties in mind, even if they give them the wrong names.

It is unfortunate that this superficial
question has sometimes been taken to mark
the boundary of 'functionalism'. Sometimes
so and sometimes not - and that's why I
have no idea whether I am a fuctionalist.

Those who take 'pain' to be a rigid design­
ator of the highly disjunctive property will
need to controvert my argument that 'pain'
is not rigid, and they will not wish to claim
that one can distinguish situations in which
the pain-role is differently occupied. Instead,
they should controvert the first step, and
deny that the actual occupant of the pain­
role is called 'pain'. I call that denial a
reductio.

CONTENT

A mind is an organ of REPRESENTATION.
Many things are true according to it; that is,
they are believed. Or better, they are more
or less probable according to it; that is, they
are believed or disbelieved to varying
degrees. Likewise, many things are desired
to varying positive or negative degrees.
What is believed, or what is desired, we call
the CONTENT of BELIEF or DESIRE.

(I think it an open question to what extent
other states with content – doubting, won­
dering, fearing, pretending, . . . – require
separate treatment. and to what extent they
can be reduced to patterns in belief and desire
and contentless feeling. Be that as it may, I
shall ignore them here (see EMOTION).)

What determines the content of belief and
desire? – The occupation of folk-psycho­
logical roles by physical states, presumably
neural states; and ultimately the pattern of
coinstantiation of fundamental physical
properties and relations. But to say just that
is to say not much. Those who agree with it
can, and do, approach the problem of
content in very different ways.

I can best present the approach I favour
by opposing it to an alternative. A crude
sketch will suffice, so in fairness I name my
opponent Strawman. I doubt there is anyone
real who takes exactly the position that
Strawman does – but very many are to be
found in his near vicinity.

Strawman says that folk psychology says
– and truly – that there is a LANGUAGE OF
THOUGHT. It has words, and it has syntactic
constructions whereby those words can be
combined into sentences. Some of these sen­
tences have a special status. Strawman says
they are 'written in the belief box' or 'in the
desire box', but even Strawman doesn't take
that altogether literally. There are folk-psy­
chological causal roles for the words, for the
syntactic constructions, and for the belief
and desire boxes. It is by occupying these
roles that the occupants deserve their folk­
psychological names.

The question what determines content
then becomes the question: what deter­
mines the semantics of the language of
thought? Strawman says that folk psychol­
ogy specifies the semantic operations that
correspond to syntactic constructions such
as predication. As for the words. Strawman
says that folk psychology includes, in its
usual tacit and unsystematic way, a causal
theory of reference (more or less as in
Kripke, 1972). There are many relations of
acquaintance that connect the mind to
things, including properties and relations, in
the external world. Some are relations of
perceptual acquaintance. Others are less
direct: you are acquainted with the thing by
being acquainted with its traces. Often, you
are acquainted with the thing by way of its
linguistic traces that is, you have heard of it
by name. Somehow, in virtue of the differ­
cent causal roles of different words of the
language of thought, different words are
associated with different relations of
acquaintance, which connect them to differ­
ent external things. Whatever a word is
thus connected to is the referent of that
word.

Once the words of the language of
thought have their referents, the sentences
have their meanings. These are structures
built up from the referents of the words in a
way that mirrors the syntactic construction
of the sentences from the words. Take pre­
dication – Strawman's favourite example. A
word F of your language of thought is con­
nected by one relation of acquaintance to
the property of being French. Another word
A is connected by another relation of
acquaintance to the man André. (The first
relation might be linguistic, for instance, and the second perceptual.) The syntactic construction of predication builds a sentence $F(A)$. Its meaning is the ordered pair of the property of being French and André. Such a pair is a 'singular proposition', true just in case its second element instantiates its first. (Other singular propositions are triples, quadruples, ... , with relations in the first place.) If you have $F(A)$ written in your belief box, you thereby believe that André is French.

Strawman's account of content is sketchy, as I said it would be. Even with help from all his allies, I doubt he will find it easy to fill the gaps. I especially wonder what he can say about how the words get hooked up to the right relations of acquaintance. A causal theory of reference for public language might usefully mention mutual expectations among language-users, intentions to instill beliefs, semantic intentions, or other such instances of mental content. But even if we had corresponding expectations or intentions about our own language of thought, Strawman could not without circularity use them in a general account of mental content.

Suppose, all the same, that Strawman's account could be completed successfully by its own lights. I would still have four objections.

First, I don't believe that folk psychology says there is a language of thought. Rather, I think it is agnostic about how mental representation works – and wisely so.

What is the issue? Of course everybody should agree that the medium of mental representation is somehow analogous to language. A raven is like a writing-desk. Anything can be analogized to anything. And of course nobody thinks the head is full of tiny writing.

A serious issue, and one on which I take folk psychology to be agnostic, concerns the relation between the whole and the parts of a representation. Suppose I have a piece of paper according to which, *inter alia*, Collingwood is east of Fitzroy. Can I tear the paper up so that I get one snippet that has exactly the content that Collingwood is east of Fitzroy, nothing more and nothing less? If the paper is covered with writing, maybe I can: for maybe 'Collingwood is east of Fitzroy' is one of the sentences written there. But if the paper is a map, any snippet according to which Collingwood is east of Fitzroy will be a snippet according to which more is true besides. For instance, I see no way to lose the information that they are adjacent, and that a street runs along the border. And I see no way to lose all information about their size and shape.

(A hologram, or famously a connectionist network (see connectionism), differs even more from a paper covered with writing. If we make a hologram of the map and break it into snippets, detail will be lost in blur. But the arrangement of *all* the suburbs, provided it was shown with sufficient prominence on the original map, will remain to the last.)

Mental representation is language-like to the extent that parts of the content are the content of parts of the representation. If our beliefs are 'a map . . . by which we steer', as Ramsey said (1931b, p. 238), then they are to that extent not language-like. And to that extent, also, it is misleading to speak in the plural of beliefs. What is one belief? No snippet of a map is big enough that, determinately, something is true according to it, and also small enough that, determinately, nothing is true according to any smaller part of it. If mental representation is map-like (let alone if it is hologram-like) then 'beliefs' is a bogus plural. You have beliefs the way you have the blues, or the mumps, or the shivers.

But if mental representation is language-like, one belief is one sentence written in the belief box, so 'beliefs' is a genuine plural. Whether the plural is bogus or genuine is not settled by rules of grammar. Rather, it is an empirical question, and a question that folk psychology leaves open. 'The shivers' might be a parallel case. Is there such a thing as one shiver? – Maybe and maybe not. I don't think one cycle of vibration should be called 'one shiver', but there might be a better candidate. What if one
firing of a control neuron would set you shivering for four seconds. and prolonged shivering is caused by this neuron firing every two seconds? If so, I think the shivering set off by one firing could well be called 'one shiver', and then it is right to say that shivering consists of a sequence of overlapping shivers. Under this hypothesis, the plural is genuine. Under other hypotheses, the plural is bogus.

Of course you might say, under the hypothesis that mental representation is map-like, that any proposition true according to the mental map is one belief. Or you might say that the one belief that Collingwood is east of Fitzroy is the highly disjunctive state of having some mental map or other according to which that *inter alia* is true. Say so if you like. But I only insist that if you say either thing, then you may not also assume that 'one belief' is the sort of thing that can occupy a causal role. You may still say '... because he believes that Collingwood is east of Fitzroy', but only if you mean by it '... because he has beliefs' – bogus plural! – 'according to which *inter alia* Collingwood is east of Fitzroy'.

If Strawman heeds the advice of some of his allies, he will respond by changing his position. He will give away conceptual analysis and folk psychology, and market his wares as 'cognitive science' (see COGNITIVE PSYCHOLOGY). No problem, then, if the folk are agnostic about the language of thought. Let it be a new hypothesis, advanced because it best explains ... What? Well-known facts about belief? – But 'belief' is a folk-psychological name for a kind of state posited by folk psychology. If Strawman leaves all that behind him, where shall he find his evidence? He can never again set up thought experiments and ask us what we want to say about them. That would only elicit our folk-psychological preconceptions. He can make a fresh start if he really wants to – I assume he will not want to – but he cannot have his cake and eat it too. (See Jackson, 1992.)

If Strawman stands his ground, on the other hand, he will insist that folk psychology is far from agnostic about the language of thought. It has plenty to say, after all, about our 'concepts' (or 'ideas') of things. Our concept of a concept, says Strawman, is just our concept of a word of the language of thought. – I doubt it. I haven't much of any concept of Elsternwick. I have little idea what the place looks like, what sort of people live there, ... All I know is that there is a place of that name, and roughly where it is. But I *do* have the word. (At least, I have the word 'Elsternwick' of our public language. If I have a language of thought, presumably this word has been borrowed into it.) My lack of a concept isn't lack of a word; rather, I lack any very rich cluster of associated descriptions.

Strawman can reply that even if I haven't much concept of Elsternwick, still I have enough of one that I can think about Elsternwick (for instance, when I think how little I know about it). It is this minimal concept of a concept, he says, that is our folk-psychological concept of a word of the language of thought. – Yes, I have a concept of Elsternwick in the minimal sense that I have whatever it takes to be able to think about it. But must the basis of such an ability, in general or even in this case, be the possession of a word? On that question, the folk and I remain agnostic.

My second objection to Strawman's account is that it delivers only wide content. Which singular propositions you believe depends upon which external things are suitably connected by relations of acquaintance to the words of your language of thought.

Strawman holds that all content is wide because he has learned the lesson of TWIN EARTH. Recall the example (Putnam, 1975, pp. 139–42). Oscar the Earthling believes that water often falls from clouds. Twoscar on Twin Earth is in no way acquainted with water, that is, with H2O. Rather, Twoscar is acquainted with XYZ, a superficially similar liquid that is abundant on Twin Earth, in exactly the way that Oscar is acquainted with H2O. There is no other relevant difference between Twoscar and Oscar. We are invited to agree that Twoscar does not believe that water falls from clouds, and
believes instead that XYZ falls from clouds. Strawman does agree.

And so do I, but with many reservations. For one thing, I think agreement is not compulsory. Like any up-to-date philosopher of 1955, I think that 'water' is a cluster concept. Among the conditions in the cluster are: it is liquid, it is colourless, it is odourless, it supports life. But, pace the philosopher of 1955, there is more to the cluster than that. Another condition in the cluster is: it is a natural kind. Another condition is indexical: it is abundant hereabouts. Another is metalinguistic: many call it 'water'. Another is both metalinguistic and indexical: I have heard of it under the name 'water'. When we hear that XYZ off on Twin Earth fits many of the conditions in the cluster but not all, we are in a state of semantic indecision about whether it deserves the name 'water'. (See Unger, 1984, pp. 79–104. But while I agree with Unger about what happens in various cases, I don't endorse all the morals he draws.) When in a state of semantic indecision, we are often glad to go either way, and accommodate our own usage temporarily to the whims of our conversational partners (Lewis, 1979b). So if some philosopher, call him Schmutnam, invites us to join him in saying that the water on Twin Earth differs in chemical composition from the water here, we will happily follow his lead. And if another philosopher, Putnam (1975), invites us to say that the stuff on Twin Earth is not water – and hence that Twoscar does not believe that water falls from clouds – we will just as happily follow his lead. We should have followed Putnam's lead only for the duration of that conversation, then lapsed back into our accommodating state of indecision. But, sad to say, we thought that instead of playing along with a whim, we were settling a question once and for all. And so we came away lastingly misled.

The example half succeeds. It is not compulsory, but certainly it is permissible, to say that Oscar does believe that water falls from clouds and differently acquainted Twoscar does not. Therefore wide content does serve a purpose. It enters into the analysis of some sentences that are about belief, or at least partly about belief; or at least it does so under some permissible disambiguations of these sentences.

Other examples are similar. Twoscar is acquainted with molybdenum as Oscar is with aluminium; with a disease of bone as Oscar is with a disease of joints; with spy robots as Oscar is with cats; and so on. It seems to matter little whether Twoscar is our neighbour, or whether he lives on a remote planet, or whether he lives in a different possible world. In each case we find that the difference in what Twoscar and Oscar are acquainted with makes a difference to the truth value, under some disambiguation, of some sentences that are at least partly about belief. But that is all we find. There is nothing here to support Strawman's thesis that wide content is the only kind of content; or that it is in any way pre-eminent or basic.

We should not jump to the conclusion that just any belief sentence is susceptible to Twin Earth examples. Oscar thinks that square pegs don't fit round holes; I don't think you can tell an even halfway convincing story of how Twoscar, just by being differently acquainted, fails to think so too. Oscar believes there's a famous seaside place called 'Blackpool'; so does differently acquainted Twoscar, though of course it may not be Blackpool – not our Blackpool – that he has in mind. Oscar believes that the stuff he has heard of under the name 'water' falls from clouds. So does Twoscar – and so does Twoscar even if you alter not only his acquaintance with water but his relations of acquaintance to other things as well. You know the recipe for Twin Earth examples. You can follow it in these cases too. But what you get falls flat even as an example of how content is sometimes wide, let alone as evidence that content is always wide.

The famous brain in a bottle is your exact duplicate with respect to brain states and their typical causal roles; but is acquainted only with aspects of the computer that fabricates its virtual reality. You and the brain share no objects of acquaintance. So,
according to Strawman, you and the brain share no common beliefs whatever.

Newborn Swampman, just this moment formed by an unlikely chance assembly of atoms, also is your exact duplicate with respect *inter alia* to brain states and their typical causal roles (Davidson, 1987). But so far, he hasn’t had time to become acquainted with much of anything. Therefore, according to Strawman, he believes not much of anything.

Strawman and his allies may think that we have here two remarkable philosophical discoveries. I think, rather, that Strawman’s thesis that all content is wide has here met with a twofold *reductio ad absurdum*. Granted, the brain in a bottle shares no wide content of belief with you. Granted, Swampman has no wide content of belief at all. Yet there must be some good sense in which both the brain and Swampman are your mental twins; some good sense in which they believe just what you do. (And in our less extreme cases, there must be some good sense in which Twoscar believes just what Oscar does.) Strawman’s position is unacceptable. Not because it posits wide content; but because it omits narrow content, content independent of what one is acquainted with. It omits the sort of content that you and the brain and Swampman, and likewise Oscar and Twoscar, have in common.

(Narrow content is independent of what you are acquainted with, but that does not mean that it is altogether intrinsic to you. For it still depends on the causal roles of your brain states; causation depends on the laws of nature; and if some sort of regularity theory of lawhood is true, living under such-and-such laws is not intrinsic to you. Further, it is the typical causal roles of your brain states that matter. But you may be an atypical member of your kind; hence what is typical of your kind is not intrinsic to you. So I can say only this: if X and Y are intrinsic duplicates, and if they live under the same laws of nature, and if they are the same in kind, then they must be exactly alike in narrow content.)

In insisting on the existence of narrow content, I am not guided by any preconcep-

tion about what sort of properties may figure in causal explanation, or in truly scientific explanation. I dare say the fundamental laws of physics must concern perfectly natural, intrinsic properties. But that’s irrelevant, since causal and scientific explanation seldom consists in subsumption under these fundamental laws. Rather, it is a matter of giving information about how things are caused (Lewis, 1986b). Such information can come in many forms, both within science and without, and there is no reason to proscribe extrinsic classifications. (Lynne Baker told me a nice example: the science of economics is all about extrinsic properties like poverty and debt. Yet there is nothing wrong, and nothing unscientific, in saying that Fred stays poor because of his burden of debt.)

I am guided, rather, by my tacit mastery of the principles of folk psychology. I said: Oscar believes that the stuff he has heard of under the name ‘water’ falls from clouds; and so does Twoscar. (And so do you, and so does the brain in a bottle, and so does Swampman.) These are ordinary folk-psychological belief sentences; but narrow ones, as witness the fact that they are not susceptible to Twin Earth examples.

This narrow content is content, rightly so-called: something is true according to the belief-system in question. The content is true on condition that the stuff the believer has heard of under the name ‘water’ does indeed fall from clouds; otherwise false. It is not ‘purely syntactic content’ – something I take to be a contradiction in terms. Nor is it a mere function that delivers genuine content as output when given circumstances of acquaintance as input. Nor is it merely phenomenalistic content, restricted in subject matter to the believer’s experience.

However, it is not content that can be given by a singular proposition, and that leads to my third objection against Strawman’s account.

Strawman’s singular propositions suffice to specify which things have which properties. If all else supervenes upon the pattern
of coinstantiation of fundamental properties, that in turn will suffice to specify the way the world is. But much of the content of our knowledge and belief is *de se*: it concerns not the world but oneself. (See Perry, 1977; Lewis, 1979a; Chisholm, 1979.) However much I may know about the things that make up the world, their properties and their arrangement, it is something extra to know which one of all these things is *me*. This is *de se* knowledge, whereby I locate myself in the world and self-ascribe the properties I think myself to possess, but is not knowledge of how the world is. Its content cannot be captured by singular propositions. What singular proposition is expressed when I say, or I think, ‘I am DL’? – Just the proposition that DL = DL. And when I self-ascribe the property F? – Just the proposition that DL is F. But I can know these propositions without knowing who I am, or whether I am F. (And you can know them too.) Strawman’s only recourse is to say that *de se* knowledge is characterized not by its *de se* content but some other way – and if he says that, he confesses that his account of content is inadequate. Belief that falls short of knowledge can likewise have *de se* content. If you take yourself to be DL, your false belief and my true belief have their *de se* content in common. Desire also has *de se* content. If you desire to be F and I believe myself to be F, again the two attitudes have their *de se* content in common.

There is also tensed content. The world is spread out over many times; but we can have knowledge, or belief or desire, about which of these times is now. Again, this is not knowledge of how the whole spread-out world is. It is something extra. Some would speak of content *de se et nunc*, but I would subsume *de nunc* under *de se*. For I think we persist through time by consisting of many time-slices, or momentary selves; and in the last analysis, it is these momentary selves that do our thinking. So when I think ‘It’s now time for lunch’, that’s one of my momentary selves self-ascribing *de se* the property of being located at lunchtime.

The ‘propositions’, if we may call them that, which make up *de se* content are true or false not absolutely, as singular propositions are, but relative to a subject. (Or to a subject at a time, if you don’t believe in momentary selves.) The content of my knowledge *de se* that I am DL is something that is true for me but not for you. Its linguistic expression requires a first-person pronoun, or some equivalent device. We could call it an ‘egocentric proposition’ (or ‘egocentric and tensed’). Or we can simply identify it with the property that I self-ascribe: the property of being DL. Likewise the *de se* content of my belief that I have F is just the property of F itself; the *de se* content of my belief that it’s lunchtime is the property (possessed not by the whole of me but by some of my momentary selves) of being located at lunchtime; and so on. A *de se* self-ascription of a property is true just on condition that the self-ascriber possesses the self-ascribed property.

(May I say, then, that *de se* belief has ‘truth conditions’? Not if Strawman has his way. He goes in for terminological piracy. He transforms one term after another into a mere synonym for ‘singular proposition’. He has taken ‘object of thought’. He has taken ‘content’. He has taken ‘proposition’. He is well on the way to taking ‘truth condition’. When he has taken all the terms for his own, dissident thoughts will be unsayable.) Since Strawman has no place for *de se* content, it makes sense that he overlooks narrow content as well. For narrow content is very often *de se*. To revisit our previous example: Oscar self-ascribes having heard, under the name of ‘water’, of a liquid that falls from clouds. He also self-ascribes the property of being at a place (and time) in the vicinity of which the most abundant liquid is one that falls from clouds. Differently acquainted Twoscar self-ascribes these same two properties, and in this way Oscar and Twoscar share the same *de se* narrow content of belief.

On my own view, it is just such *de se* narrow content that underlies wide content. The semantics of the alleged language of thought needn’t enter into it. To the extent that language enters my story at all, it is not by way of the language of thought, but
rather by way of thought about language – about the ordinary public language, whereby, for instance, Oscar heard of something under the name 'water'.

Here is one recipe (Lewis, 1979a, pp. 538-43): if R is a relation of acquaintance, and subject S self-ascribes being R-acquainted uniquely with something that has property F (the narrow part), and if S is R-acquainted uniquely with A (the wide part), S thereby widely believes the singular proposition that A has F. There are variants on the recipe. Our example of French André was a case in which property F as well as individual A enters indirectly as an object of acquaintance; we must of course let in cases where the property F gives way to a relation with two or more relata; maybe sometimes we should drop the qualification 'uniquely'; and maybe sometimes the relation R is not, or not entirely, a matter of acquaintance. But in every case, wide belief in a singular proposition derives from narrow de se self-ascription plus facts about what the subject is related to.

Often we know a lot about which singular propositions someone believes in this wide and derivative way; but we know less about how – in virtue of just which self-ascriptions and relations of acquaintance – he believes those singular propositions. So it's no surprise to find that our ordinary-language belief sentences often seem to be ascriptions of wide content. Often; but not always. In these last few paragraphs I've been talking about de se narrow content, and I've been talking about it in plain English. (Such bits of jargon as I used were first explained in plain English.)

There are still other dimensions to the semantic complexity and the multifarious ambiguity of ordinary-language belief sentences. Think of the belief sentences that show up as test cases in articles advocating one semantic analysis or another. I always want to say: 'in a sense that's true, in a sense false'. One complication is that we get direct-quotational effects even in what is ostensibly indirect quotation (see Rieber, 1992). An example: Fred knows perfectly well that the house he lives in is made of wood, but Fred also thinks that 'abode' is the English word for a house made of mud-brick. 'Fred believes that he has an abode – yes or no?' In at least some contexts (this isn't one of them) I'd be prepared to insist on 'no'. Wouldn't you? Moral: if you hope to understand the folk psychology of belief by studying the linguistic phenomenology of ordinary belief sentences, you're in for big trouble.

I've said that narrow content is very often de se, but by resorting to a cheap trick I can change 'often' to 'always'. Take an apparent exception: the narrow belief that square pegs won't fit in round holes. Take this to be the de se self-ascription of the property of inhabiting a world wherein square pegs won't fit in round holes. A peculiar property, since either all the inhabitants of the world share it or else none do; and, like many other self-ascribed properties, very far from fundamental: but in a broad enough sense of the word, a property all the same. Likewise you can self-ascribe the property of inhabiting a world where there's a famous seaside place called 'Blackpool'. And so on, until all narrow content has been included as de se. Hoky, but maybe worthwhile for the sake of uniform treatment.

My final objection is that Strawman ignores large parts of the folk psychology of belief and desire: the parts that characterize aspects of our rationality. Folk psychology says that a system of beliefs and desires tends to cause behaviour that serves the subject's desires according to his beliefs. Folk psychology says that beliefs change constantly under the impact of perceptual evidence: we keep picking up new beliefs, mostly true, about our perceptual surroundings; whereupon our other beliefs (and our instrumental desires) change to cohere with these new beliefs. Folk psychology sets presumptive limits to what basic desires we can have or lack: de gustibus non disputandum, but still a bedrock craving for a saucer of mud would be unintelligible (Anscombe, 1958, pp. 69–71). Likewise it sets limits to our sense of plausibility: which hypotheses we find credible prior to evidence, hence
which hypotheses are easily confirmed when their predictions come true. And it sets presumptive limits on what our contents of belief and desire can be. Self-ascribed properties may be 'far from fundamental', I said – but not too far. Especially gruesome gerrymanders are prima facie ineligible to be contents of belief and desire. (See Lewis, 1983a, pp. 370–7; Lewis, 1986a, pp. 38–9 and 105–8.) In short, folk psychology says that we make sense. It credits us with a modicum of rationality in our acting, believing, and desiring.

(Beware. 'Rationality' is an elastic word, and here I've stretched it to cover a lot. If you'd rather use it more narrowly – just for the serving of desires according to beliefs, say – no harm done. So long as you don't just ignore the several other departments of rationality that I listed, it doesn't matter what you call them.)

If mental states are to be analysed as occupants of folk-psychological roles, and if the folk psychology of belief and desire has a lot to say about rationality, and if what it says is framed in terms of content, then it seems that constraints of rationality are constitutive of content. Yet Strawman's account of content makes no place for constitutive rationality. Why not?

Perhaps Strawman thought, wisely, that it would be better to say too little than too much. It wouldn't do to conclude that, as a matter of analytic necessity, anyone who can be said to have beliefs and desires at all must be an ideally rational homo economicus! Our rationality is very imperfect, Strawman knows it, and he knows that the folk know it too. Of course we overlook options and hypotheses, we practice inference to the third-best explanation, we engage in double think, and so on, and on, and on.

But there is no cause for alarm. Folk psychology can be taken as a theory of imperfect, near-enough rationality, yet such rationality as it does affirm can still be constitutive. And even if folk psychology did set too high a standard – even if, to take the worst case, it were a theory of ideal rationality – still an imperfect but near-enough occupant of a folk-psychological role could thereby be an imperfect but near-enough deserver of a folk-psychological name. Remember also that the typical occupant of a role needn't occupy it in every case. In short, constitutive rationality leaves plenty of room for human folly.

(I think that systematic theories of ideal rationality – decision theory, for instance, and the theory of learning from experience by conditionalizing a subjective probability distribution – are severely idealized versions of parts of folk psychology. They are founded upon our tacit knowledge of folk psychology, elicited in the guise of 'intuition'. But folk psychology also supplies the grains of salt to be applied to these idealizations. Sometimes it supplies complementary pairs of opposite idealizations: a quantitative theory of subjective probabilities and utilities precise to however many decimal places, and alongside it a non-quantitative theory of beliefs and desires that don't admit of degree at all.)

Constitutive rationality is part of the legacy of behaviourism, and that is a second reason why Strawman mistrusts it. A behaviourist analysis might say, roughly, that a subject's beliefs and desires are those beliefs and desires, attribution of which would best make sense of how the subject is disposed to behave, and of how his changing behavioural dispositions depend on the changing perceptible features of his surroundings. But Strawman is a robust realist about beliefs and desires. He takes them to be genuine inner states, and causes of behaviour. He won't like an analysis that dispenses with efficacious inner states in favour of mere patterns of dispositions. Still less would he like it if the behaviourist went on to say that attributions of belief and desire governed by constitutive rationality were instrumentally useful, or warranted by rules of assertability, but not straightforwardly true.

I applaud these misgivings. I too am a robust realist about beliefs and desires. (About whole systems of beliefs and desires, anyway, though maybe not about all the little snippets – the sentences written in the belief and desire boxes – of which these systems may or may not be composed.) But
I say the proper remedy is not to shun constitutive rationality, but to apply it differently. The behaviourist applies it directly to the subject: I say we should apply it to the subject’s inner state. The behaviourist says that the subject has that system of beliefs and desires that best makes sense of how the subject is disposed to behave. Whereas I’d say that the inner state is that system of beliefs and desires that best makes sense of the behaviour which that state is apt for causing in subjects. Thus I’d use constitutive rationality not to dispense with causally efficacious inner states, but rather to define their content.

A third reason why Strawman shuns constitutive rationality is that sometimes it needs to be applied not to the singular propositions that are the wide content of belief and desire, but instead to the underlying de se narrow content. The furniture of the Lebenswelt which presents us with our problems of decision and learning consists, in the first instance, of objects given qua objects of acquaintance, and individuated by acquaintance. (See Hintikka, 1972; Lewis, 1983b.) That is a matter of narrow content. If you are lucky, and you’re never wrong or uncertain about whether you’re really R-acquainted with something, and you’re never wrong or uncertain about whether the thing you’re R1-acquainted with is or isn’t the same as the thing you’re R2-acquainted with, then we can talk about your beliefs and desires entirely in terms of wide content. We can safely let things simpliciter stand in for things-qua-objects-of-acquaintance. But if you’re not so lucky, that won’t work. Take unlucky Pierre (Kripke, 1979). He self-ascribes being R1-acquainted with a pretty city and being R2-acquainted with an ugly city. But in fact he is R1-acquainted and R2-acquainted with the same city, London. Thereby he believes both that London is pretty and that London is ugly. (Kripke derives this conclusion from certain premises, but I find the conclusion at least as obvious as the premises.) I take this to be a conflict in wide content: Pierre widely believes two singular propositions that predicate conflicting properties of the same thing. Folk psychology says that by careful attention we can detect and eliminate conflicts in our beliefs – especially if we’re good at logic, as Pierre is. But plainly that was never meant to apply to Pierre’s conflict of singular propositions. Mere thought can’t save him. What he needs is the information de se that he is R1-acquainted and R2-acquainted with the very same thing. (See Lewis, 1981.)

And suppose Pierre believes that by boarding the bus before him, he can be taken to London for a week of sight-seeing. Would boarding that bus serve his desires according to his beliefs? It helps not at all to know that he widely believes both that London is pretty and that London is ugly. What does help is the information already given about his narrow self-ascriptions, plus one further thing: he also self-ascribes having a bus before him that would take him to the place he is R1-acquainted with. (See Lewis, 1986a, p. 58.)

(On constitutive rationality, see Stalnaker, 1984, pp. 1–42; Lewis, 1974, 1986a, pp. 27–40. But see Lewis, 1974, with caution: it began as a conversation with Donald Davidson, and I went rather too far in granting undisputed common ground. (1) I gave an important place to the subject Karl’s beliefs as expressed in Karl’s own language; that certainly suggests language-of-thoughtism, though I hope I committed myself to nothing more than the safe thesis that Karl’s medium of mental representation is somehow analogous to language. (2) I was too individualistic: I ignored the possibility that deviant Karl might believe something in virtue of the causal role of his inner state not in Karl himself but in others who are more typical members of Karl’s kind. (3) I had not yet come to appreciate the role of de se content. Also see Lewis, 1986a, pp. 27–40, with caution: besides endorsing constitutive rationality, I also stated it within a controversial framework of realism about unactualized possibilia. I still think that’s a good way to state it; but I never said it was the only way. Constitutive rationality and realism about possibilia needn’t be a package deal!)

429
This completes my list of objections against Strawman’s program for explaining content. Doubtless you can think of ever so many ways of amending Strawman’s theses to get around my objections. Some lists of amendments would take us to the positions really held by real people. Of course I can’t show that no version of Strawman-amended can work. But for myself, I pin my hopes on a more radical reversal of Strawman’s position.

With Strawman for a foil, my own approach can be summed up quickly. The contentful unit is the entire system of beliefs and desires. (Maybe it divides up into contentful snippets, maybe not.) That system is an inner state that typically causes behaviour, and changes under the impact of perception (and also spontaneously). Its content is defined, insofar as it is defined at all, by constitutive rationality on the basis of its typical causal role. This content is in the first instance narrow and de se (or de se et nunc if you’d rather steer clear of momentary selves). Wide content is derivative, a product of narrow content and relationships of acquaintance with external things.

See also Davidson; Dennett; Fodor; Philosophy and Psychology; Propositional Attitudes; Reasons and Causes; Thoughts; Thought and Language.

BIBLIOGRAPHY


—— 1983a. New work for a theory of uni-


Thanks to the Boyce Gibson Memorial Library and the philosophy department of Birkbeck College; and to the editors, Ned Block, Alex Byrne, Mark Crimmins, Allen Hazen, Ned Hall, Elijah Millgram, Thomas Nagel, and especially Frank Jackson.

DAVID LEWIS
ELUSIVE KNOWLEDGE

David Lewis

We know a lot. I know what food penguins eat. I know that phones used to ring, but nowadays squeal, when someone calls up. I know that Essendon won the 1993 Grand Final. I know that here is a hand, and here is another.

We have all sorts of everyday knowledge, and we have it in abundance. To doubt that would be absurd. At any rate, to doubt it in any serious and lasting way would be absurd; and even philosophical and temporary doubt, under the influence of argument, is more than a little peculiar. It is a Moorean fact that we know a lot. It is one of those things that we know better than we know the premises of any philosophical argument to the contrary.

Besides knowing a lot that is everyday and trite, I myself think that we know a lot that is interesting and esoteric and controversial. We know a lot about things unseen: tiny particles and pervasive fields, not to mention one another’s underwear. Sometimes we even know what an author meant by his writings. But on these questions, let us agree to disagree peacefully with the champions of ‘post-knowledgeism’. The most trite and ordinary parts of our knowledge will be problem enough.

For no sooner do we engage in epistemology – the systematic philosophical examination of knowledge – than we meet a compelling argument that we know next to nothing. The sceptical argument is nothing new or fancy. It is just this: it seems as if knowledge must be by definition infallible. If you claim that S knows that P, and yet you grant that S cannot eliminate a certain possibility in which not-P, it certainly seems as if you have granted that S does not after all know that P. To speak of fallible knowledge, of knowledge despite uneliminated possibilities of error, just sounds contradictory.

Blind Freddy can see where this will lead. Let your paranoid fantasies rip – CIA plots, hallucinogens in the tap water, conspiracies to deceive, old Nick himself – and soon you find that uneliminated possibilities of error are everywhere. Those possibilities of error are far-fetched, of course, but possibilities all the same. They bite into even our most everyday knowledge. We never have infallible knowledge.

Never – well, hardly ever. Some say we have infallible knowledge of a few simple, axiomatic necessary truths; and of our own present experience. They say that I simply cannot be wrong that a part of a part of something is itself a part of that thing; or that it seems to me now (as I sit here at the keyboard) exactly as if I am hearing clicking noises on top of a steady whirring. Some say so. Others deny it. No matter; let it be granted, at least for the sake of the argument. It is not nearly enough. If we have only that much

Thanks to many for valuable discussions of this material. Thanks above all to Peter Unger; and to Stewart Cohen, Michael Devitt, Alan Hajek, Stephen Hetherington, Denis Robinson, Ernest Sosa, Robert Stalnaker, Jonathan Vogel, and a referee for this Journal. Thanks also to the Boyce Gibson Memorial Library and to Ormond College.
infallible knowledge, yet knowledge is by definition infallible, then we have very little knowledge indeed – not the abundant everyday knowledge we thought we had. That is still absurd.

So we know a lot; knowledge must be infallible; yet we have fallible knowledge or none (or next to none). We are caught between the rock of fallibilism and the whirlpool of scepticism. Both are mad!

Yet fallibilism is the less intrusive madness. It demands less frequent corrections of what we want to say. So, if forced to choose, I choose fallibilism. (And so say all of us.) We can get used to it, and some of us have done. No joy there – we know that people can get used to the most crazy philosophical sayings imaginable. If you are a contented fallibilist, I implore you to be honest, be naive, hear it afresh. ‘He knows, yet he has not eliminated all possibilities of error.’ Even if you’ve numbed your ears, doesn’t this overt, explicit fallibilism still sound wrong?

Better fallibilism than scepticism; but it would be better still to dodge the choice. I think we can. We will be alarmingly close to the rock, and also alarmingly close to the whirlpool, but if we steer with care, we can – just barely – escape them both.

Maybe epistemology is the culprit. Maybe this extraordinary pastime robs us of our knowledge. Maybe we do know a lot in daily life; but maybe when we look hard at our knowledge, it goes away. But only when we look at it harder than the sane ever do in daily life; only when we let our paranoid fantasies rip. That is when we are forced to admit that there always are uneliminated possibilities of error, so that we have fallible knowledge or none.

Much that we say is context-dependent, in simple ways or subtle ways. Simple: ‘it’s evening’ is truly said when, and only when, it is said in the evening. Subtle: it could well be true, and not just by luck, that Essendon played rottenly, the Easybeats played brilliantly, yet Essendon won. Different contexts evoke different standards of evaluation. Talking about the Easybeats we apply lax standards, else we could scarcely distinguish their better days from their worse ones. In talking about Essendon, no such laxity is required. Essendon won because play that is rotten by demanding standards suffices to beat play that is brilliant by lax standards.

Maybe ascriptions of knowledge are subtly context-dependent, and maybe epistemology is a context that makes them go false. Then epistemology would be an investigation that destroys its own subject matter. If so, the sceptical argument might be flawless, when we engage in epistemology – and only then!1

If you start from the ancient idea that justification is the mark that distinguishes knowledge from mere opinion (even true opinion), then you well might conclude that ascriptions of knowledge are context-dependent because standards for adequate justification are context-dependent. As follows: opinion, even if true, deserves the name of

---

1 The suggestion that ascriptions of knowledge go false in the context of epistemology is to be found in Barry Stroud, ‘Understanding Human Knowledge in General’ in Marjorie Clay and Keith Lehrer (eds.), *Knowledge and Skepticism* (Boulder: Westview Press, 1989); and in Stephen Hetherington, ‘Lacking Knowledge and Justification by Theorising About Them’ (lecture at the University of New South Wales, August 1992). Neither of them tells the story just as I do, however it may be that their versions do not conflict with mine.
knowledge only if it is adequately supported by reasons; to deserve that name in the especially demanding context of epistemology, the arguments from supporting reasons must be especially watertight; but the special standards of justification that this special context demands never can be met (well, hardly ever). In the strict context of epistemology we know nothing, yet in laxer contexts we know a lot.

But I myself cannot subscribe to this account of the context-dependence of knowledge, because I question its starting point. I don’t agree that the mark of knowledge is justification. First, because justification is not sufficient: your true opinion that you will lose the lottery isn’t knowledge, whatever the odds. Suppose you know that it is a fair lottery with one winning ticket and many losing tickets, and you know how many losing tickets there are. The greater the number of losing tickets, the better is your justification for believing you will lose. Yet there is no number great enough to transform your fallible opinion into knowledge – after all, you just might win. No justification is good enough – or none short of a watertight deductive argument, and all but the sceptics will agree that this is too much to demand.

Second, because justification is not always necessary. What (non-circular) argument supports our reliance on perception, on memory, and on testimony? And yet we do gain knowledge by these means. And sometimes, far from having supporting arguments, we don’t even know how we know. We once had evidence, drew conclusions, and thereby gained knowledge; now we have forgotten our reasons, yet still we retain our knowledge. Or we know the name that goes with the face, or the sex of the chicken, by relying on subtle visual cues, without knowing what those cues may be.

The link between knowledge and justification must be broken. But if we break that link, then it is not – or not entirely, or not exactly – by raising the standards of justification that epistemology destroys knowledge. I need some different story.

To that end, I propose to take the infallibility of knowledge as my starting point. Must infallibilist epistemology end in scepticism? Not quite. Wait and see. Anyway, here is the definition. Subject S knows proposition P iff P holds in every possibility left uneliminated by S’s evidence; equivalently, iff S’s evidence eliminates every possibility in which not-P.

The definition is short, the commentary upon it is longer. In the first place, there is the proposition, P. What I choose to call ‗propositions‘ are individuated coarsely, by necessary equivalence. For instance, there is only one necessary proposition. It holds in

2 Unless, like some, we simply define ‘justification’ as ‘whatever it takes to turn true opinion into knowledge’ regardless of whether what it takes turns out to involve argument from supporting reasons.

3 The problem of the lottery was introduced in Henry Kyburg, Probability and the Logic of Rational Belief (Middletown, CT: Wesleyan University Press, 1961), and in Carl Hempel, ‘Deductive-Nomological vs. Statistical Explanation’ in Herbert Feigl and Grover Maxwell (eds.), Minnesota Studies in the Philosophy of Science, Vol. II (Minneapolis: University of Minnesota Press, 1962). It has been much discussed since, as a problem both about knowledge and about our everyday, non-quantitative concept of belief.

4 The case of testimony is less discussed than the others; but see C.A.J. Coady, Testimony: A Philosophical Study (Oxford: Clarendon Press, 1992) pp. 79-129.

every possibility; hence in every possibility left uneliminated by S's evidence, no matter who S may be and no matter what his evidence may be. So the necessary proposition is known always and everywhere. Yet this known proposition may go unrecognised when presented in impenetrable linguistic disguise, say as the proposition that every even number is the sum of two primes. Likewise, the known proposition that I have two hands may go unrecognised when presented as the proposition that the number of my hands is the least number n such that every even number is the sum of n primes. (Or if you doubt the necessary existence of numbers, switch to an example involving equivalence by logic alone.) These problems of disguise shall not concern us here. Our topic is modal, not hyperintensional, epistemology.

Next, there are the possibilities. We needn't enter here into the question whether these are concreta, abstract constructions, or abstract simples. Further, we needn't decide whether they must always be maximally specific possibilities, or whether they need only be specific enough for the purpose at hand. A possibility will be specific enough if it cannot be split into subcases in such a way that anything we have said about possibilities, or anything we are going to say before we are done, applies to some subcases and not to others. For instance, it should never happen that proposition P holds in some but not all sub-cases; or that some but not all sub-cases are eliminated by S's evidence.

But we do need to stipulate that they are not just possibilities as to how the whole world is; they also include possibilities as to which part of the world is oneself, and as to when it now is. We need these possibilities de se et nunc because the propositions that may be known include propositions de se et nunc. Not only do I know that there are hands in this world somewhere and somewhen. I know that I have hands, or anyway I have them now. Such propositions aren't just made true or made false by the whole world once and for all. They are true for some of us and not for others, or true at some times and not others, or both.

Further, we cannot limit ourselves to 'real' possibilities that conform to the actual laws of nature, and maybe also to actual past history. For propositions about laws and history are contingent, and may or may not be known.

Neither can we limit ourselves to 'epistemic' possibilities for S - possibilities that S does not know not to obtain. That would drain our definition of content. Assume only that knowledge is closed under strict implication. (We shall consider the merits of this assumption later.) Remember that we are not distinguishing between equivalent propositions. Then knowledge of a conjunction is equivalent to knowledge of every conjunct. P is the conjunction of all propositions not-W, where W is a possibility in which not-P. That suffices to yield an equivalence: S knows that P iff, for every possibility W in which not-P, S knows that not-W. Contraposing and cancelling a double negation: iff every possibility which S does not know not to obtain is one in which P. For short: iff P holds throughout S's epistemic possibilities. Yet to get this far, we need no substantive definition of knowledge at all! To turn this into a substantive definition, in fact the very definition we gave before, we need to say one more thing: S's epistemic possibilities are

---


just those possibilities that are uneliminated by $S$'s evidence.

So, next, we need to say what it means for a possibility to be eliminated or not. Here I say that the uneliminated possibilities are those in which the subject's entire perceptual experience and memory are just as they actually are. There is one possibility that actually obtains (for the subject and at the time in question); call it actuality. Then a possibility $W$ is uneliminated iff the subject's perceptual experience and memory in $W$ exactly match his perceptual experience and memory in actuality. (If you want to include other alleged forms of basic evidence, such as the evidence of our extrasensory faculties, or an innate disposition to believe in God, be my guest. If they exist, they should be included. If not, no harm done if we have included them conditionally.)

Note well that we do not need the 'pure sense-datum language' and the 'incorrigible protocol statements' that for so long bedevilled foundationalist epistemology. It matters not at all whether there are words to capture the subject's perceptual and memory evidence, nothing more and nothing less. If there are such words, it matters not at all whether the subject can hit upon them. The given does not consist of basic axioms to serve as premises in subsequent arguments. Rather, it consists of a match between possibilities.

When perceptual experience $E$ (or memory) eliminates a possibility $W$, that is not because the propositional content of the experience conflicts with $W$. (Not even if it is the narrow content.) The propositional content of our experience could, after all, be false. Rather, it is the existence of the experience that conflicts with $W$: $W$ is a possibility in which the subject is not having experience $E$. Else we would need to tell some fishy story of how the experience has some sort of infallible, ineffable, purely phenomenal propositional content . . . Who needs that? Let $E$ have propositional content $P$. Suppose even — something I take to be an open question — that $E$ is, in some sense, fully characterized by $P$. Then I say that $E$ eliminates $W$ iff $W$ is a possibility in which the subject's experience has content different from $P$. I do not say that $E$ eliminates $W$ iff $W$ is a possibility in which $P$ is false.

Maybe not every kind of sense perception yields experience; maybe, for instance, the kinaesthetic sense yields not its own distinctive sort of sense-experience but only spontaneous judgements about the position of one's limbs. If this is true, then the thing to say is that kinaesthetic evidence eliminates all possibilities except those that exactly resemble actuality with respect to the subject's spontaneous kinaesthetic judgements. In saying this, we would treat kinaesthetic evidence more on the model of memory than on the model of more typical senses.

Finally, we must attend to the word 'every'. What does it mean to say that every possibility in which not-$P$ is eliminated? An idiom of quantification, like 'every', is normally restricted to some limited domain. If I say that every glass is empty, so it's time for another round, doubtless I and my audience are ignoring most of all the glasses there are in the whole wide world throughout all of time. They are outside the domain. They are irrelevant to the truth of what was said.

Likewise, if I say that every uneliminated possibility is one in which $P$, or words to that effect, I am doubtless ignoring some of all the uneliminated alternative possibilities that there are. They are outside the domain, they are irrelevant to the truth of what was said.
But, of course, I am not entitled to ignore just any possibility I please. Else true ascriptions of knowledge, whether to myself or to others, would be cheap indeed. I may properly ignore some uneliminated possibilities; I may not properly ignore others. Our definition of knowledge requires a *sotto voce* proviso. *S knows that P iff S's evidence eliminates every possibility in which not-P*—Psst!—except for those possibilities that we are properly ignoring.

Unger suggests an instructive parallel. Just as *P* is known iff there are no uneliminated possibilities of error, so likewise a surface is flat iff there are no bumps on it. We must add the proviso: Psst!—except for those bumps that we are properly ignoring. Else we will conclude, absurdly, that nothing is flat. (Simplify by ignoring departures from flatness that consist of gentle curvature.)

We can restate the definition. Say that we *presuppose* proposition *Q* iff we ignore all possibilities in which not-*Q*. To close the circle: we *ignore* just those possibilities that falsify our presuppositions. *Proper* presupposition corresponds, of course, to proper ignoring. Then *S* knows that *P* iff *S*’s evidence eliminates every possibility in which not-*P*—Psst!—except for those possibilities that conflict with our proper presuppositions.

The rest of (modal) epistemology examines the *sotto voce* proviso. It asks: what may we properly presuppose in our ascriptions of knowledge? Which of all the uneliminated alternative possibilities may not properly be ignored? Which ones are the ‘relevant alternatives’?—relevant, that is, to what the subject does and doesn’t know? In reply, we can list several rules. We begin with three prohibitions: rules to tell us what possibilities we may not properly ignore.

First, there is the *Rule of Actuality*. The possibility that actually obtains is never properly ignored; actuality is always a relevant alternative; nothing false may properly be presupposed. It follows that only what is true is known, wherefore we did not have to include truth in our definition of knowledge. The rule is ‘externalist’–the subject himself may not be able to tell what is properly ignored. In judging which of his ignorings are proper, hence what he knows, we judge his success in knowing—not how well he tried.

When the Rule of Actuality tells us that actuality may never be properly ignored, we can ask: *whose* actuality? Ours, when we ascribe knowledge or ignorance to others? Or the subject’s? In simple cases, the question is silly. (In fact, it sounds like the sort of pernicious nonsense we would expect from someone who mixes up what is true with

---


11 Some of them, but only some, taken from the authors just cited.
what is believed.) There is just one actual world, we the ascribers live in that world, the subject lives there too, so the subject's actuality is the same as ours.

But there are other cases, less simple, in which the question makes perfect sense and needs an answer. Someone may or may not know who he is; someone may or may not know what time it is. Therefore I insisted that the propositions that may be known must include propositions de se et nunc; and likewise that the possibilities that may be eliminated or ignored must include possibilities de se et nunc. Now we have a good sense in which the subject's actuality may be different from ours. I ask today what Fred knew yesterday. In particular, did he then know who he was? Did he know what day it was? Fred's actuality is the possibility de se et nunc of being Fred on September 19th at such-and-such possible world; whereas my actuality is the possibility de se et nunc of being David on September 20th at such-and-such world. So far as the world goes, there is no difference: Fred and I are worldmates, his actual world is the same as mine. But when we build subject and time into the possibilities de se et nunc, then his actuality yesterday does indeed differ from mine today.

What is more, we sometimes have occasion to ascribe knowledge to those who are off at other possible worlds. I didn't read the newspaper yesterday. What would I have known if I had read it? More than I do in fact know. (More and less: I do in fact know that I left the newspaper unread, but if I had read it, I would not have known that I had left it unread.) I-who-did-not-read-the-newspaper am here at this world, ascribing knowledge and ignorance. The subject to whom I am ascribing that knowledge and ignorance, namely I-as-I-would-have-been-if-I-had-read-the-newspaper, is at a different world. The worlds differ in respect at least of a reading of the newspaper. Thus the ascriber's actual world is not the same as the subject's. (I myself think that the ascriber and the subject are two different people: the subject is the ascriber's otherworldly counterpart. But even if you think the subject and the ascriber are the same identical person, you must still grant that this person's actuality qua subject differs from his actuality qua ascriber.)

Or suppose we ask modal questions about the subject: what must he have known, what might he have known? Again we are considering the subject as he is not here, but off at other possible worlds. Likewise if we ask questions about knowledge of knowledge: what does he (or what do we) know that he knows?

So the question 'whose actuality?' is not a silly question after all. And when the question matters, as it does in the cases just considered, the right answer is that it is the subject's actuality, not the ascriber's, that never can be properly ignored.

Next, there is the Rule of Belief. A possibility that the subject believes to obtain is not properly ignored, whether or not he is right to so believe. Neither is one that he ought to believe to obtain – one that evidence and arguments justify him in believing – whether or not he does so believe.

That is rough. Since belief admits of degree, and since some possibilities are more specific than others, we ought to reformulate the rule in terms of degree of belief, compared to a standard set by the unspecificity of the possibility in question. A possibility may not be properly ignored if the subject gives it, or ought to give it, a degree of belief that is sufficiently high, and high not just because the possibility in question is unspecific.
How high is ‘sufficiently high’? That may depend on how much is at stake. When error would be especially disastrous, few possibilities may be properly ignored. Then even a low degree of belief may be ‘sufficiently high’ to bring the Rule of Belief into play. The jurors know that the accused is guilty only if his guilt has been proved beyond reasonable doubt.\(^\text{12}\)

Yet even when the stakes are high, some possibilities still may be properly ignored. Disastrous though it would be to convict an innocent man, still the jurors may properly ignore the possibility that it was the dog, marvellously well-trained, that fired the fatal shot. And, unless they are ignoring other alternatives more relevant than that, they may rightly be said to know that the accused is guilty as charged. Yet if there had been reason to give the dog hypothesis a slightly less negligible degree of belief — if the world’s greatest dog-trainer had been the victim’s mortal enemy — then the alternative would be relevant after all.

This is the only place where belief and justification enter my story. As already noted, I allow justified true belief without knowledge, as in the case of your belief that you will lose the lottery. I allow knowledge without justification, in the cases of face recognition and chicken sexing. I even allow knowledge without belief, as in the case of the timid student who knows the answer but has no confidence that he has it right, and so does not believe what he knows.\(^\text{13}\) Therefore any proposed converse to the Rule of Belief should be rejected. A possibility that the subject does not believe to a sufficient degree, and ought not to believe to a sufficient degree, may nevertheless be a relevant alternative and not properly ignored.

Next, there is the Rule of Resemblance. Suppose one possibility saliently resembles another. Then if one of them may not be properly ignored, neither may the other. (Or rather, we should say that if one of them may not properly be ignored in virtue of rules other than this rule, then neither may the other. Else nothing could be properly ignored; because enough little steps of resemblance can take us from anywhere to anywhere.) Or suppose one possibility saliently resembles two or more others, one in one respect and another in another, and suppose that each of these may not properly be ignored (in virtue of rules other than this rule). Then these resemblances may have an additive effect, doing more together than any one of them would separately.

We must apply the Rule of Resemblance with care. Actuality is a possibility uneliminated by the subject’s evidence. Any other possibility \(W\) that is likewise uneliminated by the subject’s evidence thereby resembles actuality in one salient respect: namely, in respect of the subject’s evidence. That will be so even if \(W\) is in other respects very dissimilar to actuality — even if, for instance, it is a possibility in which the subject is radically deceived by a demon. Plainly, we dare not apply the Rules of Actuality and Resemblance to conclude that any such \(W\) is a relevant alternative — that would be capitulation to skepticism. The Rule of Resemblance was never meant to apply to this resemblance! We seem to have an \textit{ad hoc} exception to the Rule, though one that makes

\(^{12}\) Instead of complicating the Rule of Belief as I have just done, I might equivalently have introduced a separate Rule of High Stakes saying that when error would be especially disastrous, few possibilities are properly ignored.

good sense in view of the function of attributions of knowledge. What would be better, though, would be to find a way to reformulate the Rule so as to get the needed exception without ad hocery. I do not know how to do this.

It is the Rule of Resemblance that explains why you do not know that you will lose the lottery, no matter what the odds are against you and no matter how sure you should therefore be that you will lose. For every ticket, there is the possibility that it will win. These possibilities are saliently similar to one another: so either every one of them may be properly ignored, or else none may. But one of them may not properly be ignored: the one that actually obtains.

The Rule of Resemblance also is the rule that solves the Gettier problems: other cases of justified true belief that are not knowledge.14

(1) I think that Nogot owns a Ford, because I have seen him driving one; but unbeknownst to me, he does not own the Ford he drives, or any other Ford. Unbeknownst to me, Havit does own a Ford, though I have no reason to think so because he never drives it, and in fact I have often seen him taking the tram. My justified true belief is that I looked at it at 4:22 while it was stopped saying 4:39. That possibility was not properly ignored. It resembles actuality perfectly so far as the stopped clock goes.

(2) The stopped clock is right twice a day. It says 4:39, as it has done for weeks. I look at it at 4:39; by luck I pick up a true belief. I have ignored the uneliminated possibility that it was stopped saying 4:39. That possibility was not properly ignored. It resembles actuality perfectly so far as the stopped clock goes.

(3) Unbeknownst to me, I am travelling in the land of the bogus barns; but my eye falls on one of the few real ones. I don’t know that I am seeing a barn, because I may not properly ignore the possibility that I am seeing yet another of the abundant bogus barns. This possibility saliently resembles actuality in respect of the abundance of bogus barns, and the scarcity of real ones, hereabouts.

(4) Donald is in San Francisco, just as I have every reason to think he is. But, bent on


Though the lottery problem is another case of justified true belief without knowledge, it is not normally counted among the Gettier problems. It is interesting to find that it yields to the same remedy.
deception, he is writing me letters and having them posted to me by his accomplice in Italy. If I had seen the phoney letters, with their Italian stamps and postmarks, I would have concluded that Donald was in Italy. Luckily, I have not yet seen any of them. I ignore the uneliminated possibility that Donald has gone to Italy and is sending me letters from there. But this possibility is not properly ignored, because it resembles actuality both with respect to the fact that the letters are coming to me from Italy and with respect to the fact that those letters come, ultimately, from Donald. So I don’t know that Donald is in San Francisco.

Next, there is the Rule of Reliability. This time, we have a presumptive rule about what may be properly ignored; and it is by means of this rule that we capture what is right about causal or reliabilist theories of knowing. Consider processes whereby information is transmitted to us: perception, memory, and testimony. These processes are fairly reliable.15 Within limits, we are entitled to take them for granted. We may properly presuppose that they work without a glitch in the case under consideration. Defeasibly—very defeasibly!—a possibility in which they fail may properly be ignored.

My visual experience, for instance, depends causally on the scene before my eyes, and what I believe about the scene before my eyes depends in turn on my visual experience. Each dependence covers a wide and varied range of alternatives.16 Of course, it is possible to hallucinate—even to hallucinate in such a way that all my perceptual experience and memory would be just as they actually are. That possibility never can be eliminated. But it can be ignored. And if it is properly ignored—as it mostly is—then vision gives me knowledge. Sometimes, though, the possibility of hallucination is not properly ignored; for sometimes we really do hallucinate. The Rule of Reliability may be defeated by the Rule of Actuality. Or it may be defeated by the Rules of Actuality and of Resemblance working together, in a Gettier problem: if I am not hallucinating, but unbecknownst to me I live in a world where people mostly do hallucinate and I myself have only narrowly escaped, then the uneliminated possibility of hallucination is too close to actuality to be properly ignored.

We do not, of course, presuppose that nowhere ever is there a failure of, say, vision. The general presupposition that vision is reliable consists, rather, of a standing disposition to presuppose, concerning whatever particular case may be under consideration, that we have no failure in that case.

In similar fashion, we have two permissive Rules of Method. We are entitled to presuppose—again, very defeasibly—that a sample is representative; and that the best explanation of our evidence is the true explanation. That is, we are entitled properly to ignore possible failures in these two standard methods of non-deductive inference. Again, the general rule consists of a standing disposition to presuppose reliability in whatever particular case may come before us.


Yet another permissive rule is the Rule of Conservatism. Suppose that those around us normally do ignore certain possibilities, and it is common knowledge that they do. (They do, they expect each other to, they expect each other to expect each other to, . . . ) Then – again, very defeasibly! – these generally ignored possibilities may properly be ignored. We are permitted, defeasibly, to adopt the usual and mutually expected presuppositions of those around us.

(It is unclear whether we need all four of these permissive rules. Some might be subsumed under others. Perhaps our habits of treating samples as representative, and of inferring to the best explanation, might count as normally reliable processes of transmission of information. Or perhaps we might subsume the Rule of Reliability under the Rule of Conservatism, on the ground that the reliable processes whereby we gain knowledge are familiar, are generally relied upon, and so are generally presupposed to be normally reliable. Then the only extra work done by the Rule of Reliability would be to cover less familiar – and merely hypothetical? – reliable processes, such as processes that relied on extrasensory faculties. Likewise, mutatis mutandis, we might subsume the Rules of Method under the Rule of Conservatism. Or we might instead think to subsume the Rule of Conservatism under the Rule of Reliability, on the ground that what is generally presupposed tends for the most part to be true, and the reliable processes whereby this is so are covered already by the Rule of Reliability. Better redundancy than incompleteness, though. So, leaving the question of redundancy open, I list all four rules.)

Our final rule is the Rule of Attention. But it is more a triviality than a rule. When we say that a possibility is properly ignored, we mean exactly that; we do not mean that it could have been properly ignored. Accordingly, a possibility not ignored at all is ipso facto not properly ignored. What is and what is not being ignored is a feature of the particular conversational context. No matter how far-fetched a certain possibility may be, no matter how properly we might have ignored it in some other context, if in this context we are not in fact ignoring it but attending to it, then for us now it is a relevant alternative. It is in the contextually determined domain. If it is an uneliminated possibility in which not-P, then it will do as a counter-example to the claim that P holds in every possibility left uneliminated by S’s evidence. That is, it will do as a counter-example to the claim that S knows that P.

Do some epistemology. Let your fantasies rip. Find uneliminated possibilities of error everywhere. Now that you are attending to them, just as I told you to, you are no longer ignoring them, properly or otherwise. So you have landed in a context with an enormously rich domain of potential counter-examples to ascriptions of knowledge. In such an extraordinary context, with such a rich domain, it never can happen (well, hardly ever) that an ascription of knowledge is true. Not an ascription of knowledge to yourself (either to your present self or to your earlier self, untainted by epistemology); and not an ascription of knowledge to others. That is how epistemology destroys knowledge. But it does so only temporarily. The pastime of epistemology does not plunge us forevermore into its special context. We can still do a lot of proper ignoring, a lot of knowing, and a lot of true ascribing of knowledge to ourselves and others, the rest of the time.

What is epistemology all about? The epistemology we’ve just been doing, at any rate, soon became an investigation of the ignoring of possibilities. But to investigate the
ignoring of them was *ipso facto* not to ignore them. Unless this investigation of ours was an altogether atypical sample of epistemology, it will be inevitable that epistemology must destroy knowledge. That is how knowledge is elusive. Examine it, and straightforwardly it vanishes.

Is resistance useless? If you bring some hitherto ignored possibility to our attention, then straightway we are not ignoring it at all, so *a fortiori* we are not properly ignoring it. How can this alteration of our conversational state be undone? If you are persistent, perhaps it cannot be undone – at least not so long as you are around. Even if we go off and play backgammon, and afterward start our conversation afresh, you might turn up and call our attention to it all over again.

But maybe you called attention to the hitherto ignored possibility by mistake. You only suggested that we ought to suspect the butler because you mistakenly thought him to have a criminal record. Now that you know he does not – that was the previous butler – you wish you had not mentioned him at all. You know as well as we do that continued attention to the possibility you brought up impedes our shared conversational purposes. Indeed, it may be common knowledge between you and us that we would all prefer it if this possibility could be dismissed from our attention. In that case we might quickly strike a tacit agreement to speak just as if we were ignoring it; and after just a little of that, doubtless it really would be ignored.

Sometimes our conversational purposes are not altogether shared, and it is a matter of conflict whether attention to some far-fetched possibility would advance them or impede them. What if some far-fetched possibility is called to our attention not by a sceptical philosopher, but by counsel for the defence? We of the jury may wish to ignore it, and wish it had not been mentioned. If we ignored it now, we would bend the rules of cooperative conversation; but we may have good reason to do exactly that. (After all, what matters most to us as jurors is not whether we can truly be said to know; what really matters is what we should believe to what degree, and whether or not we should vote to convict.) We would ignore the far-fetched possibility if we could – but can we? Perhaps at first our attempted ignoring would be make-believe ignoring, or self-deceptive ignoring; later, perhaps, it might ripen into genuine ignoring. But in the meantime, do we know? There may be no definite answer. We are bending the rules, and our practices of context-dependent attributions of knowledge were made for contexts with the rules unbent.

If you are still a contented fallibilist, despite my plea to hear the sceptical argument afresh, you will probably be discontented with the Rule of Attention. You will begrudge the sceptic even his very temporary victory. You will claim the right to resist his argument not only in everyday contexts, but even in those peculiar contexts in which he (or some other epistemologist) busily calls your attention to far-fetched possibilities of error. Further, you will claim the right to resist without having to bend any rules of cooperative conversation. I said that the Rule of Attention was a triviality: that which is not ignored at all is not properly ignored. But the Rule was trivial only because of how I had already chosen to state the *sotto voce* proviso. So you, the contented fallibilist, will think it ought to have been stated differently. Thus, perhaps: ‘Psst! – except for those possibilities we *could* properly have ignored’. And then you will insist that those far-fetched
possibilities of error that we attend to at the behest of the sceptic are nevertheless possibilities we could properly have ignored. You will say that no amount of attention can, by itself, turn them into relevant alternatives.

If you say this, we have reached a standoff. I started with a puzzle: how can it be, when his conclusion is so silly, that the sceptic’s argument is so irresistible? My Rule of Attention, and the version of the proviso that made that Rule trivial, were built to explain how the sceptic manages to sway us – why his argument seems irresistible, however temporarily. If you continue to find it eminently resistible in all contexts, you have no need of any such explanation. We just disagree about the explanandum phenomenon.

I say S knows that P iff P holds in every possibility left uneliminated by S’s evidence – Psst! – except for those possibilities that we are properly ignoring. ‘We’ means: the speaker and hearers of a given context; that is, those of us who are discussing S’s knowledge together. It is our ignorings, not S’s own ignorings, that matter to what we can truly say about S’s knowledge. When we are talking about our own knowledge or ignorance, as epistemologists so often do, this is a distinction without a difference. But what if we are talking about someone else?

Suppose we are detectives; the crucial question for our solution of the crime is whether S already knew, when he bought the gun, that he was vulnerable to blackmail. We conclude that he did. We ignore various far-fetched possibilities, as hard-headed detectives should. But S does not ignore them. S is by profession a sceptical epistemologist. He never ignores much of anything. If it is our own ignorings that matter to the truth of our conclusion, we may well be right that S already knew. But if it is S’s ignorings that matter, then we are wrong, because S never knew much of anything. I say we may well be right; so it is our own ignorings that matter, not S’s.

But suppose instead that we are epistemologists considering what S knows. If we are well-informed about S (or if we are considering a well-enough specified hypothetical case), then if S attends to a certain possibility, we attend to S’s attending to it. But to attend to S’s attending to it is ipso facto to attend to it ourselves. In that case, unlike the case of the detectives, the possibilities we are properly ignoring must be among the possibilities that S himself ignores. We may ignore fewer possibilities than S does, but not more.

Even if S himself is neither sceptical nor an epistemologist, he may yet be clever at thinking up far-fetched possibilities that are uneliminated by his evidence. Then again, we well-informed epistemologists who ask what S knows will have to attend to the possibilities that S thinks up. Even if S’s idle cleverness does not lead S himself to draw sceptical conclusions, it nevertheless limits the knowledge that we can truly ascribe to him when attentive to his state of mind. More simply: his cleverness limits his knowledge. He would have known more, had he been less imaginative.17

Do I claim you can know P just by presupposing it?! Do I claim you can know that a

---

possibility $W$ does not obtain just by ignoring it? Is that not what my analysis implies, provided that the presupposing and the ignoring are proper? Well, yes. And yet I do not claim it. Or rather, I do not claim it for any specified $P$ or $W$. I have to grant, in general, that knowledge just by presupposing and ignoring is knowledge; but it is an especially elusive sort of knowledge, and consequently it is an unclaimable sort of knowledge. You do not even have to practise epistemology to make it vanish. Simply mentioning any particular case of this knowledge, aloud or even in silent thought, is a way to attend to the hitherto ignored possibility, and thereby render it no longer ignored, and thereby create a context in which it is no longer true to ascribe the knowledge in question to yourself or others. So, just as we should think, presuppositions alone are not a basis on which to claim knowledge.

In general, when $S$ knows that $P$ some of the possibilities in which not-$P$ are eliminated by $S$'s evidence and others of them are properly ignored. There are some that can be eliminated, but cannot properly be ignored. For instance, when I look around the study without seeing Possum the cat, I thereby eliminate various possibilities in which Possum is in the study; but had those possibilities not been eliminated, they could not properly have been ignored. And there are other possibilities that never can be eliminated, but can properly be ignored. For instance, the possibility that Possum is on the desk but has been made invisible by a deceiving demon falls normally into this class (though not when I attend to it in the special context of epistemology).

There is a third class: not-$P$ possibilities that might either be eliminated or ignored. Take the far-fetched possibility that Possum has somehow managed to get into a closed drawer of the desk — maybe he jumped in when it was open, then I closed it without noticing him. That possibility could be eliminated by opening the drawer and making a thorough examination. But if uneliminated, it may nevertheless be ignored, and in many contexts that ignoring would be proper. If I look all around the study, but without checking the closed drawers of the desk, I may truly be said to know that Possum is not in the study — or at any rate, there are many contexts in which that may truly be said. But if I did check all the closed drawers, then I would know better that Possum is not in the study. My knowledge would be better in the second case because it would rest more on the elimination of not-$P$ possibilities, less on the ignoring of them.\textsuperscript{18, 19}

Better knowledge is more stable knowledge: it stands more chance of surviving a

---

\textsuperscript{18} Mixed cases are possible: Fred properly ignores the possibility $W_1$ which Ted eliminates; however Ted properly ignores the possibility $W_2$ which Fred eliminates. Ted has looked in all the desk drawers but not the file drawers, whereas Fred has checked the file drawers but not the desk. Fred's knowledge that Possum is not in the study is better in one way, Ted's is better in another.

\textsuperscript{19} To say truly that $X$ is known, I must be properly ignoring any uneliminated possibilities in which not-$X$; whereas to say truly that $Y$ is better known than $X$, I must be attending to some such possibilities. So I cannot say both in a single context. If I say 'X is known, but Y is better known', the context changes in mid-sentence: some previously ignored possibilities must stop being ignored. That can happen easily. Saying it the other way around — 'Y is better known than X, but even X is known' is harder, because we must suddenly start to ignore previously unignored possibilities. That cannot be done, really; but we could bend the rules and make believe we had done it, and no doubt we would be understood well enough. Saying 'X is flat, but Y is flatter' (that is, 'X has no bumps at all, but Y has even fewer or smaller bumps') is a parallel case. And again, 'Y is flatter, but even X is flat' sounds clearly worse — but not altogether hopeless.
shift of attention in which we begin to attend to some of the possibilities formerly ignored. If, in our new shifted context, we ask what knowledge we may truly ascribe to our earlier selves, we may find that only the better knowledge of our earlier selves still deserves the name. And yet, if our former ignorings were proper at the time, even the worse knowledge of our earlier selves could truly have been called knowledge in the former context.

Never – well, hardly ever – does our knowledge rest entirely on elimination and not at all on ignoring. So hardly ever is it quite as good as we might wish. To that extent, the lesson of scepticism is right – and right permanently, not just in the temporary and special context of epistemology.\(^\text{20}\)

What is it all for? Why have a notion of knowledge that works in the way I described? (Not a compulsory question. Enough to observe that we do have it.) But I venture the guess that it is one of the messy short-cuts – like satisficing, like having indeterminate degrees of belief – that we resort to because we are not smart enough to live up to really high, perfectly Bayesian, standards of rationality. You cannot maintain a record of exactly which possibilities you have eliminated so far, much as you might like to. It is easier to keep track of which possibilities you have eliminated if you – Psst! – ignore many of all the possibilities there are. And besides, it is easier to list some of the propositions that are true in all the uneliminated, unignored possibilities than it is to find propositions that are true in all and only the uneliminated, unignored possibilities.

If you doubt that the word ‘know’ bears any real load in science or in metaphysics, I partly agree. The serious business of science has to do not with knowledge per se; but rather, with the elimination of possibilities through the evidence of perception, memory, etc., and with the changes that one’s belief system would (or might or should) undergo under the impact of such eliminations. Ascriptions of knowledge to yourself or others are a very sloppy way of conveying very incomplete information about the elimination of possibilities. It is as if you had said:

The possibilities eliminated, whatever else they may also include, at least include all the not-P possibilities; or anyway, all of those except for some we are presumably prepared to ignore just at the moment.

The only excuse for giving information about what really matters in such a sloppy way is that at least it is easy and quick! But it is easy and quick; whereas giving full and precise information about which possibilities have been eliminated seems to be extremely difficult, as witness the futile search for a ‘pure observation language’. If I am right about how ascriptions of knowledge work, they are a handy but humble approximation. They may yet be indispensable in practice, in the same way that other handy and humble approximations are.

If we analyse knowledge as a modality, as we have done, we cannot escape the conclusion that knowledge is closed under (strict) implication.\(^\text{21}\) Dretske has denied that

---

\(^{20}\) Thanks here to Stephen Hetherington. While his own views about better and worse knowledge are situated within an analysis of knowledge quite unlike mine, they withstand transplantation.

\(^{21}\) A proof-theoretic version of this closure principle is common to all ‘normal’ modal logics: if
knowledge is closed under implication; further, he has diagnosed closure as the fallacy that drives arguments for scepticism. As follows: the proposition that I have hands implies that I am not a handless being, and *a fortiori* that I am not a handless being deceived by a demon into thinking that I have hands. So, by the closure principle, the proposition that I know I have hands implies that I know that I am not handless and deceived. But I don’t know that I am not handless and deceived – for how can I eliminate that possibility? So, by *modus tollens*, I don’t know that I have hands. Dretske’s advice is to resist scepticism by denying closure. He says that although having hands does imply not being handless and deceived, yet knowing that I have hands does not imply knowing that I am not handless and deceived. I do know the former, I do not know the latter.22

What Dretske says is close to right, but not quite. Knowledge is closed under implication. Knowing that I have hands does imply knowing that I am not handless and deceived. Implication preserves truth – that is, it preserves truth in any given, fixed context. But if we switch contexts midway, all bets are off. I say (1) pigs fly; (2) what I just said had fewer than three syllables (true); (3) what I just said had fewer than four syllables (false). So ‘less than three’ does not imply ‘less than four’? No! The context switched midway, the semantic value of the context-dependent phrase ‘what I just said’ switched with it. Likewise in the sceptical argument the context switched midway, and the semantic value of the context-dependent word ‘know’ switched with it. The premise ‘I know that I have hands’ was true in its everyday context, where the possibility of deceiving demons was properly ignored. The mention of that very possibility switched the context midway. The conclusion ‘I know that I am not handless and deceived’ was false in *its* context, because that was a context in which the possibility of deceiving demons was being mentioned, hence was not being ignored, hence was not being properly ignored. Dretske gets the phenomenon right, and I think he gets the diagnosis of scepticism right; it is just that he misclassifies what he sees. He thinks it is a phenomenon of logic, when really it is a phenomenon of pragmatics. Closure, rightly understood, survives the test. If we evaluate the conclusion for truth not with respect to the context in which it was uttered, but instead with respect to the different context in which the premise was uttered, then truth is preserved. And if, *per impossibile*, the conclusion could have been said in the same unchanged context as the premise, truth would have been preserved.

A problem due to Saul Kripke turns upon the closure of knowledge under implication. *P* implies that any evidence against *P* is misleading. So, by closure, whenever you know that *P*, you know that any evidence against *P* is misleading. And if you know that evidence is misleading, you should pay it no heed. Whenever we know – and we know a lot, remember – we should not heed any evidence tending to suggest that we are wrong. But that is absurd. Shall we dodge the conclusion by denying closure? I think not. Again, I diagnose a change of context. At first, it was stipulated that *S* knew, whence it

21 Continued. . .

the logic validates an inference from zero or more premises to a conclusion, then also it validates the inference obtained by prefixing the necessity operator to each premise and to the conclusion. Further, this rule is all we need to take us from classical sentential logic to the least normal modal logic. See Brian Chellas, *Modal Logic: An Introduction* (Cambridge: Cambridge University Press, 1980) p. 114.
followed that $S$ was properly ignoring all possibilities of error. But as the story continues, it turns out that there is evidence on offer that points to some particular possibility of error. Then, by the Rule of Attention, that possibility is no longer properly ignored, either by $S$ himself or by we who are telling the story of $S$. The advent of that evidence destroys $S$’s knowledge, and thereby destroys $S$’s licence to ignore the evidence lest he be misled.

There is another reason, different from Dretske’s, why we might doubt closure. Suppose two or more premises jointly imply a conclusion. Might not someone who is compartmentalized in his thinking – as we all are? – know each of the premises but fail to bring them together in a single compartment? Then might he not fail to know the conclusion? Yes; and I would not like to plead idealization-of-rationality as an excuse for ignoring such cases. But I suggest that we might take not the whole compartmentalized thinker, but rather each of his several overlapping compartments, as our ‘subjects’. That would be the obvious remedy if his compartmentalization amounted to a case of multiple personality disorder; but maybe it is right for milder cases as well.

A compartmentalized thinker who indulges in epistemology can destroy his knowledge, yet retain it as well. Imagine two epistemologists on a bushwalk. As they walk, they talk. They mention all manner of far-fetched possibilities of error. By attending to these normally ignored possibilities they destroy the knowledge they normally possess. Yet all the while they know where they are and where they are going! How so? The compartment in charge of philosophical talk attends to far-fetched possibilities of error. The compartment in charge of navigation does not. One compartment loses its knowledge, the other retains its knowledge. And what does the entire compartmentalized thinker know? Not an altogether felicitous question. But if we need an answer, I suppose the best thing to say is that $S$ knows that $P$ iff any one of $S$’s compartments knows that $P$. Then we can say what we would offhand want to say: yes, our philosophical bushwalkers still know their whereabouts.

Context-dependence is not limited to the ignoring and non-ignoring of far-fetched possibilities. Here is another case. Pity poor Bill! He squanders all his spare cash on the pokies, the races, and the lottery. He will be a wage slave all his days. We know he will never be rich. But if he wins the lottery (if he wins big), then he will be rich. Contrapositively: his never being rich, plus other things we know, imply that he will lose. So, by closure, if we know that he will never be rich, we know that he will lose. But when we discussed the case before, we concluded that we cannot know that he will lose. All the possibilities in which Bill loses and someone else wins saliently resemble the possibility in which Bill wins and the others lose; one of those possibilities is actual; so by the Rules of Actuality and of Resemblance, we may not properly ignore the possibility that Bill wins. But there is a loophole: the resemblance was required to be salient. Salience, as well as ignoring, may vary between contexts. Before, when I was explaining how the Rule of Resemblance applied to lotteries, I saw to it that the resemblance between the many possibilities associated with the many tickets was sufficiently salient. But this time, when we were busy pitying poor Bill for his habits and not for his luck, the resemblance of the many possibilities was not so salient. At that point, the possibility of

Bill's winning was properly ignored; so then it was true to say that we knew he would never be rich. Afterward I switched the context. I mentioned the possibility that Bill might win, wherefore that possibility was no longer properly ignored. (Maybe there were two separate reasons why it was no longer properly ignored, because maybe I also made the resemblance between the many possibilities more salient.) It was true at first that we knew that Bill would never be rich. And at that point it was also true that we knew he would lose – but that was only true so long as it remained unsaid! (And maybe unthought as well.) Later, after the change in context, it was no longer true that we knew he would lose. At that point, it was also no longer true that we knew he would never be rich.

But wait. Don’t you smell a rat? Haven’t I, by my own lights, been saying what cannot be said? (Or whistled either.) If the story I told was true, how have I managed to tell it? In trendyspeak, is there not a problem of reflexivity? Does not my story deconstruct itself?

I said: $S$ knows that $P$ iff $S$’s evidence eliminates every possibility in which not-$P$ – Psst! – except for those possibilities that we are properly ignoring. That ‘psst’ marks an attempt to do the impossible – to mention that which remains unmentioned. I am sure you managed to make believe that I had succeeded. But I could not have done.

And I said that when we do epistemology, and we attend to the proper ignoring of possibilities, we make knowledge vanish. First we do know, then we do not. But I had been doing epistemology when I said that. The uneliminated possibilities were not being ignored – not just then. So by what right did I say even that we used to know?24

In trying to thread a course between the rock of fallibilism and the whirlpool of skepticism, it may well seem as if I have fallen victim to both at once. For do I not say that there are all those uneliminated possibilities of error? Yet do I not claim that we know a lot? Yet do I not claim that knowledge is, by definition, infallible knowledge?

I did claim all three things. But not all at once! Or if I did claim them all at once, that was an expository shortcut, to be taken with a pinch of salt. To get my message across, I bent the rules. If I tried to whistle what cannot be said, what of it? I relied on the cardinal principle of pragmatics, which overrides every one of the rules I mentioned: interpret the message to make it make sense – to make it consistent, and sensible to say.

When you have context-dependence, ineffability can be trite and unmysterious. Hush! [moment of silence] I might have liked to say, just then, ‘All of us are silent’. It was true. But I could not have said it truly, or whistled it either. For by saying it aloud, or by whistling, I would have rendered it false.

I could have said my say fair and square, bending no rules. It would have been tiresome, but it could have been done. The secret would have been to resort to ‘semantic ascent’. I could have taken great care to distinguish between (1) the language I use when I talk about knowledge, or whatever, and (2) the second language that I use to talk about

24 Worse still: by what right can I even say that we used to be in a position to say truly that we knew? Then, we were in a context where we properly ignored certain uneliminated possibilities of error. Now, we are in a context where we no longer ignore them. If now I comment retrospectively upon the truth of what was said then, which context governs: the context now or the context then? I doubt there is any general answer, apart from the usual principle that we should interpret what is said so as to make the message make sense.
the semantic and pragmatic workings of the first language. If you want to hear my story
told that way, you probably know enough to do the job for yourself. If you can, then my
informal presentation has been good enough.

Princeton University

Received October 1995
Revised August 1996
I. MANIFESTO

An adequate theory of colour must be both materialistic and commensensical. The former demand is non-negotiable. The latter can be compromised to some degree. We need not be 'ever so inclusive' in advancing all our offhand folk-theoretical opinions as conditions of adequacy on a theory. Imperfect occupants of the folk-theoretical role of colour will be imperfect deservers of the name, but may nevertheless deserve it quite well enough.

But compromise has its limits. It won't do to say that colours do not exist; or that we are unable to detect them; or that they never are properties of material things; or that they go away when things are...
unilluminated or unobserved; or that they change with every change in the illumination, or with every change in an observer’s visual capacities; or that the same surface of the same thing has different colours for different observers. Compromise on these points, and it becomes doubtful whether the so-called ‘colours’ posited in your theory are rightly so-called. Yet it is a Moorean fact that there are colours rightly so-called. Deny it, and the most credible explanation of your denial is that you are in the grip of some philosophical (or scientific) error.

II. FOLK PSYCHOPHYSICS

In other words, it is a Moorean fact that the folk psychophysics of colour is close to true.

Like other folk theories, the folk psychophysics of colour is a generally shared body of tacit belief. It concerns not only colours themselves, but also the inner states of colour experience that the colours tend to cause in us, and further inner states and behaviour that these colour experiences cause in turn. It is not to be supposed that we go around with careful formulations of folk psychophysics in mind, though perhaps with enough patience it might be possible to elicit such formulations Meno-fashion.

Folk psychophysics should, for instance, afford an explanation of why we can sort dyed bits of wool, and then shuffle them together, and then sort them again into just the same heaps as before. This folk-psychophysical explanation will involve many causal chains, each running from the colour of a bit of wool to a colour experience in the sorter, and thence to a desire to put that bit of wool on a certain heap, and thence to the desired behaviour.

The folk psychophysics of colour is doubtless a fragment of some larger folk theory. Elsewhere in that larger theory we might find the explanation of why our wool-sorter wanted to succeed at his task. We need not worry about demarcation, so long as we agree to include all that will be needed for what follows.

The folk psychophysics of colour is common knowledge among us. In the same tacit way in which we believe the theory itself, we

likewise believe that others around us all believe it too; and that they
in turn ascribe belief in it to those around them in the same way we
do; and so on ad infinitum, or at least as far as is humanly possible.\(^5\)

A theory implicitly defines its theoretical terms. If, without benefit
of any prior definition of ‘entropy’, thermodynamics says that entropy
does this, that, and the other, we may factor that into two parts. There
is an existential claim – a ‘Ramsey sentence’ – to the effect that there
exists some quantity which does this, that, and the other (or near
enough). And there is a semantic stipulation: let that which does this,
that, and the other (or near enough), if such there be, bear the name
‘entropy’.\(^6\) Here is another way to say it: the theory associates with
the term ‘entropy’ a certain theoretical role. It claims that this role is
occupied. And it implicitly defines ‘entropy’ as a name for the occu-
pant of the role.

What thermodynamics can do, the folk psychophysics of colour can
likewise do. The situation is messier, of course, because the theory is
tacit. Lacking an authoritative codification, we have no sharp way to
tell what is non-negotiably part of the theory, what is provisionally part
of it, and what is no part of it at all. Nor are these matters of secret fact.
When folk psychophysics defines its theoretical terms, we run a risk of
semantic indecision. Yet within limits, that indecision is harmless. We
can hope – indeed we can be confident, as a matter of Moorean fact! –
that the safe limits are not exceeded.

Central to the folk psychophysics of colour are a multitude of prin-
ciples connecting colours and corresponding colour experiences. They
go roughly as follows.

When a red thing is before someone’s eyes, it typically causes in
him an experience of redness

or better

5 See my Convention: A Philosophical Study (Cambridge, MA: Harvard University Press,
197 and elsewhere.
pp. 427–446.
When a red thing illuminated by normal light is a suitable distance before the eyes of someone with normal visual capacities in normal surroundings, it typically causes in him an experience of redness.

'Normal light' can be explained in terms of the range of illumination that most people — actually, nowadays, and hereabouts — mostly encounter. Likewise, mutatis mutandis, for normal capacities and normal surroundings. Some examples of the abnormal are: sodium vapour light; eyes adapted to sodium vapour light; a room with purple walls, floor, and ceiling, filled with purple furniture, with purple curtains over the windows. I do not suppose it is incumbent on folk psychophysics to do much by way of listing abnormal cases — if it tried, it would soon outrun common knowledge. A statistical conception, rigidified to actuality and nowadays and hereabouts, should suffice. It is otherwise with distance: I dare say that folk psychophysics has a lot to tell us about which distances are 'suitable', but for me that part of the theory remains tacit.

Further, it should really have been

When a red thing is at such-and-such direction before someone's eyes, it typically causes in him an experience of redness-in-such-and-such-direction

or

When a red thing fills such-and-such solid angle . . .

None of these refinements will prove relevant to the principal topic of this paper, so for the most part I shall ignore them.

Drawing on these connecting principles, and also on the part of folk psychophysics that classifies colours as properties (for the most part) of the surfaces of opaque things, and colour experiences as inner states of people (and perhaps other animals), we have folk-psychophysical roles for the colours and for the colour experiences. When we take the theoretical terms to name the occupants of the theoretical roles, we arrive at 'definitions' such as these.
D1 Red is the surface property of things which typically causes experience of red in people who have such things before their eyes.

D2 Experience of red is the inner state of people which is the typical effect of having red things before the eyes.

If, as a matter of contingent fact, the surface property that causes experience of red is a certain reflectance property — that is, a property that supervenes on the thing’s reflectance spectrum — then D1 may serve as a premise for a chromophysical identification: red is, as a matter of contingent fact, that reflectance property. And if, as a matter of contingent fact, the inner state which is the typical effect of red things before the eyes is a certain pattern of neuron firings in the visual cortex, then D2 may serve as a premise for a psychophysical identification: experience of red is, as a matter of contingent fact, that pattern of neural firings.

D1 and D2 are all very well as truths; as widely shared tacit beliefs, and even as items of our common knowledge; and as premises for chromophysical and psychophysical identifications. But as a pair of definitions they are almost totally useless, by reason of circularity. Anyone who needed definitions both of ‘red’ and of ‘experience of red’ would be little wiser if we gave him both D1 and D2. That circularity, its remedy, and the new problem that arises out of that remedy, shall be our principal business in the rest of this paper.

III. OBJECTIONS AND REPLIES

But first I stop to address some other likely worries, apart from circularity, about what I have already said.

Objection. Our colour experiences do not depend solely on the colours of things before the eyes. They depend also on the illumination, the visual capacities of the observer, and the surroundings. And it is not good enough just to make an exception for ‘abnormal’ cases, understood simply as infrequent cases (infrequent actually, nowadays, and hereabouts). There is no range of ‘normal’ cases such that we only
seldom find ourselves outside that range, and also such that within that range the effects of differing illumination, etc., are entirely absent.7

Reply. No; but there is a range such that we seldom find ourselves outside it, and such that within it the effects of differing illumination, etc., are so small that they are hard to notice. Go from noonday sun to a cloudy afternoon to sunset; go from any of these to incandescent lighting to fluorescent lighting (but not to sodium or mercury vapour lighting); go from the room painted yellow to the room painted purple (but not to a room where almost everything in sight is purple); go from one to another so suddenly that your eyes are maladapted at first. Take with you a book with a green cover. It will continue throughout to cause experience of green, and will not start to cause experience of blue instead. Indeed, if it is the right book, it will continue throughout to cause experience of some fairly specific shade of green: Brunswick green, not malachite green or apple green.

(To be sure, it does not continue throughout to cause precisely the same sort of colour experience. And if our topic were not the naming of the colours that we name in daily life, but instead the naming – or rather, numbering – of the myriad precise shades that must be distinguished by matchers of paint, we would have to be much more fussy.)

In short, the barely noticeable differences in colour experience due to frequently encountered differences in illumination are not a problem. The much more striking differences in experience that are caused by much less frequently encountered differences in illumination are not a problem either: we need only say, for instance, that sodium vapour light causes colour illusions. It is only if we conflate these two non-problems, and we try to discuss the entire range of dependence of experience on illumination in a uniform way, that we get a problem. It would be daft to deny that sodium vapour light causes illusions. It would be still more daft to say that all light causes colour illusions.

with the sole exception of cloudy Scottish daylight (or some other precisely and arbitrarily chosen standard).

Another way to create an illusory problem is to ignore our daily experience, follow where theory leads, and then back an oversimple theory. Imagine that the eye is simply a spectrometer that measures the intensities at various wavelengths of the light coming from the thing seen, and produces colour experiences that depend just on those intensities. Then the experience produced by something with an unchanging reflectance spectrum would depend a lot on illumination - even for illumination within the frequently encountered range. The colours of things could not seem as constant as they manifestly do.

So should we assume severe and ubiquitous inconstancy when we philosophize about colour? Of course not! Rather, we should stop assuming that the eye is simply a spectrometer. We should think of it, perhaps, as a calibrated spectrometer. If the spectrometer's measurements are corrected to compensate for differences in illumination, they will indeed measure reflectance properties, and experience that depends on the corrected measurements will indeed exhibit constancy of colour. And if the spectrometer shares information with similar spectrometers aimed in different directions (or if it spends some of its own time scanning around the scene), and if it has definite preconceptions about the sort of distribution of reflectance values to be expected, then it will be in a position to calculate the proper correction factors. Of course the method has its limitations. No amount of calibration will permit measurement of reflectance values in a band of wavelengths that is missing from the illumination; when the preconceptions used to calculate correction factors are mistaken, the reflectance measurements will likewise be mistaken; and recalibration after a change of illumination will not be instantaneous. Nevertheless, this fallible method may be good enough to yield approximate constancy throughout a wide range of frequently encountered cases. And that suffices.

8 Here I follow the lead of David Hilbert, Color and Color Perception (Chicago: Center for the Study of Language and Information, 1987). One theory that treats the eye as a calibrated spectrometer capable of measuring reflectance properties is Land's well-known 'retinex' theory. But it is by no means the only such theory on offer.
Objection. A brown expanse on the television is composed of diversely coloured pixels – none of them brown. The brown expanse and a certain blue pixel are before the eye in the same direction, at the same distance, at the same time. Should they not cause a colour experience of blue and brown in the very same direction? Yet they do not.

Reply. The pixel and the expanse are not both at a ‘suitable distance’, even though they are at the same distance, because the suitability of a viewing distance depends on the size of the thing viewed.

Follow-up objection. Then consider a pixel-sized part of the brown expanse. Now the brown thing and the blue thing are the same in size as well as distance, so if one is at a suitable viewing distance then so is the other.

Reply. A pixel-sized part of a brown expanse is not necessarily brown. In this case it is blue. In fact, it is the blue pixel already considered. If the pixel is at a suitable distance, no brown part of the brown expanse is at a suitable distance and also in the same direction as the pixel. Colours are imperfectly ‘dissective’ properties: they are not always shared between a thing and its proper parts.9

Objection. Some ostensible facts about the colours – for instance, that there cannot be a reddish green, or that there cannot be a shade of yellow that is closer to various shades of blue than it is to any other shade of yellow – are best explained in terms of the way our colour vision works, rather than in terms of relations between physical properties of surfaces. Then if colours are physical properties of surfaces, how can these facts of exclusion and proximity be facts about colours?

Reply. Our account provides a correspondence between colours and colour experiences. (Various rival accounts also provide such correspondences.) For now, that correspondence takes the form of a set of definitional circles, but soon we shall give it a more satisfactory form. Whatever form it takes, the correspondence yields relations among colours in the image of relations among colour experiences (or vice versa). So no matter where the relations of exclusion and

9 See Hilbert, Color and Color Perception, pp. 29–42.
proximity may originate, in the end we have them twice over: as relations among colour experiences and also as relations among the corresponding colours.

(We might have had an offhand opinion that these relations originated as relations among surface properties. If so, we were wrong. But I am not sure we had any such opinion at all; and if we did, we have no business elevating it into a Moorean fact of folk psychophysics.)

If it is absolutely impossible to have an experience of reddish green (or if it is nomically impossible, or if it is unimaginable, or if it is just very difficult to arrange); and if reddish green is by definition the surface property that typically causes experience of reddish green; then the desired exclusion among colours follows. Then it is absolutely impossible (or nomically impossible, or unimaginable, or difficult) for there to be a reddish green surface.\textsuperscript{10} \textit{Quod erat demonstrandum.} Likewise for the yellow in the midst of blue.

\textit{Objection.} Not only surfaces of opaque things are coloured, but also transparent volumes, light sources, the sky, . . . Colours are not always surface properties. \textit{A fortiori} they are not always reflectance properties.

\textit{Reply.} I have no opinion about whether the blue of the sky and the blue of an opaque picture of the sky are one shared non-disjunctive property, or whether they are two different, but saliently related, properties.

But even in the first case there is a colour property – \textit{surface blue}, or rather a certain shade thereof – that the picture has and that the sky itself lacks. (It is the conjunction of the shared property with the property of being a surface of an opaque thing.) And in the second case also, we have a surface blue that the picture has and the sky lacks. It is these surface colours that are my present concern, and that well might be identical to reflectance properties. My restriction of topic

\textsuperscript{10} It is not absolutely or nomically impossible to arrange for an experience of reddish green. But it is difficult, at least by the method presently known: that of H. Crane and T. P. Plantida, 'On Seeing Reddish Green and Yellowish Blue', \textit{Science} 221 (1983) pp. 1078–1080. That method does not involve any special surface property before the eyes. I do not know whether those who have the experience acquire an ability to imagine it afterward when they are no longer having it.
leaves unfinished business. But I doubt that it leads to error within my restricted topic.

*Objection.* If colours are reflectance properties, they supervene on reflectance spectra. A reflectance spectrum is a disposition, or a bundle of dispositions, to reflect various proportions of light at various wavelengths. A disposition requires a ‘causal basis’. For instance, something is disposed to break when struck if it has some property – some intrinsic property, at least in the clearest cases – which, together with striking, would cause breaking. In the case of reflectance properties, the causal basis is a matter of surface microstructure. It turns out that a given reflectance property can have any of many microstructural bases, different ones in different actual cases. Now we face a nasty four-way choice.\(^{11}\)

Shall we say (1) that the colour is not the reflectance property after all, but rather it is the microstructural causal basis thereof? Or shall we say (2) that the colour is the reflectance property, but that a reflectance property – and any other disposition likewise – is identical to its causal basis? Either way, we end up concluding that what we thought was one single colour, indeed one single shade of one colour, is really many different microstructural properties, different ones in different cases. That is seriously uncommonsensical. Or shall we rather say that a colour is a reflectance property, a reflectance property is a disposition, a disposition is not identical to any of its causal bases, but rather a disposition is the property of having some suitable causal basis or other? Then what shall we say about how the selective reflection of light – or more generally, the manifestation of any disposition – is caused? We cannot deny that the causal basis does the causing. That is how we defined ‘causal basis’, after all. Shall we say (3) that only the causal basis does the causing, and the disposition itself is impotent? In the present case, that would mean denying that surface colours cause selective reflection of light; and denying also that they cause more distant effects, such as colour experiences or colour-discriminating behaviour, via causal chains that begin with selective reflection of light. And it is

safe to say that colours do not cause colour experiences some other way! But to deny that colours cause colour experiences is seriously uncommonsensical.

Or shall we say that a disposition and its basis, for instance a colour and a microstructure, both cause the manifestation? What kind of multiple causation is this? It seems wrong to say that the basis causes the manifestation by first causing the disposition, or vice versa, or that the basis and the disposition jointly cause the manifestation. The remaining alternative is to say (4) that we must have a case of over-determination: the disposition and its basis redundantly cause the manifestation. For instance, the colour and the microstructure redundantly cause the selective reflection of light, and thereby they redundantly cause the colour experience. This alternative, too, is seriously uncommonsensical.

Reply. Alternatives (1), (2), and (3) should be discarded. Alternative (4) is close to right, but needs correction. The key to sorting out this mess is that we should not have been talking about properties as causes in the first place. That is loose talk. (Most of the time it is harmlessly loose talk, and I do not propose to give it up just because it got us into trouble this once.) Really, the causing is done by particular havings of properties — particular ‘events’, as we call them. (But that name is not meant to imply that all events are short-lived or involve change. If something in a frozen world is changeless throughout some short or long interval, or indeed throughout all time, its changelessness during the interval is still called an ‘event’.) The very same event that is essentially a having of some causal basis of a certain disposition is also accidentally a having of the disposition itself.\(^\text{12}\) So an effect of this event is caused by a having of the basis, and caused also by a having of the disposition. But since these havings are one and the same event, there is no redundant causation. So a colour experience may be caused by a colour, and also by the microstructural causal basis of the reflectance values that comprise that colour, without having a case of causal over-determination.

Objection. If we do not want to say that what we thought was one single colour is really many different properties, different ones in different cases, then we also have a problem about metamers: different reflectance spectra that cause the same colour experience.¹³ What to say about them?

Reply. At worst, we have a problem about some metamers. I note, first, that even if we think that the colours will turn out to be reflectance properties, we have no reason to think that all reflectance properties—all properties that supervene on reflectance spectra—will turn out to be colours. Metamers that differ in their reflectance spectra, but do not differ in those reflectance properties that participate in causing colour experience, are no problem. We have the same colour, exactly as in the case where there are different microstructural causal bases of the same reflectance spectrum.

Suppose, for example, that there is a certain division of the visible spectrum into long-wave, middle-wave, and short-wave intervals such that, if we take the triple of integrals of reflectance over the three intervals, the colours will turn out to supervene not only on reflectance spectra but on the triples of integrals.¹⁴ When the spectra differ but the triples do not differ, we have the same colour. Only if different triples caused the same colour experience would we have a problem; only then would we face a choice between denying the identity of colours with triples of reflectance integrals and admitting, uncommonsensically, that some colours cannot be distinguished by sight. (Not without varying the illumination, anyway. Metamers under one illumination may look different under another, especially if we resort to illuminations not frequently encountered.)

I note, second, that it would not be so very bad to acknowledge

---


¹⁴ Perhaps the three intervals should be ‘fuzzy’ intervals. These are taken as functions—bell curves, or similar—measuring the degree to which a given point is deemed to fall within the interval. To integrate a quantity over a fuzzy interval, we multiply the quantity by the function just mentioned, and we integrate that product *simpliciter*. 
some cases of indistinguishable colours, so long as such cases were uncommon. David Hilbert reports an encouraging consequence of one well-developed model of colour constancy (that is, of how the spectrometer gets calibrated in accordance with preconceptions):

Although there are possible differences of colour that are undetectable in normal circumstances, the actual occurrence of such differences appears to be relatively rare.\[15\]

\textit{Objection.} When the purple people eater comes out of hiding to stalk its prey, it uses its powers of telehypnosis to disrupt the colour vision of victim and bystanders. In this way it manages to be mistaken for a benign brown beast. The deception always works. The purple people eater, though it is purple from tail to teeth, never causes experience of purple. Is this a counter-example?

\textit{Reply.} No. The question what the eater itself typically causes is the wrong question. We were supposed instead to ask what the surface properties of the eater typically cause. Does it have a surface property such that this property – or rather, particular havings thereof – will typically cause experience of purple? It does. The same property that occurs on the people eater occurs, or at least could occur, elsewhere: before the eyes of a normal perceiver with undisrupted colour vision. In that case it typically causes, or would cause, experience of purple. Accordingly, that property is the colour purple, and the purple people eater is rightly so called.

\textit{Objection.} The yellow killer is a more problematic predator. It too disrupts the colour vision of anyone who sets eyes on it; and it disrupts all other brain processes as well, thereby causing instant death. But in this case, it is the killer’s colour that does the damage. Its special shade of yellow, ‘killer yellow’, is fatal regardless of what the coloured thing may be. This colour does not typically cause colour experience. It never does, and never could so long as we retain our vulnerability to it.\[16\]

\[15\] Hilbert, Color and Color Perception, pp. 130–131.
\[16\] The example of killer yellow is due to Saul Kripke, in lectures many years ago. I
Reply. The case is consistent if, but only if, we can subsume killer yellow as a special case under some broader property that does typically cause experience of yellow; and that is not an artificial gerrymander, but rather is unified by appropriate resemblances between its various subcases. If so, we can repeat the strategy that worked in the case of the purple people eater: we distinguish a broader property which deserves the colour name from a narrower property which, if considered just on its own, does not.

To put the point another way: killer yellow deserves to be so called, if indeed it does, in virtue of its resemblance to other shades of yellow that do cause experience of yellow. For instance, it might resemble other shades of yellow in respect of the values of its triple of long-wave, middle-wave, and short-wave integrals of reflectance, if those turn out to be the reflectance properties on which colour experiences normally depend.

In that case, the relation that unites the killer property with non-lethal shades of yellow, and thereby justifies us in classifying the killer property itself as a shade of yellow, is colour resemblance in a doubly derivative sense. First, we have resemblance among colour experiences. That resemblance does not apply directly, since there are no experiences of killer yellow. Second, we have the image among colours of this first relation among experiences, under the causal correspondence between surface properties and the experiences they typically cause: the derivative relation that holds between colours just when the colour experiences they typically cause stand in the first relation. This second resemblance also does not apply directly, since killer yellow does not cause any colour experience at all. Finally, it may be that this second relation typically correlates with a third relation: resemblance in respect of the triple of reflectance integrals. If that turns out to be so (and whether it is so is an empirical question, even supposing that the colours do turn out to be triples of reflectance integrals) then in a still more derivative sense we may regard this third relation also as a relation of colour resemblance. So, finally, we have a relation that can unite killer yellow with other yellows.

I am obliged to note that what I say here may not correspond to the whole of what Kripke said in those lectures.
IV. FROM CIRCULARITY TO MULTIPLICITY

I return at last to our principal difficulty: the definitional circle between the name of a colour and the name of the corresponding colour experience. Red is the surface property apt for causing experience of red, which is the inner state that red things before the eyes are apt for causing. Likewise for green, for magenta, . . .

A manoeuvre due mainly to Carnap often works to cure such circularities, but in the present case it proves disappointing.17 Here is how it works, when it works. Suppose thermodynamics has two theoretical terms, 'entropy' and 'temperature'; and without benefit of prior definition of either, it says that entropy does this, that, and the other; temperature does so and so; temperature and entropy are related thus. Then any ordered pair of $X$ and $Y$ is called a realization of thermodynamics iff $X$ does this, that, and the other; $Y$ does so and so; $X$ and $Y$ are related thus. (Equivalently: $X$ and $Y$ make thermodynamics true, if taken as referents of the respective terms 'entropy' and 'temperature'.) Now we define our terms without circularity: entropy is the first component of the unique realization of thermodynamics, if such there be; temperature is the second component of the unique realization.

What to do if there is no realization, not even a near-enough imperfect realization? Here different versions of the method disagree. Carnap's version falls silent about what referents, if any, the theoretical terms have; my version and Bedard's version say that the terms fail to refer. What to do if there are two (or more) different realizations, say the pair of $X_1$ and $Y_1$ and also the pair of $X_2$ and $Y_2$? Again there is disagreement. Carnap's version says that either the terms refer respectively to $X_1$ and $Y_1$ or else they refer respectively to $X_2$ and $Y_2$, but falls silent about which. My version says again that the terms fail to refer. Bedard's version says that the terms become indeterminate in reference, so that there is no fact of the matter about whether they refer respectively to $X_1$ and $Y_1$ or instead to $X_2$ and $Y_2$; and as usual

we resort to van Fraassen's method of supervaluations and say that what is true (or false) on both resolutions of the indeterminacy alike is true (or false) simpliciter.\textsuperscript{18} Nowadays I am inclined to split the difference between my version and Bedard's: when the two realizations are sufficiently different we get reference failure; when they are sufficiently alike we get indeterminacy of reference; in between it is indeterminate which one we get.

Now suppose our term-introducing theory is a fragment of the folk psychophysics of colour. It has just one pair of a colour name and a colour-experience name: 'red' and 'experience of red'. It says nothing to distinguish this pair of a colour and a corresponding colour experience from all other corresponding pairs. So we have a very severe problem of multiple realization: every corresponding pair of a colour and a colour experience is a realization of our theory. But it will not do to fall silent about the reference of 'red' and 'experience of red'; or to say that these terms fail to refer; or to say that these terms are radically indeterminate in reference. Our method for curing circularities, in all its versions, has been defeated by the multiplicity of realizations.

But at least we are now in a position to do some useful redefining. A corresponding pair is any realization of the theory just considered. A colour is any first component of a corresponding pair. A colour experience is any second component.

If we take a bigger fragment of folk psychophysics that mentions \( n \) different colours and their corresponding experiences, we make matters worse. Now we have as many realizations as there are ordered \( n \)-tuples of corresponding pairs!

Instead of our definitional circle, we have no definitions at all. We are worse off than before.

V. PAROCHIAL SOLUTIONS

The remedy is that the folk psychophysics of colour, as we have envisaged it so far, is too abstract and general. It has a lot to say about

the causal relations between colours and their corresponding colour experiences. But it needs another chapter to tell us what distinguishes each colour from all the others (or each colour experience from all the others). Then the problem of multiplicity will be solved. Once we add the missing chapter, the folk psychophysics of colour will have a unique realization. (Or near enough. Maybe there will still be multiple realizations, but they will differ so little from one another that if we follow Bedard and say that the theoretical terms are indeterminate in reference, their indeterminacy will be far from radical. Some mild indeterminacy about the exact boundaries of the colours is credible enough. In fact, it is perfect determinacy that would be hard to believe.)

When we add the chapter that distinguishes the colours by name, how shall we do it? Most simply, we could give examples. ‘Red is the colour of pillar boxes’ has been a philosophers’ favourite, in England at least. Others might prefer to mention the people’s flag. I myself would rather say that red is the colour of the diagonal stripe on an Essendon Football Club jumper.

Likewise for more specific shades. Brunswick green, for instance, was the colour of locomotives of the Great Western Railway. (Of faded ones and freshly painted ones; of those painted before and after certain changes in the recipe for the paint; wherefore it cannot be a perfectly specific shade.)

Sometimes the examples are particular and soon forgotten. I ask you ‘What’s magenta?’ You make a mark with a crayon and say ‘This is.’ I trust you to have been cooperative and truthful in answering my question. So now I can say that magenta is the colour of the mark you made when I asked you what magenta was.

A week later I can no longer say that: I have forgotten the lesson. No harm done, if I retain the mental abilities which the lesson imparted to me. Any time I like, I can put myself into a state in which I can truly say ‘Magenta is the colour that typically causes the colour experience I am right now imagining’. Or I can use myself (and others can use me) as a living instrument: magenta is the colour such that I am disposed to say ‘magenta’ if you point to it and ask ‘What colour is that?’; and if I am disposed to be cooperative and truthful, and if the
magenta thing is at a suitable distance, and if the light and my visual capacities are normal.

(In calling myself a 'living instrument', I do not at all mean to suggest that my responses are mere mindless reflexes, rather than the reasoned judgements they might seem to be. No – bizarre cases aside, surely they are the latter. But for present purposes it does not matter which they are, so long as they are reliable indicators that magenta is present.)

The general chapter of the folk psychophysics of colour told us about the relations of resemblance among colour experiences, and about the derivative relations of resemblance among the colours themselves. So once we have distinguished some colours by the methods already mentioned, we can distinguish further colours by interpolation or triangulation: orange is reddish yellow; lemon yellow is ever-so-slightly greenish, and not at all reddish, yellow; and so on.

In these ways, I can add a chapter to the folk psychophysics of colour that will distinguish all the colours to which I have been properly introduced. And so can you, and so can almost anyone. (Even a blind man, except that he cannot use the methods that depend on his own abilities to imagine colour and to serve as a living instrument.) Once the chapter on distinguishing the colours has been added, we no longer have a severe problem of multiple realization. Now it is safe to say that our colour names and our colour-experience names refer to the approximately unique occupants of theoretical roles. Such moderate multiplicity as remains will result at worst in semantic indeterminacy regarding the exact boundary between one colour (or one colour experience) and another. All is well.

Or is it? The trouble is that I (or you, or almost anyone) drew upon ever so much parochial knowledge: information that is not common knowledge throughout the linguistic community. In Ballybunnion where the pillar boxes are green, or in New Haven where they are blue and not in the shape of pillars, people may or may not know about the red pillar boxes in far-off lands. And even if they do, does everyone else in the English-speaking world know that they do? Does everyone know that everyone else knows that everyone else knows that they do? Does . . . Surely not!
Likewise it is common knowledge among certain people—call them *footy people*—that Essendon wears a red diagonal stripe. But this is probably not known at all to certain other people, call them *rugby people*. So it is not common knowledge throughout the linguistic community. It takes all kinds to make the English-speaking world.

Surely very few other people know of the lesson whereby I was introduced to magenta, especially if I myself forgot it in a week. Nor do people who have never heard of me know anything about my ability to imagine magenta or to serve as an instrument for indicating its presence. Here too, the chapter on distinguishing the colours was built not upon common knowledge but upon parochial knowledge.

Maybe it is common knowledge that orange is reddish yellow. But interpolations and triangulations are useless by themselves. They only serve to distinguish more colours after we have distinguished some already. How to get started?

Our parochial solution is not good enough because language is a conventional, rational system of semantic coordination. I say something I take to be true under semantic interpretation *I*; you trust me to be truthful under *I* and also to be well-informed; and in this way you come to share my information—something that both of us wanted to happen. Coordination of truthfulness under *I* with trust under *I* would have worked just as well, even if *I* and *I* assigned opposite truth conditions to the same sentences. But miscoordination—of truthfulness under *I* with trust under *I* or vice versa—would have deprived us of the benefits of communication.

Any conventional system of coordination, semantic or otherwise, is rationalised by a potentially endless system of mutual expectations. Why do you drive on the left? Because you expect me to. Why do you expect me to? Because you expect me to expect you to. Why . . . In short: we both drive on the left because it is common knowledge between us that we both will. (Likewise *mutatis mutandis* if there are more drivers sharing the roads than two.) Our coordination is rationally sustained. It is not hard-wired and it is not mere luck. Even if we drive on the left by habit, as surely we do, the habit is sustained by

---

19 See my *Convention*. 

350
reason. If our common knowledge were eroded by doubt, or undermined by counter-evidence, we would swiftly change our ways.

So here is the current state of the problem: if we distinguish and name the colours by recourse to parochial information, how can we have the semantic common knowledge that is required if our talk about colour is to be part of a conventional, rational system of coordination?

VI. A SOLUTION THAT RELIES TOO MUCH ON LUCK

A few examples of a few colours are common knowledge throughout the linguistic community. Blood is red, well-watered leaves are green, the sky is blue, and flames are yellow. (Or rather, since we agreed to confine ourselves to surface colours, pictures of the sky are blue and pictures of flames are yellow. But let us waive the point.) These few examples will do to begin the job of distinguishing the colours, and maybe we can do the rest by interpolation and triangulation. Then the chapter of folk psychophysics that distinguishes the colours will not rely on parochial knowledge. It will be common knowledge throughout the linguistic community. So when we name the colours and colour experiences as occupants of theoretical roles, that naming will be a matter of semantic common knowledge. Our problem will be solved.

This solution is all very well for today, but consider the future. Leaves, well-watered or otherwise, will be long gone and forgotten. We will resort to brute force methods, powered by fusion, to regain oxygen from carbon dioxide. Flames will be forbidden. The polluted sky will occasionally be blue, when the smog clears, but just as often it will be red with nitric oxide, yellow with sulphur dust, or green with noxious scum never yet seen in our own time. People will have to spend their lives encased in protective armour, and so their blood never will be seen except by police and medical robots that have neither language nor colour vision. All records of better times will long since have been destroyed by the Hedonic Legion because of their tendency to sadden those who read them. Yet there will still be colour — lots of it. Graffiti will cover everything.
Could there still be semantic common knowledge of colour language? I think so. That means that our presently proposed solution, though it may well succeed under fortunate conditions, cannot be the only possible solution.

VII. A SOLUTION THAT IS UNAVAILABLE TO MATERIALISTS

Some philosophers think that each sort of colour experience has a simple, ineffable, unique essence that is instantly revealed to anyone who has that experience. When I was shown the crayon mark and told that it was magenta (and I believed what I was told, and it was true) straightway I knew all there is to know about experience of magenta. I knew that it was the experience with the simple, ineffable, unique essence $E$. And that is all there is to it.\textsuperscript{20} (Or perhaps it is the colour magenta itself that has the simple, ineffable, unique essence that is instantly revealed to each beholder, or anyway to each beholder with normal visual capacities in normal light.)

If this doctrine of revelation were true, presumably it would be obviously true. Even those philosophers who denied it would know it in their hearts, once they had seen a few colours and experienced the workings of revelation for themselves. Thanks to its obviousness, the general doctrine of revelation could readily become common knowledge throughout the linguistic community. Then if it were also common knowledge that everyone in the community becomes acquainted with magenta early in life (and if the community were properly dismissive of sceptical doubts about inverted spectra, etc.) it would be common knowledge throughout the community that magenta is the colour that typically causes experience with essence $E$. (Or perhaps, according to the other version of the story, that magenta is the colour that itself has the simple, ineffable, unique essence $E$.) The same would go, of course, for all the other colours that we become acquainted with.

and have names for. So here we would have the requisite common knowledge of the semantics of the language of colour. It would be ineffable common knowledge, but what is the harm in that? The doctrine of revelation would solve our problem.

The only remaining difficulty is that the doctrine is false. At any rate, it is false for colour experiences (and colours themselves). At any rate, it is false by materialist lights – and we have pledged ourselves non-negotiably to materialism. The essence of a colour experience is not at all simple, not at all ineffable, not at all easily known. Probably it is a matter of neural firing patterns; but if not that, something equally esoteric. Likewise for the essences of the colours themselves.

The doctrine of revelation is tailor-made to solve our problem. But we materialists must dismiss this ‘solution’ as a useless piece of wishful thinking.

21 Maybe revelation is true in some other cases – as it might be for the part–whole relation.

22 You should have noticed that my continuing difficulties result from a continuing effort to associate descriptive senses with the names of the colours and the colour experiences. Is that effort misguided? Did not Kripke and his allies refute the description theory of reference, at least for names of people and places? Then why should we expect descriptivism to work any better for names of colours and colour experiences? For the supposed refutation of descriptivism, see *inter alia* Saul Kripke, ‘Naming and Necessity’ and Keith Donnellan, ‘Proper Names and Identifying Descriptions’, both in *Semantics of Natural Language* (eds.) Donald Davidson and Gilbert Harman (Dordrecht: Reidel, 1972); and Michael Devitt, *Designation* (New York: Columbia University Press, 1981) pp. 3–25.

I disagree. What was well and truly refuted was a version of descriptivism in which the descriptive senses were supposed to be a matter of famous deeds and other distinctive peculiarities. A better version survives the attack: *causal descriptivism*. The descriptive sense associated with a name might for instance be ‘the place I have heard of under the name “Taromeo”’ or maybe ‘the causal source of this token: Taromeo’, and for an account of the relation being invoked here, just consult the writings of the causal theorists of reference. A brief mention of causal descriptivism, with credit to Robert Nozick, is to be found in ‘Naming and Necessity’ itself (p. 349, fn. 38); see also my ‘Putnam’s Paradox’, *Australasian Journal of Philosophy* 62 (1984) pp. 226–227 (reprinted in this volume as Chapter 2); and Fred Kroon, ‘Causal Descriptivism’, *Australasian Journal of Philosophy* 65 (1987) pp. 1–17.

Causal descriptivism explains, as the causal theory itself does not, (1) how the way reference is fixed by means of causal chains is part of our semantic common

353
So much the worse for materialism? Not if we have an alternative solution. And we do.

VIII. HOW MUCH COMMON KNOWLEDGE DO WE REALLY NEED?

Consider again the case of the footy people and the rugby people: two subcommunities of the English-speaking linguistic community. For simplicity, I suppose that every English-speaker is either a footy person or a rugby person. Or both? I do not need to say whether or not the two subcommunities overlap. But they do not overlap much. Plenty of footy people know nothing of rugby, and vice versa.

The two subcommunities mingle. Often, you cannot tell whether you are talking to a footy person or a rugby person, so long as you are not talking about sport.

It is common knowledge among the footy people that the diagonal stripe on the Essendon jumper is red, that the widest horizontal stripe on the Footscray jumper is red, that Swans are mostly red, . . . None of these things are known to rugby people, except for those of them, if any, who are footy people as well.

Likewise, some examples of red things are common knowledge among the rugby people. But a footy person like me — a footy person who is not also a rugby person — has no idea what these examples are.

The footy people have their version of the added chapter of folk psychophysics that distinguishes the colours. And when they define theoretical terms as names for the occupants of theoretical roles, they arrive at their version of a definition of ‘red’. This definition is part of the semantic common knowledge of the footy people, but it is not common knowledge among the English-speaking community as a whole.

Likewise the rugby people have their version of the chapter and the definition. But the two definitions are parochial. In fact, no definition knowledge, and (2) how it depends on our contingent semantic conventions. As it does: distinctive-peculiarity descriptivism could have been true instead, inconvenient though that would have been, but our actual conventions of language are such as to render it false.
of ‘red’ is part of the semantic common knowledge of the entire community.

(Not even ‘the colour of blood’? After all, the present example is not set in the nightmare future previously imagined. Still, these rugby people are peculiar! Look at their funny ideas about the proper shape for goalposts. Might they not also have funny ideas about blood? And even if you know they do not, do you know that everyone knows? Do you know that everyone knows that everyone knows? Do you know . . .)

When footy people and rugby people mingle, they talk. Sometimes they talk about the colours of things. It never seems to happen that there is a misunderstanding. Why not? You might expect that sometimes a footy person, wrongly thinking that his conversational partner was a footy person too, would call something red because it is red in the footy sense. Then the rugby person he was talking to, wrongly thinking that the speaker was another rugby person, would believe that the thing was red in the rugby sense — and he would get a surprise when it turned out not to be.

If such conversational mishaps occurred, we would notice them. (Or at least we would hear from others who noticed them.) And if conversationalists took precautions to prevent mishaps, we would notice that. And we would hope that those who belong to both subcommunities — if there are any of those — would warn the rest of us about the risk of misunderstandings. And yet we never notice mishaps or precautions or warnings. Why? The best explanation is that there is in fact no risk of misunderstanding, because exactly the same things that are red in the footy sense are red also in the rugby sense.

Exactly the same things; or near enough. Disagreement on a few things that both subcommunities alike would treat as borderline cases would be harmless and hard to notice.

In this way, even though neither definition was common knowledge throughout the entire linguistic community, there might still be existential common knowledge: common knowledge (1) that some definition of ‘red’ was common knowledge among the footy people, and (2) that some definition of ‘red’ was common knowledge among the rugby people, and (3) that these two definitions agreed (or near enough) about which things were red.

355
If there were only that much common knowledge throughout the entire community, that would suffice. Existential common knowledge, even without fully specific semantic common knowledge, would sustain rational semantic coordination, even between footy people and rugby people conversing about what is red.

If the footy and rugby definitions of 'red' agree in extension, that will avoid the simplest sort of mishap in communication. But if the two definitions differ in intension, is there not still some risk of a more subtle sort of mishap? Suppose two people are talking; and unbeknownst to them, one is a footy person and one is a rugby person. So long as they talk about which things are red, they have no problem. But what happens if they talk instead about which things would be red under various counterfactual circumstances? Suppose the footy definition of red is simply: 'the colour of the Essendon diagonal stripe'. Now if the Essendon Football Club were taken over by the Environmentalist Party, who are no respecters of tradition and never pass up a chance for self-advertisement, what would then be red: blood or leaves? If leaves would then be red in the footy sense yet blood would remain red in the rugby sense, our conversationalists would be in trouble if they were to try to discuss this hypothetical case. Their mismatched definitions would cause a failure of communication after all.

Yet this sort of mishap also is never observed. Why not? Because the footy definition of 'red' is not simply 'the colour of the Essendon stripe'. Rather it is: 'the colour of the Essendon stripe actually, nowadays, and hereabouts'. The latter definition is 'rigidified'. Unlike the simple one, it can be relied upon not to shift its reference when we talk about unactualized hypothetical cases, or about actualities at remote times and places.23 The rugby definition, whatever it may be, is likewise rigidified. And it is common knowledge throughout the entire linguistic community that both definitions are rigidified. In this

way mishaps of communication are averted even when conversation ranges beyond the actual or beyond the here and now.

(I do not say that the simple phrase 'the colour of the Essendon stripe' is definitely unrigidified. Rather, I suppose it to be ambiguous with respect to rigidification. Hence the custom of rigidification has two separate advantages. One is that, as we have already seen, it can prevent failures of communication between conversational partners who use different definitions. The other is that it prevents ambiguity in the speech of a single person. The cost of a custom of rigidification is that it makes unrigidified things harder to say. To gain the advantage, and pay the cost, it is not of course required that we always signal rigidification with some special form of words.)

Do the footy people’s rigidified 'red' and the rugby people’s rigidified 'red' have the same intension or do they not? A straight answer would be unwise. The two have something beyond their extension in common, and something not in common. But whether ‘intension’ is the word for what they have in common or the word for what they have not in common is an unsettled matter.

It is a simplification to imagine that the footy people define ‘red’ as ‘the colour of the Essendon stripe’ (duly rigidified). Many examples of red are common knowledge among them. Any selected one of these, or any selection of several of these, would do for a definition. It would serve no good purpose for them to make an official choice. If they have not done so, that is only to be expected.

Or perhaps there is a default choice: the biggest class of examples that does not overstep the bound of their common knowledge. If all the examples that are common knowledge among them are mentioned in their version of the chapter of folk psychophysics that distinguishes the colours, and if red is defined to be the occupant of the resulting folk-theoretical role, that is just what will happen.

But there might be a few rotten apples in the barrel: items of common knowledge that are not knowledge at all but rather error. I said that one example of red was the widest horizontal stripe on the Footscray jumper. But that begged a question about what is a stripe and what is a stripe-shaped part of the background. Arguably, the widest stripe of all is not the red stripe, but rather the very wide blue stripe right at the bottom.
As already noted, we should be prepared to say that red is a *near-enough* occupant of a role. But which one? Once we retreat from demanding perfection, there might in principle be competing near-enough occupants. So even given a default choice, there is still some risk of semantic indecision in naming the colours. Doubtless the risk is slight, doubtless the indecision risked is also slight. Yet we can still ask what to do about it. Answer: nothing. Not unless there turns out to be a real problem. The advantage of being prepared is not worth the bother of solving countless problems in advance when most of them will never arise.

The footy people and the rugby people were two linguistic subcommunities, both large. Of course the subcommunities could be smaller and more numerous, so long as the necessary existential common knowledge was available. In the extreme case, I do not see why we could not have very many subcommunities of one person each, each one with his own private collection of coloured box-beetles. At that limit, the specific common knowledge within the subcommunity would vanish. Existential common knowledge would have to do the whole job of sustaining a rational system of linguistic coordination. I see no reason why this extreme case would be impossible in principle. But certainly it would be a far cry from the real world.
MILL AND MILQUETOAST

David Lewis

1. Toleration

We are fortunate to live under institutions of toleration. Opinions that many of us deem false and pernicious are nevertheless held, and even imparted to others, with impunity. This is so in part because we hold legal rights to freedom of thought and freedom of expression. Not only do these legal rights exist; they enjoy widespread support. Any effort to revoke them would be widely opposed. Those whose opinions were threatened with suppression would find many allies, even among those who most deplored their opinions.

But legal rights are far from the whole story. The institutions of toleration are in large part informal, a matter not of law but of custom, habits of conduct and thought. Even when the law lets us do as we like, many of us do not like to do anything that would make people suffer for the opinions they hold, or hinder their expression of their opinions. We may choose our friends and our casual acquaintances as we please, and we are certainly free to shun those whose opinions we find objectionable; but many of us exercise this freedom half-heartedly, or with a bad conscience, or not at all. An editor or a bookseller has plenty of discretion to assist in the spreading of some opinions and not others, and might weigh many different considerations in deciding what to publish or what to sell; but might very well think it wrong to give any weight at all to whether an author’s opinions are true or false, beneficial or dangerous.

Not only do customs of toleration complement legal rights; to some extent, the customs may even substitute for the rights. Doubtless it is a good idea to entrench toleration by writing it into the constitution and the statutes. But the measure of toleration need not be legalistic. The real test is: what can you get away with? What opinions can you express without fear of reprisal? To what extent can you reach your audience, if it wants to be reached? What can you read or hear without fear of reprisal? If the samizdat circulate freely, and you needn’t be a hero to write or produce or read them, that is not yet good enough. But it is very much more than nothing. A country where banned books become contraband best-sellers is worse off than a country where books cannot be banned at all; but their difference is not great when we compare them both with a country where banned books really do disappear.

Toleration need not be everywhere to be effective. An atheist is not welcome everywhere—who is?—and if he cannot find toleration in the place he most wants to be, to that extent he suffers for his opinions. But if there are many
and varied places where an atheist is perfectly welcome, then he doesn't suffer much. Likewise, it is essential that there should be some magazines where atheism may be published; it matters little that there are many others where it may not. Even a handful of urban and rural bohemias can go a long way toward making toleration available to those who have need of it. So if an intolerant majority do not bestir themselves to clean up the bohemias, then even they are participating in the institutions of toleration.

2. Mill's Project

That is what toleration is. Now, what is it good for? In his On Liberty, Mill undertakes to give it a utilitarian defence.¹ That is, he undertakes to show that its expected benefits outweigh its expected costs. But he is no simplistic Benthamite: 'I regard utility as the ultimate appeal on all ethical questions; but it must be utility in the largest sense, grounded on the permanent interests of man as a progressive being.' (p.14) So whatever commitments Mill may incur elsewhere, here we needn't worry whether matters of human flourishing somehow translate into a common currency of pleasure and pain.

All the same, we had better not take utility in too large a sense. 'I forego any advantage from the idea of an abstract right as a thing independent of utility.' (p.14) So it will not do to claim that the infringement of such 'abstract' rights is itself one cost to be weighed in the balance as a component of 'utility', whether with infinite weight (as a 'side constraint') or just as one consideration among others.

There seems to be another rule to Mill's game, unannounced but manifest in his practice. Let us make it explicit. It is the rule of neutralism. Suppose we have a dispute, say between believers and atheists, and suppose the believers want to suppress what they take to be the false and dangerous opinions of the atheists. Some utilitarian atheist might defend toleration thus: in the first place there is no God, therefore no harm can come of holding beliefs offensive to God. Nor can the spread of atheism do harm in any other way. Therefore suppressing atheism has no benefits to match its costs. Therefore toleration would be better. This defence is utilitarian, sure enough; but unMillian. The Millian defender of toleration makes his case without taking sides in the dispute. Of course he may argue from factual premises—no utilitarian could go far without them!—but not from factual premises that are part of the very dispute between the suppressors and the suppressed. It is Mill's ambition to defend toleration even when questions remain disputed, therefore it will not do to require some settlement of the dispute before the case for toleration can be completed.

The neutralism of Mill's practice goes further. Some utilitarian might say to the believers that according to their opinion toleration maximises utility because God is offended more by the cruelty of inquisitors than by the impudence of atheists; and might argue to the atheists that according to

their opinion toleration maximises utility because there is no God to be offended. This playing both sides of the street is a valid argument by separation of cases: if A then toleration maximises utility, if B then toleration maximises utility, therefore toleration maximises utility in either case. But however valid it may be, this too is unMillian. In a Millian defence of toleration, not only must the factual premises be common ground between the two sides; also a uniform and non-disjunctive argument must be addressed to both. The Millian invites both sides to assent to a single, common list of the benefits of toleration and costs of suppression. This common list is supposed to have decisive weight in favour of toleration. One or the other side may have in mind some further costs and benefits that obtain according to its own disputed opinions, perhaps including some that count in favour of suppression; but if so, these considerations are supposed to be outweighed by the considerations on the neutral common list.

Why do I ascribe a rule of neutralism to Mill? Only because I never see him violate it. Not because he states and defends it—he does not. And not because it is in any way essential to his project of defending toleration by appeal to utility. On the contrary. To decide whether he himself should think that toleration maximises utility, Mill must sum up all the relevant costs and benefits according to his own opinions. To persuade me that toleration maximises utility, he must sum them up according to my opinions (perhaps my original opinions, or perhaps my new opinions after he is done persuading me). It is irrelevant whether the opinions are disputed or undisputed. But Mill is not doing his private sums, nor is On Liberty addressed to some one person in particular. It is meant to persuade an audience with varied opinions. It's hard to play both sides of the street when you're writing for both sides at once! Better for Mill if he can address the whole of his case to the whole of his audience. He can do so, if a neutral common list suffices to outweigh whatever other disputed costs and benefits there may be. Hence the rule of neutralism. It makes no sense as a constraint on utilitarian argument per se, but plenty of sense as part of Mill's strategy of persuasion.

3. Self- and Other-Regarding

The main principle of On Liberty, second only to the ultimate appeal to utility in the largest sense, is that ‘the sole end for which mankind are warranted . . . in interfering with the liberty of action of any of their number is self-protection.’ (p.13) It is notoriously difficult to get clear about the requisite line between self- and other-regarding action. But it is worth a digression to see why the principle and the difficulty need not concern us here.

First, and decisively, because the protection of self-regarding conduct is in any case derived from the ultimate appeal to utility. It has no force of its own to justify toleration if the direct appeal to utility fails.

Second, in addition, because if an opinion is not held secretly, but is expressed in a way that might persuade others, that is other-regarding: both because of the effect that the opinion may have on the life of the convert
and because of what the convert might do, premised on that opinion, which might affect third parties.

Mill is confusing on this point. 'The liberty of expressing and publishing opinions may seem to fall under a different principle, since it belongs to that part of the conduct of an individual which concerns other people; but, being almost of as much importance as the liberty of thought itself and resting in great part on the same reasons, is practically inseparable from it.' (p.16) What kind of argument is this? Other-regarding conduct is not in general protected by reason of inseparability from private thought, as will be plain if someone's religion demands human sacrifice.

4. Mill's Tally

I do not believe that a utilitarian defence of toleration, constrained by Mill's rule of neutralism, has any hope of success. I make no fundamental objection to broadly utilitarian reasoning, at least in such matters as this. It's just that I think the balance of costs and benefits will too easily turn out the wrong way. When we tally up the benefits of toleration that can be adduced in a neutral and uniform way, they will just not be weighty enough. They will fall sadly short of matching the benefits of suppression, calculated according to the opinions of the would-be suppressors.

I begin the tally with the items Mill himself lists.

**Risk of error.** True and beneficial opinion might be suppressed in the mistaken belief that it is false and harmful.

Mill says just 'true' and 'false'; the utilitarian argument requires that we say 'beneficial' and 'harmful'; there's no guarantee that these coincide, but for simplicity let's suppose they do.

**Mixture.** Truth and error may be found combined in one package deal, so that there's no way of suppressing the error without suppressing truth as well.

**Dead dogma (reasons).** Unless received opinion 'is suffered to be, and actually is, vigorously and earnestly contested, it will . . . be held in the manner of a prejudice, with little comprehension of its rational grounds.' (p.64)

**Dead dogma (meaning).** Further, 'the meaning of the doctrine itself will be in danger of being lost or enfeebled, and deprived of its vital effect on the character and conduct . . . cumbering the ground and preventing the growth of any real and heartfelt conviction from reason or personal experience.' (p.64)

Mill counts deadness of dogma as a harm only in case received opinion is true. But perhaps he should also think it worse, from the standpoint of human flourishing, that error should be held as dead dogma rather than in a real and heartfelt and reasoned way.

Mill's guess about what will happen if received opinion is vigorously contested seems remarkably optimistic. Will there be debate at all, and not
just warfare? If there is debate, will it help the debaters think through their positions, or will they rather throw up a cloud of sophistries? If they think things through, will they discover unappreciated reasons or bedrock disagreement?

5. The Tally Extended

Mill’s list so far seems too short. Why not borrow from the next chapter of On Liberty also? Then we could add—

Individuality. If diversity is of value, and thinking for oneself, and thoughtful choice, why aren’t these things of some value even when people think up, and thoughtfully choose among, diverse errors?

Building character. The more chances you get to think and choose, the better you get at it; and being good at thinking and choosing is one big part of human flourishing. Freedom as a social condition offers exercises which conduce to freedom as a trait of character. Practice makes perfect.

This too seems more a piece of armchair psychology than a firm empirical result. Travelers’ tales suggest that the hard school of the east sometimes does better than the free and easy west at building just such character as Mill rightly values. If we like guessing, we might guess that when it comes to building character, freedom and competent repression both take second place—what does best is repression bungled, with gratuitous stupidity and cruelty. That speculation seems at least as likely as Mill’s—but responsible utilitarian calculation should put little faith in either one.

We noted that truth and error might be found combined in a package deal. Then if we suppress the error, we lose truth as well. But the same thing can happen even if the error we suppress is unmixed with truth.

Transformation. Future thinkers may turn our present errors into truth not just by filtering out the false parts but in more complicated ways. They may find us standing on our heads, and turn us on our feet. They may attend to old questions and give them new answers. They may borrow old ideas and transplant them into new and better settings. They may put the old errors to use in metaphors and analogies. If we suppress errors that might have been the germ of better things to come, we block progress. Does progress conduce to utility?—We may hope so, at least if it is ‘utility in the largest sense, grounded on the permanent interests of man as a progressive being’.

Mill’s lists of harms and benefits feature the high-faluting, interesting, speculative ones. He omits the obvious.

The insult of paternalism. If I paternalise over you, and in particular if I prevent you from being exposed to some seductive heresy, my action is manifestly premised on doubt of your competence, and on confidence in my own. You are likely to take offence both at my low opinion of you and at my pretension of superiority. No less so, if you acknowledge
that I am indeed more competent than you are to govern your life. Bad enough it should be true! Do I have to rub it in?

(This is a different thing from the alleged insult of denying that you have rights. For (1) no similar insult is given when Bentham tells you that natural rights are nonsense upon stilts, yet he denies that you have rights more clearly than any paternalist does; (2) the insult may still be there even if you too are of Bentham’s opinion; or (3) if you think that you once had rights but have freely given them away to me.)

More obviously still, there are —

*The secret police.* To do an effective job of suppression, it is necessary to build a system of informers and dossiers. Once in place, the means of suppression may be taken over and turned to new purposes. They might be used to advance the ambitions of a would-be tyrant—something all would agree (before it began, at least) in counting as a cost.2

*The dungeon.* If you wish to express or study proscribed opinions, and someone stops you, you will be displeased that your desires are frustrated. And if you are determined to go ahead, the only effective means of stopping you—the dungeon, the gulag, the asylum, the gallows—may prove somewhat unfelicific.

This completes our neutralist tally, our list of considerations that are meant to be accepted by all parties to disputed questions. One way or another, and even if we receive Mill’s armchair psychology with all the doubt it deserves, we still have some rather weighty benefits of toleration and costs of suppression. But of course that’s not enough. Mill wins his case only if the benefits of toleration outweigh the costs—and not only according to his own opinions, but according to the opinions of those he seeks to dissuade from suppressing. The cost of toleration, lest we forget, is that dangerous errors may flourish and spread.

6. *The Inquisitor Reads Mill*

McCloskey has written that ‘many Christian liberals appear to be especially muddled, for, as Christians, they regard eternal salvation and moral living as being of tremendous importance and as being goods as valuable as freedom. Yet many of them deny the state even the abstract right to aid truth, morality, and religion and to impede error and evil, while at the same time they insist on its duty to promote the good of freedom. Their implicit value judgment is so obviously untenable that one cannot but suspect that it has not been made explicit and considered in its own right.’ In the same vein, Quine: ‘If someone firmly believes that eternal salvation and damnation hinge on embracing his particular religion, he would be callous indeed to sit

2. Another possibility is that the means of suppression might be turned to a new purpose which, like the original suppression, serves utility according to the opinions of some but not of others. The Informer of Bray, like the Vicar, might serve his new masters as willingly as he served the old. But this danger, however weighty it might seem to some, is inadmissible under the rule of neutralism.
tolerantly back and watch others go to hell. To dramatise their point, I imagine the Inquisitor: a thoughtful Christian, benevolent by his own lights, far from muddled and far from liberal. Can Mill persuade him to change his intolerant ways?

The Inquisitor, as I shall imagine him, is the very man Mill ought to be addressing. He agrees completely with Mill that the ultimate appeal is to utility in the largest sense. He claims no infallibility. Indeed his faith is infirm, and he is vividly aware that he just might be making a tragic mistake. He is satisfied—too quickly, perhaps—that Mill is an expert social psychologist, who knows whereof he speaks concerning the causes of dead dogma and the causes of excellent character. In short, he grants every item in the neutralist tally of costs and benefits.

His only complaint is that the tally is incomplete. He believes, in fact, that the included items have negligible weight compared to the omitted item. Heresy, so the Inquisitor believes, poisons the proper relationship between man and God. The heretic is imperfectly submissive, or sees God as nothing but a powerful sorcerer, or even finds some trace of fault in God's conduct. The consequence is eternal damnation. That is something infinitely worse than any evil whatever in this life; infinitely more weighty, therefore, than the whole of the neutralist tally. Further, damnation is not just a matter of pain. (Hellfire is no part of it, just an inadequate metaphor for what really happens.) Damnation is harm along exactly the dimension that Mill wanted us to bear in mind: it is the utter absence and the extreme opposite of human excellence and flourishing.

The Inquisitor also believes that heresy is contagious. The father of lies has fashioned it with all his cunning to appeal to our weaknesses. There is nothing mechanical about it—those never exposed to heretical teachings sometimes reinvent heresy for themselves, those who are exposed may withstand temptation—but still, those who are not exposed are a great deal safer than those who are.

The Inquisitor also believes that if he is ruthless enough in suppressing heresy, he may very well succeed. Not, of course, in eradicating heresy for all time; but in greatly reducing the incidence of exposure, and consequently in saving a great many souls from damnation.

Note well that the Inquisitor does not think that he could save the souls of heretics by forced conversion. He accepts the common wisdom that this cannot be done: forced conversion would be insincere, so it would be worthless in the sight of God. He knows no way to save the heretics themselves. What he could do by suppressing heresy, so he thinks, is to save many of those who are not yet heretics, but would succumb if exposed to heretical teachings.

The Inquisitor does not relish the suffering of heretics. As befits a utilitarian, he is moved by benevolence alone. He hates cruelty. But he heeds the warning:

‘if you hate cruelty, remember that nothing is so cruel in its consequences as the toleration of heresy.’

Therefore the Inquisitor concludes, even after discounting properly for his uncertainty, that the balance of cost and benefit is overwhelmingly in favour of suppression. Mill’s case for tolerating heresy is unpersuasive. In fact it is frivolous—serious matters are at stake! You might as well oppose the suppression of heresy on the ground that dungeons cost too much money.

Mill has lost his case.

This is not to say that the Inquisitor stumps utilitarianism itself. Mill was trying to bring off a tour de force: to abide by his self-imposed rule of neutralism, and yet win the argument against all comers. A more modest utilitarian might proceed in any of three ways.

One way for the utilitarian to deal with the Inquisitor is not to argue with him at all. You don’t argue with the sharks; you just put up nets to keep them away from the beaches. Likewise the Inquisitor, or any other utilitarian with dangerously wrong opinions about how to maximise utility, is simply a danger to be fended off. You organise and fight. You see to it that he cannot succeed in his plan to do harm in order—as he thinks and you do not—to maximise utility.

A second way is to fight first and argue afterward. When you fight, you change the circumstances that afford the premises of a utilitarian argument. First you win the fight, then you win the argument. If you can make sure that the Inquisitor will fail in his effort to suppress heresy, you give him reason to stop trying. Though he thinks that successful persecution maximises utility, he will certainly agree that failed attempts are nothing but useless harm.

Finally, a modest utilitarian might dump the rule of neutralism. He might argue that, according to the Inquisitor’s own opinions, there are advantages of toleration which are more weighty than those on the neutralist tally and which the Inquisitor had not appreciated. Or he might start by trying to change the Inquisitor’s mind about the facts of theology, and only afterward try to demonstrate the utility of toleration. He might try to persuade the Inquisitor to replace his present theological opinions by different ones: atheism, perhaps, or a religion of sweetness and light and salvation for all. Or he might only try to persuade the Inquisitor to be more sceptical: to suspend judgement on matters of theology, or near enough that the uncertain danger of damnation no longer outweighs the more certain harms that are done when heresy is suppressed.

7. The Assumption of Infallibility

Mill does at one point seem to be doing just that—supporting toleration by supporting scepticism. If he did, he would not be observing the rule of neutralism. He would be putting forward not an addition to whatever

his reader might have thought before, but rather a modification. And he would be a fine old pot calling the kettle black. Part of his own case rests on far-from-certain psychological premises.

But the appearance is deceptive. Mill's point when he says that 'all silencing of discussion is an assumption of infallibility' (pp.21-22) is not that we should hesitate to act on our opinions—for instance by silencing discussion we believe to be harmful—out of fear that our opinions may be wrong. For Mill very willingly agrees with the hypothetical objector who says that 'if we were never to act on our opinions, because those opinions may be wrong, we should leave all our interests uncared for, and all our duties unperformed . . . There is no such things as absolute certainty, but there is assurance sufficient for the purposes of human life. We may, and must, assume our opinion to be true for the guidance of our own conduct.' (pp.23-24) Mill's real point is that if we are duly modest and do not assume ourselves infallible, we should have confidence in our opinions only when they have withstood the test of free discussion. A sceptic is like a traffic cop: he admonishes us to slow down in our believing. Whereas Mill is like the traffic cop in the tire advertisement: 'If you're not riding on Jetzon tires—slow down!' That cop doesn't want us to slow down—he wants us to buy Jetzon tires. Free discussion is the Jetzon tire that gives us license to speed, fallible though we be. To dare to do without Jetzon tires is to overrate your skill as a driver; to do without free discussion is to assume yourself infallible. 'Complete liberty of contradicting and disproving our opinion is the very condition which justifies us in assuming its truth for purposes of action; and on no other terms can a being with human faculties have any rational assurance of being right.' (p.24) Mill thus assures us that if we do meet the condition, then we are justified in acting on our opinions.

Our Inquisitor, if he takes Mill's word for this as he does on other matters, will not dare suppress heresy straightaway. First he must spend some time in free discussion with the heretics. Afterward, if they have not changed his mind, then he will deem himself justified in assuming the truth of his opinion for purposes of action; which he will do when he goes forward to suppress heresy, and burns his former partners in discussion at the stake.

Compare Herbert Marcuse, who advocated 'withdrawal of tolerance from regressive movements before they can become active; intolerance even toward thought, opinion, and word, and finally, intolerance . . . toward the self-styled conservatives, to the political Right' during the present 'emergency situation'. If tolerance is withdrawn only after Marcuse has enjoyed it for many years, Mill cannot complain that Marcuse has not yet earned the right to act on his illiberal opinions.

8. Dangerous Opinions

The Inquisitor, apart from his anachronistic utilitarianism, is just an ogre out of the past. Might Mill's defence work well enough, if not against just any imaginable foe of toleration, at least against any we are likely to meet in the present day? I doubt it. To be sure, some of us nowadays are sanguine about dangerous opinions. Whatever harm opinions may do under other conditions, we think they pose no present danger in our part of the world. The neutralist tally is all the defence of toleration we need. But others of us think otherwise: they think that some of the people around them hold opinions that are not only false but harmful. I predict that for many pairs of my readers—perhaps a majority of pairs—one of the pair holds some opinion that the other would find profoundly dangerous.

It might be a religious or irreligious opinion that conduces, in the opinion of the other, to contempt for oneself, for other people, for the natural world, or for God.

It might be a political opinion favouring some social arrangement which, in the opinion of the other, is a trap—an arrangement which makes most people's lives degraded and miserable, but which gives a few people both a stake in its continuation and the power to prevent change.

It might be an opinion belittling some supposed danger which, in the opinion of the other, requires us to take urgent measures for our protection. It might be the opinion that we need not worry about environmental hazards, or nuclear deterrence, or Soviet imperialism, or AIDS, or addictive drugs. It might be an opinion which, in the opinion of the other, is racist or sexist and thereby fosters contempt and oppressive conduct.

It might be a moral opinion (say, about abortion) which, in the opinion of the other, either condones and encourages wickedness or else wrongly condemns what is innocent and sometimes beneficial.

In each of these cases, important matters are at stake. In each case, the stakes involve a great deal of 'utility in the largest sense, grounded on the permanent interests of man as a progressive being.' To be sure, these cases are less extreme than that of the Inquisitor and the heretics. We have no infinite outweighing. Still, they are extreme enough. In each case, the disutility that is feared from the dangerous opinion seems enough to outweigh all the advantages of toleration according to the neutralist tally. And this remains so even if we discount all around for uncertainty, duly acknowledging that we are fallible.

In each case, therefore, if effective suppression were feasible, it would seem frivolous for the foe of the dangerous opinion to stay his hand because of any consideration Mill has an offer. If he does stay his hand, it seems as if he lets geniality or custom or laziness stand in the way of his wholehearted pursuit of maximum utility.

9. Morris

Take our contemporary, Henry M. Morris. He thinks, for one thing, that 'Evolution is the root of atheism, of communism, nazism, behaviorism, racism,
economic imperialism, militarism, libertinism, anarchism, and all manner of anti-Christian systems of belief and practice.'7 He thinks, for another thing, that in history and the social sciences, 'it is especially important . . . that the teacher gives a balanced presentation of both points of view [evolutionist and creationist] to students. Otherwise the process of education for living becomes a process of indoctrination and channelization, and the school degenerates into a hatchery of parrots.'8 At any rate, he says both these things, and let us take him at his word. Doubtless he mainly has in mind the 'balanced treatment' versus purely evolutionist teaching. But what he says, and his argument for it, apply equally to the 'balanced treatment' versus the purely creationist teaching we might have expected him to favour. So evolution is dangerous in the extreme, yet it is not to be suppressed—it is not even to be left out of the curriculum for schoolchildren—lest we hatch parrots! ('Parrots', I take it, are the same thing as those who hold their opinions as dead dogma.) How can Morris possibly think that the harm of hatching parrots is remotely comparable to the harm done by 'balanced presentation' that spreads evolutionist ideas? How dare he give this feeble Millian reason for tolerating, and even spreading, such diabolically dangerous ideas? Surely, by his own lights, he is doing the Devil's work when he favours balance over suppression.

10. Milquetoast

Mill's defence—who needs it? Perhaps the sceptical who, when told any story about the harmful effects of dangerous opinions, will find it too uncertain to serve as a basis of action? Or perhaps the apathetic, who may believe the story but not think the harm really matters very much? No, because the sceptical and the apathetic will be equally unimpressed by Mill's own story about the harmful effects of suppression. Nor would Mill have wanted to address his argument to the sceptical or the apathetic. That is not how he wants us to be. He wants us to have our Jetzon tires exactly so that we may speed. He favours vigour, dedication, moral earnestness.

I suggest that Mill's defence of toleration might best be addressed to Caspar Milquetoast, that famous timid soul.9 Doubtless he too is not the pupil Mill would have chosen, but at least he is in a position to put the lesson to use.

Milquetoast does have opinions about important and controversial matters. And he does care. He cares enough to raise his voice and bang the table in the privacy of his own house: asked if he wants Russian dressing on his salad, the answer is 'NO!' He isn't always timid. (p.185) But when he is out and about, his main goal is to avoid a quarrel. All else takes second place. He knows better than to talk to strangers on vital topics: asked what he thinks of the Dodgers' chances, he'd 'rather not say, if you don't mind'.

(p.162) And when his barber, razor in hand, asks how he's going to vote, Milquetoast fibs: 'Why-uh-er-I don't get a vote. I've been in prison—stir I mean—and I've lost my citizenship'. (p.183)

Milquetoast thinks, let us suppose, that it is a dangerous mistake to ignore the threat of Soviet imperialism. He would be hard put to explain why a rosy view of the evil empire is not dangerous enough to be worth suppressing. But he knows that this opinion is controversial. He knows that others think that the Soviet threat is bogus, and that the only real threat comes from our efforts to resist the bogus threat. How horrid to have to dispute these matters—as he surely would if he dared to suggest that the dangerous mistake should be suppressed. What to do? — Solution: bracket the controversial opinions. Keep them as opinions, somehow, in some compartment of one's mind, but ignore them in deciding what is to be done. In questions of suppression and toleration, in particular, appeal to uncontroversial considerations only. Conduct the discussion according to Mill's rule of neutralism. Then all hands can perhaps agree that the neutralist tally is right so far as it goes. And without the airing of disagreeable disagreement, we can go no further. Settle the question without acrimony, then, and we must settle it in favour of toleration. Those compartments of the mind that fear the dangerous consequences of the tolerated opinions should hold their tongues, lest they get us into strife.10

Milquetoast, of course, is an incompetent maximiser of utility. His conduct may be fortunate enough, if there turn out to be better reasons for toleration than we have yet considered. But his thought is simply shocking—he systematically declines to be guided by the whole of his system of opinions, ignoring the part that would engage him in unpleasant dispute. Nor is he at all keen to improve the quality of his thought by entering into discussion. That is why Mill should not be proud to have Milquetoast as his star pupil.

11. **A Treaty of Toleration**

To see how toleration can find a better utilitarian foundation, let us return to our story of the Inquisitor and the heretics. The Inquisitor thinks that the heretics hold a dangerous opinion—dangerous enough to be well worth suppressing, despite all the considerations on the neutralist tally. Because the Inquisitor thinks this, he in turn is a danger to the heretics. Not only does he menace their personal safety; also, if the heretics think that the spreading of their word will benefit all who embrace it, then they must see the Inquisitor as bringing disutility to all mankind. And the more there are of the orthodox, who think as the Inquisitor does, the worse it will be.

10. Milquetoast may resemble the sort of liberal portrayed in Thomas Nagel, 'Moral Conflict and Political Legitimacy', Philosophy & Public Affairs 16 (1987), pp.215-240: 'The defense of liberalism requires that a limit somehow be drawn to appeals to the truth in political argument' (p.227). True liberalism 'must depend on a distinction between what justifies individual belief and what justified appealing to that belief in support of the exercise of political power' (p.229). But of course Nagel's liberal is moved not by timidity but by high principle.
It would be best, indeed, if none were left who might someday reinfect mankind with the old darkness. Important matters are at stake. And now let us suppose that the heretics, no less than the Inquisitor, are wholehearted pursuers of utility as they see it. (Utility in the largest sense.) In this way the heretics think that the Inquisitor, and all of the orthodox, hold a dangerous opinion—dangerous enough to be well worth suppressing, despite all the considerations on the neutralist tally.

I suppose that some such rough symmetry is a common, though not a necessary, feature of situations in which someone thinks that someone else’s opinion is dangerous enough to be worth suppressing.

Devoted as both sides are to utility, and disagreeing as they do about where utility is to be found, what is there to do but fight it out? According to the Inquisitor’s opinion, the best outcome will be victory: to vanquish the heretics and suppress their heresy. If this outcome is within reach, going for it is required. Not only is toleration not required by any appeal to utility; it is forbidden. Any restraint or mercy would be wrong. It would be self-indulgent neglect of ‘the permanent interests of man as a progressive being’, since the foremost of these interests is salvation. Suppose further that there is no hope of changing the Inquisitor’s mind about the causes of salvation and damnation. Then there is no way—Millian or unMillian—to persuade him that it is a utilitarian mistake to suppress heresy. He has done his sums correctly, by his lights; we cannot fault them. Of course we can, and we should, fault his premises. They are both false and harmful. But there is no further mistake about what follows.

Likewise, mutatis mutandis, according to the heretics’ opinion.

If one side has victory within reach, the utilitarian defence of toleration fails. But now suppose instead that the two sides are more or less equally matched. Victory is not so clearly within reach. Neither side can have it just for the asking. Resort to war means taking a gamble. One side or the other will win, and then the winners will suppress the dangerous opinions of the losers. Orthodoxy will triumph and heresy will vanish, at least for a time. Or else heresy will triumph and orthodoxy will vanish. Who can tell which it will be?

In deciding what he thinks of a state of toleration, the Inquisitor must compare it not just with one possible outcome of war but with both. Toleration means that both creeds go unsuppressed, they flourish side by side, they compete for adherents. Many are lost, but many are saved. How many?—It depends. The fear is that the heretics will not scruple to advance their cause by cunning deceit; the hope is that truth will have an inherent advantage, and will benefit from God’s favour. Let us suppose that the Inquisitor takes a middling view of the prospect, not too pessimistic and not too optimistic. Then just as he finds victory vastly better than toleration, from the standpoint of salvation and therefore from the standpoint of utility, so he finds defeat vastly worse. According to the Inquisitor’s opinion, the triumph of heresy would be a catastrophic loss of utility. The considerations on the neutralist tally have negligible weight, given the enormous amount of utility at stake. Even the pleasures of peace and the horrors of war have negligible weight.
But the risk of defeat is far from negligible.

Likewise, *mutatis mutandis*, according to the heretics' opinion.

The Inquisitor's fear of defeat might outweigh his hope of victory. It might seem to him that suppression of orthodoxy would be more of a loss than suppression of heresy would be a gain (more lasting, perhaps); or he might take a pessimistic view of the gamble of war, and think it more likely than not that the heretics would win. Or he might take a moderately optimistic view of how many souls could be won under toleration. One way or another, he might have reason to prefer mutual toleration, unsatisfactory stalemate though it be, to war. His reason is a utilitarian reason. But it rests entirely on what he takes to be the weighty benefits and harms at stake—not the lightweight benefits and harms on the neutralist tally.

It might happen for the heretics likewise that the fear of defeat outweighs the hope of victory. If both sides think defeat more likely than victory, one side must be mistaken, but even a thoughtful utilitarian might well make such a mistake. If both sides think defeat would be more of a loss than victory would be a gain, there needn't be any mistake on either side—except, of course, the underlying mistake that one or both are making all along about what conduces to utility.

Or the heretics also might hope to do well at winning souls under toleration. The orthodox and the heretics can expect alike to win the most souls, if they believe alike that truth, or the creed God favours, will have the advantage. Their expectations are opposite, and one side or the other will be disappointed, but they can face competition with a common optimism.

It may happen, then, that each side prefers toleration to defeat more than it prefers victory to toleration, and therefore prefers toleration to the gamble of fighting it out. Then we have a utilitarian basis for a treaty of toleration. Conditional toleration—toleration so long as the other side also practices toleration—would be an equilibrium. It would be the best that either side could do, if it were what the other side was doing. Toleration is everyone's second choice. The first choice—to suppress and yet be tolerated, to gain victory without risking defeat—is not available; the other side will see to that. The third choice is the gamble of war, and we have supposed that both sides find the odds not good enough. War would be another equilibrium, but a worse one in the opinions of both sides. The worst choice is unconditional unilateral toleration, which means letting the other side have their way unopposed.

In such a case, with two equilibria and a preference on both sides for one over the other—toleration over war—it is neither automatic nor impossible that both sides will find their way to the equilibrium they both prefer. They might get there formally, by bilateral negotiation and

11. I shall be speaking almost as if there were a conflict of opposed aims. Strictly speaking, there is not. Both sides are, *ex hypothesi*, wholehearted in their pursuit of utility. But their fundamental disagreement about how to pursue their common aim is no different, strategically, from a fundamental conflict of aims. We may speak for short of a gain for one side, versus a gain for the other. But what that really means is a gain for utility according to the opinion of one side, versus a gain for utility according to the opinion of the other.
agreement. They might get there by unilateral initiatives and invitations to reciprocate. They might drift there, gradually developing a tacit understanding. They might get there under the influence of non-utilitarian reasons, and only afterward find that they had reached the outcome that maximised utility by the lights of both sides. They might have been there all along, in accordance with ancient custom. In each case, I will say that they have arrived at a treaty of toleration—maybe explicit and formal, maybe tacit.

Some treaties need to be sustained by trust and honour, lest a cheater gain advantage. It is hard to see how such a treaty could work between strict utilitarians; because if a utilitarian thinks it will maximise utility if he gains the upper hand, and if he thinks he can gain the upper hand by breaking his sworn word, then that is what he must do. But if there are no opportunities for secret preparation and a surprise breakout, then unutilitarian means of commitment are not required. The utility of the treaty is incentive enough to keep it. Neither side wants to withdraw toleration, lest the other side should have nothing to lose by withdrawing its reciprocal toleration. Often enough, contractarian and utilitarian defences of social institutions are put forward as rivals. Not so this time—here we have a contract for utilitarians.

The hopes and fears of the two sides may or may not be such as to permit a treaty of toleration. If they are, toleration may or may not be forthcoming—war is still an equilibrium, it takes two to make the switch. But now a utilitarian friend of toleration has a case to make. This time, it is a case meant not for the sceptical or the apathetic, not for the dismayed irreligious bystanders, not for Milquetoast, but for the Inquisitor himself. It is a thoroughly utilitarian case, but it is unMillian because it flouts the rule of neutralism. It plays both sides of the street. We say to the Inquisitor that a treaty of toleration affords his best hope for preventing the suppression of orthodoxy; we say to the heretics that it affords their best hope for preventing the suppression of heresy. Thereby we say to both that it affords the best hope for maximising utility, according to their very different lights. But there is no common list of benefits and costs. On the contrary, what we offer to each side as the greatest benefit of toleration is just what the other must see as its greater cost.13

12. Formal treaties of toleration, specifically between Catholic and Protestant powers, played a great part in the origins of the institutions of toleration we know today. But we can very well question whether those treaties were equilibria in the pursuit of utility in the largest sense, or whether they were just an escape from the horrors of war in the short term.

13. Unfortunately, a parallel case might be made out for a treaty that not only enjoins toleration between the orthodox and the heretics, but also bans proselytising. That might offer the orthodox their best hope for preventing the slow and peaceful extinction of orthodoxy, and likewise offer the heretics their best hope for preventing the slow and peaceful extinction of heresy. It would be bad for toleration, since each side would have to sustain the treaty by curbing its own zealots. But while this might be a third equilibrium, preferred both to war and to toleration with proselytising, it needn’t be. Only if neither side has much confidence in its powers of persuasion will it be an equilibrium at all, let alone a preferred one.
12. Closing the Gap

While a utilitarian defence of some sort of toleration has been accomplished, or so I claim, it seems not yet to be the right sort. This grudging truce between enemies, who would be at each other’s throats but for their fear of defeat, is a far cry from the institutions of toleration we know and love. Our simple story of the orthodox and the heretics differs in several ways from the real world of toleration.

**Cheerful toleration.** If we want to uphold a treaty of toleration, and doing our part means letting harmful error flourish, then we have to do it; but we don’t have to like it. Why should we? Whereas we are proud of our institutions of toleration, and pleased to see the spectrum of diverse opinions that flourish unsuppressed. Without the ones we take to be harmful errors, the diversity would be less and we would be less well pleased. Our feelings are mixed, of course. We do not wholeheartedly welcome the errors. But we do to a significant, and bizarre, degree.

**Thoughtless toleration.** In the story, the defence depends on the details of the strategic balance between the two sides. Whereas in the real world, we never stop to think how the fortunes of war might go before we take for granted that toleration is better.

**Tolerating the weak.** In particular, we tolerate the weak. If our Inquisitor had the chance to nip heresy in the bud, long before there was any chance that the heretics might have the strength to win and suppress orthodoxy, of course he would do it. Whereas we treasure the liberty of the weak, and proclaim that the minority of one means as much to us as any other minority.

**Tolerating the intolerant.** There is no sense in making a treaty with someone who declares that he will not abide by it. If we tolerate harmful error as a quid pro quo, so that others will reciprocate by tolerating beneficial truths, why continue after they announce that they will not reciprocate? Whereas we tolerate the intolerant, no less than the tolerant. We do it; and almost everyone who cares for toleration thinks we ought to do it. After Marcuse said that the time had come to withdraw tolerance, his books were no harder to buy than they were before.

**Tolerating the extra-dangerous.** In the story, the utilitarian defence may depend also on the exact balance of good and harm that we expect from the several opinions that will be protected by a treaty of toleration. The more danger heresy seems to pose, the less likely our Inquisitor is to conclude that a treaty with the heretics might be advantageous. Whereas we, for the most part, favour tolerating all dangerous opinions alike, without seeking exceptions for the very most dangerous.

One difference between our simple story and the real world, of course, is that in the real world we are not all utilitarians. We may be content to mind our own business, and insist that it is not our business to protect mankind against the harm done by dangerous opinions. Or we may be
devotees of the 'abstract rights' foresworn by Mill; then we may think that the rights of others constrain us not to serve utility by suppressing dangerous opinions, no matter how high the stakes. (Or they may constrain us to renounce only the harshest methods of suppression. But if only the harshest methods could succeed, we will not need any very weighty utilitarian reasons to dissuade us from trying the ineffective milder methods.)

These differences certainly work in favour of toleration—cheerful and thoughtless toleration, and toleration even of the weak, the intolerant, and the extra-dangerous. But let us not rely on them. Let us rather stay with the fiction of a population of wholehearted utilitarians, so that we may retain as much common ground with Mill as possible. Even so, I think we can close the gap between toleration as we find it in the simple story and toleration as we find it in the real world. We need not abandon the idea of a treaty of toleration. Instead, we must find the right way to extend the idea from our simple two-sided case to a complicated case, many-sided and always changing.

In the real world, there are many different factions. They differ in their opinions, they differ in their opinion about one another's opinions, and they differ in strength. As time goes by, factions wax and wane, and split and merge. The weak may suddenly gang up in a strong alliance, or an alliance may break up and leave the former allies weak. The people who comprise the factions change their minds. Circumstances also change. As the complicated situation changes, understanding of it will lag. Nobody will know very well who deplores whose opinions how much, and with how much strength to back up his deploring. In this complicated world, no less than in the simple case, some will find the opinions of others dangerous, and worthy of suppression; and some will think their own opinions beneficial, and will seek to protect them from suppression. Many would think it worthwhile to tolerate the most deplorable opinions, if they could thereby secure reciprocal toleration from others. They would welcome toleration by treaty. But how can they arrange it?

There might be a vast network of little treaties, each one repeating in miniature our story of the treaty between the orthodox and the heretics. Each faction would have protection from its treaty partners, and if it had chosen its partners well, that would give it the protection it needs. Each faction would extend toleration so far as its treaties require, and no farther. Two factions would enter into a treaty only when both thought it advantageous, given the strategic balance between them, their estimate of the fortunes of war, and their estimate of the potential for good or harm of the opinions that would be protected. The weak, who can offer no reciprocal toleration worth seeking, and the fanatically intolerant, who will not offer reciprocal toleration, would of course be left out of the network of treaties. Those whose opinions were thought to be extra-dangerous also would tend to be left out, other things being equal. A treaty would end when either side thought it no longer advantageous, or when either side thought (rightly or wrongly) that the other side was breaking it.
The trouble is plain to see. It would be enormously difficult for any faction to see to it that, at every moment in the changing course of events, it had exactly the treaties that would be advantageous. There would be abundant opportunities to be mistaken: to overestimate one threat and underestimate another; to be taken by surprise in a realignment of alliances; to see violation where there is compliance or compliance where there is violation; to think it open season on some weakling, unaware that your treaty partner regards that weakling as an ally. Too much care not to tolerate deplorable opinions without an adequate *quid pro quo* is unwise, if it makes the whole arrangement unworkable. Then the desired protection cannot be had.

There might instead be one big simple treaty, loose in its terms, prescribing indiscriminate toleration all around. Exceptions to a treaty of toleration—for the weak, for the intolerant, or for the extra-dangerous—seem at first to make sense. But they threaten to wreck the treaty. As new opinionated factions arise, and old ones wax and wane and merge and split, there will be occasion for endless doubt and haggling about what the exceptions do and don’t cover. If some suppression is a violation and some falls under the exceptions, then the first can be masked as the second and the second can be misperceived as the first; all the more so, if most of the cases that arise are unclear ones. Then who can know how well the treaty is really working? How confident can anyone be that his own toleration will be reciprocated in the cases that matter? It will be all too easy to doubt whether it makes good sense to remain in compliance.

Therefore, beware exceptions. Keep it simple, stupid—that which is not there cannot go wrong. First, some toleration of dangerous opinions is justified as a *quid pro quo*; then other toleration is justified because it makes the first transaction feasible.

Ought we to say, simply: *no exceptions*? It seems as if an exception that works even-handedly, and not to the permanent disadvantage of any opinion, ought to be safe. If we regulate only the manner of expression and not the content, why should anybody think that he has nothing to reciprocate because his own opinion is beyond toleration? Nobody has an opinion that he can express only by falsely shouting fire in a theatre, or only by defamation, or only by obscenity. Yet we know that even such exceptions as these can be abused. Some clever analogiser will try to erase the line between the innocent even-handed exception and the dangerous discriminatory one. He will claim that denouncing conscription is like shouting fire in a theatre, because both create a clear and present danger. Or he will claim that sharp criticism of the conduct of high officials is defamatory. Or he will claim that common smut is not half so obscene as the disgusting opinions of his opponent. If we put any limit to toleration, it is not enough to make sure that the line as drawn will not undermine the treaty. We also need some assurance that the line will stay in place where it was drawn, and not shift under pressure.

14. The second half is quoted from the instructions for a Seagull outboard motor.
No exceptions are altogether safe; maybe some are safe enough. That is a question only to be answered by experience, and experience seems to show that some exceptions—the few we have now—are safe enough. They have not yet undermined the treaty, despite all the efforts of mischievous analogisers, and there is no obvious reason why they should become more dangerous in future. We needn’t fear them much, and perhaps we can even welcome such benefits as they bring. But to try out some new and different exceptions would be foolhardy.

A simple, nearly exceptionless, well-established treaty of toleration could in time become not just a constraint of conduct, but a climate of thought. If, in the end, you will always decide that the balance of cost and benefit comes out in favour of complying with the treaty, why should you ever stop to think about the harm done by tolerating a dangerous error? Eventually you will be tolerant by habit, proudly, cheerfully, and without thought of the costs. You will proceed as if the neutralist tally were the whole story about the costs and benefits of suppression. You will bracket whatever you may think about the harm done by others’ opinions. You might still think, in some compartment of your mind, that certain opinions are false and harmful. If the treaty of toleration has become second nature, you might be hard put to explain why these opinions are not dangerous enough to be worth suppressing. But you will never think of the danger as any reason to suppress.

This habit of bracketing might be not just a consequence of a treaty but part of its very content. Not so if the treaty is a formal one, to be sure; that had better regulate action, not thought, so that it can be exact and verifiable enough to permit confident agreement. But insofar as the treaty is an informal understanding, uncodified, growing up gradually, it may prescribe not only tolerant conduct but also habits of thought conducive to toleration. In particular, it may prescribe bracketing. If your compartmentalised habits of thoughts are to some extent within your control—not indeed at every moment, but at those moments when you don’t bother to think things through as thoroughly as you might—then you may compartmentalise for a utilitarian reason. You may see, dimly, that when you bracket your fear of others’ dangerous opinions, you participate in a custom that serves utility by your lights because it protects opinions you deem beneficial, and that would not long persist if the bracketing that conduces to toleration were not mostly reciprocated.

If a treaty of toleration tends to turn us into Milquetoasts and Millians, that is not wholly a bad thing. It is too bad if we become compartmentalised in our thinking, repressing at some times what we believe at other times about the harm opinions can do. But if we forget the costs of toleration, that makes toleration more robust. And if toleration is beneficial on balance, the more robust the better.
13. Conclusion

What is toleration good for? A proper utilitarian answer need not omit the neutralist tally. After all, it does carry some weight in favour of toleration. But the principal part of the answer cannot be neutral. The main benefit of toleration is that it protects so-and-so particular opinions, true and beneficial, which would be in danger of suppression were it not for the institutions of toleration. When reciprocal toleration protects such-and-such other opinions, false and harmful, that is a cost to be regretted, and not to be denied. When a utilitarian favours toleration, of course, it is because he reckons that the benefits outweigh the costs.

If you think it would serve utility to 'withdraw tolerance' from such-and-such dangerous opinions, you'd better think through all the consequences. Your effort might be an ineffective gesture; in which case, whatever you might accomplish, you will not do away with the danger. Or it might be not so ineffective. To the extent that you succeed in withdrawing toleration from your enemy, to that extent you deprive him of his incentive to tolerate you. If toleration is withdrawn in all directions, are you sure the opinions that enhance utility will be better off? When we no longer renounce the argumentum ad baculum, are you sure it will be you that carries the biggest stick?15

Princeton University

Received July 1988
Revised September 1988

15. I thank audiences on several occasions for helpful discussions. Thanks are due especially to D. M. Armstrong, Geoffrey Brennan, Keith Campbell, Philip Kitcher, Martin Krygier, Stephanie Lewis, Michael Mahoney, Thomas Nagel, H. J. McCloskey, T. M. Scanlon, D. W. Skubik, and Kim Sterelny.
We are accustomed to punish criminal attempts much more severely if they succeed than if they fail. We are also accustomed to wonder why. It is hard to find any rationale for our leniency toward the unsuccessful. Leniency toward aborted attempts, or mere preparation, might be easier to understand. (And whether easy or hard, it is not my present topic.) But what sense can we make of leniency toward a completed attempt—one that puts a victim at risk of harm, and fails only by luck to do actual harm?

Dee takes a shot at his enemy, and so does Dum. They both want to kill; they both try, and we may suppose they try equally hard. Both act out of malice, without any shred of justification or excuse. Both give us reason to fear that they might be ready to kill in the future. The only difference is that Dee hits and Dum misses. So Dee has killed, he is guilty of murder, and we put him to death. Dum has not killed, he is guilty only of attempted murder, and he gets a short prison sentence.

Why? Dee and Dum were equally wicked in their desires. They were equally uninhibited in pursuing their wicked desires. Insofar as the wicked deserve to be punished, they deserve it equally. Their conduct was equally dangerous: they inflicted equal risks of death on their re-

This article arose out of discussion of a lecture by Judith Thomson about the guilt of successful and unsuccessful attempters. I am grateful for comments by John Broome, Stephanie Lewis, T. M. Scanlon, Thomas Schelling, and Jonathan Suzman; by the Editors of Philosophy & Public Affairs; and by audiences at Monash University, the Australian National University, and the Russellian Society (Sydney).

1. I do not wish to enter the debate about whether the traditional death penalty is ever justified. If you think not, substitute throughout whatever you think is the correct maximum penalty; my argument will go through almost without change.
spective victims. Insofar as those who act dangerously deserve to be pun-
ished, again they deserve it equally. Maybe Dee's act was worse than
Dum's act, just because of Dee's success; but it is not the act that suffers
punishment, it is the agent. Likewise, if we want to express our abhor-
rence of wickedness or of dangerous conduct, either exemplar of what
we abhor is fit to star in the drama of crime and punishment. Further,
Dee and Dum have equally engaged in conduct we want to prevent by
deterrence. For we prevent successful attempts by preventing attempts
generally. We cannot deter success separately from deterring attempts,
since attempters make no separate choice about whether to succeed.
Further, Dee and Dum have equally shown us that we might all be safer
if we defended ourselves against them; and one function of punishment
(at any rate if it is death, imprisonment, or transportation) is to get dan-
gerous criminals off the streets before they do more harm. So how does
their different luck in hitting or missing make any difference to consid-
erations of desert, expression, deterrence, or defense? How can it be just,
on any credible theory of just punishment, to punish them differently?

Here is one rationale for our peculiar practice. If the gods see innocent
blood shed, they will be angry; if they are angry, none of us will be safe
until they are propitiated; and to propitiate the gods, we must shed guilty
blood. Whereas if by luck no innocent blood is shed, the gods will not be
angered just by the sight of unsuccessful wickedness, so there will be no
need of propitiation.—This rationale would make sense, if its premises
were true. And if we put “the public” or “the victim's kin” for “the gods”
throughout it still makes sense; and that way, maybe the premises are
true, at least sometimes and to some extent. But this rationale does noth-
ing at all to defend our practice as just. If our practice is unjust, then
the ways of the gods (or the public, or the kin) are unjust, although if
the powers that be want to see injustice done, it might be prudent to
ignore justice and do their bidding.

A purely conservative rationale is open to the same complaint. Maybe
it is a good idea to stay with the practice we have learned how to operate,
lest a reform cause unexpected problems. Maybe it is good for people to
see the law go on working as they are accustomed to expect it to. Maybe
a reform would convey unintended and disruptive messages: as it might
be, that we have decided to take murder less seriously than we used to.
These considerations may be excellent reasons why it is prudent to leave
well enough alone, and condone whatever injustice there may be in our present practice. They do nothing at all to defend our practice as just.

Another rationale concerns the deterrence of second attempts. If at first you don’t succeed, and if success would bring no extra punishment, then you have nothing left to lose if you try, try again. “If exactly the same penalty is prescribed for successes as for attempts, there will be every reason to make sure that one is successful.” It cannot hurt to have some deterrence left after deterrence has failed. Maybe the experience of having tried once will make the criminal more deterrable than he was at first.—But why is this any reason for punishing successful attempts more severely? It might as well just be a reason for punishing two attempts more severely than one, which we could do regardless of success. If each separate attempt is punished, and if one share of punishment is not so bad that a second share would be no worse, then we have some deterrence against second attempts.

Another rationale sees punishment purely as a deterrent, and assumes that we will have deterrence enough if we make sure that crime never pays. If so, there is no justification for any more penal harm than it takes to offset the gains from a crime. Then a failed attempt needs no punishment: there are no gains to be offset, so even if unpunished it still doesn’t pay.—I reply that in the first place, this system of minimum deterrence seems likely to dissuade only the most calculating of criminals. In the second place, punishment is not just a deterrent. I myself might not insist on retribution per se, but certainly the expressive and defensive functions of punishment are not to be lightly forsaken.

Another rationale invokes the idea of “moral luck.” Strange to say, it can happen by luck alone that one person ends up more wicked than another. Perhaps that is why the successful attempter, by luck alone, ends up deserving more severe punishment?—I rely, first, that to some extent this suggestion merely names our problem. We ask how Dee can deserve more severe punishment just because his shot hits the mark.

Call that "moral luck" if you will; then we have been asking all along how this sort of moral luck is possible. But, second, it may be misleading to speak of the moral luck of the attempter, since it may tend to conflate this case with something quite different. The most intelligible cases of moral luck are those in which the lucky and the unlucky alike are disposed to become wicked if tempted, and only the unlucky are tempted. But then, however alike they may have been originally, the lucky and the unlucky do end up different in how they are and in how they act. Not so for the luck of hitting or missing. It makes no difference to how the lucky and the unlucky are, and no difference to how they act.4

Finally, another rationale invokes the difference between wholehearted and halfhearted attempts.5 Both are bad, but wholehearted attempts are worse. A wholehearted attempt involves more careful planning, more precautions against failure, more effort, more persistence, and perhaps repeated tries. Ceteris paribus, a wholehearted attempt evinces more wickedness—stronger wicked desires, or less inhibition about pursuing them. Ceteris paribus, a wholehearted attempt is more dangerous. It is more likely to succeed; it subjects the victim, knowingly and wrongfully, to a greater risk. Therefore it is more urgently in need of prevention by deterrence. Ceteris paribus, the perpetrator of a wholehearted attempt is more of a proven danger to us all, so it is more urgent to get him off the streets. So from every standpoint—desert, expression, deterrence, defense—it makes good sense to punish attempts more severely when they are wholehearted. Now, since wholehearted attempts

4. The luck of hitting and missing does make a difference to how their actions of shooting may be described: Dee's is a killing, Dum's is not. Dee's causes harm and thereby invades the victim's rights in a way that Dum's does not. (Dee invades the victim's right not to be harmed, as well as his right not to be endangered; Dum invades only the latter right.) But this is no difference in how they act, since the description of an action in terms of what it causes is an extrinsic description. The actions themselves, events that are finished when the agent has done his part, do not differ in any intrinsic way.

You might protest that a killing is not over when the killer has done his part; it is a more prolonged event that ends with the death of the victim; so there is, after all, an intrinsic difference between Dee's action of killing and Dum's action of shooting and missing.—No; an action of killing is different from the prolonged event of someone's getting killed, even though "the killing" can denote either one.

are more likely to succeed, success is some evidence that the attempt was wholehearted. Punishing success, then, is a rough and ready way of punishing wholeheartedness.

I grant that it is just to punish wholehearted attempts more severely—or better, since “heartedness” admits of degrees, to proportion the punishment to the heartedness of the attempt. And I grant that in so doing we may take the probability of success—in other words, the risk inflicted on the victim—as our measure of heartedness. That means not proportioning the punishment simply to the offender’s wickedness, because two equally wicked attempters may not be equally likely to succeed. One may be more dangerous than the other because he has the advantage in skill or resources or information or opportunity. Then if we proportion punishment to heartedness measured by risk, we may punish one attempter more severely not because he was more wicked, but because his conduct was more dangerous. From a purely retributive standpoint, wickedness might seem the more appropriate measure; but from the expressive standpoint, we may prefer to dramatize our abhorrence not of wickedness per se but of dangerous wickedness; and from the standpoint of deterrence or defense, clearly it is dangerous conduct that matters.

So far, so good; but I protest that it is unjust to punish success as a rough and ready way of punishing wholeheartedness. It’s just too rough and ready. Success is some evidence of wholeheartedness, sure enough. But it is very unreliable evidence: the wholehearted attempt may very well be thwarted, the half- or quarterhearted attempt may succeed. And we can have other evidence that bears as much or more on whether the attempt was wholehearted. If what we really want is to punish wholeheartedness, we have no business heeding only one unreliable fragment of the total evidence, and then treating that fragment as if it were conclusive. Suppose we had reason—good reason—to think that on average the old tend to be more wholehearted than the young in their criminal attempts. Suppose even that we could infer wholeheartedness from age somewhat more reliably than we can infer it from success. Then if we punished attempters more severely in proportion to their age, that would be another rough and ready way of punishing wholeheartedness. Ex hypothesi, it would be less rough and ready than what we do in punishing success. It would still fall far short of our standards of justice.
In what follows, I shall propose a new rationale. I do not say that it works. I do say that the new rationale works better than the old ones. It makes at least a prima facie case that our peculiar practice is just, and I do not see any decisive rebuttal. All the same, I think that the prima facie case is probably not good enough, and probably there is no adequate justification for punishing attempts more severely when they succeed.

Our present practice amounts to a disguised form of penal lottery—a punishment that leaves something to chance. Seen thus, it does in some sense punish all attempts alike, regardless of success. It is no less just, and no more just, than an undisguised penal lottery would be. Probably any penal lottery is seriously unjust, but it is none too easy to explain why.

By a penal lottery, I mean a system of punishment in which the convicted criminal is subjected to a risk of punitive harm. If he wins the lottery, he escapes the harm. If he loses, he does not. A pure penal lottery is one in which the winners suffer no harm at all; an impure penal lottery is one in which winners and losers alike suffer some harm, but the losers suffer more harm. It is a mixture: part of the punishment is certain harm, part is the penal lottery.

An overt penal lottery is one in which the punishment is announced explicitly as a risk—there might be ways of dramatizing the fact, such as a drawing of straws on the steps of the gallows. A covert penal lottery is one in which the punishment is not announced as a risk, but it is common knowledge that it brings risk with it. (A covert lottery must presumably be impure.)

A historical example of an overt penal lottery is the decimation of a regiment as punishment for mutiny. Each soldier is punished for his part in the mutiny by a one-in-ten risk of being put to death. It is a fairly pure penal lottery, but not entirely pure: the terror of waiting to see who must die is part of the punishment, and this part falls with certainty on all the mutineers alike.

Covert and impure penal lotteries are commonplace in our own time. If one drawback of prison is that it is a place where one is exposed to capricious violence, or to a serious risk of catching AIDS, then a prison

sentence is in part a penal lottery. If the gulag is noted for its abysmal standards of occupational health and safety, then a sentence of forced labor is in part a penal lottery.

III

What do we think, and what should we think, of penal lotteries? Specifically, what should we think of a penal lottery, with death for the losers, as the punishment for all attempts at murder, regardless of success? Successful or not, the essence of the crime is to subject the victim, knowingly and wrongfully, to a serious risk of death. The proposed punishment is to be subjected to a like risk of death.

We need a standard of comparison. Our present system of leniency toward the unsuccessful is too problematic to make a good standard, so let us instead compare the penal lottery with a hypothetical reformed system. How does the lottery compare with a system that punishes all attempts regardless of success, by the certain harm of a moderate prison term? A moderate term, because if we punished successful and unsuccessful attempts alike, we would presumably set the punishment somewhere between our present severe punishment of the one and our lenient punishment of the other. (Let the prison be a safe one, so that in the comparison case we have no trace of a penal lottery.) Both for the lottery and for the comparison case, I shall assume that we punish regardless of success. In the one case, success per se makes no difference to the odds; in the other case, no difference to the time in prison. This is not to say that every convicted criminal gets the very same sentence. Other factors might still make a difference. In particular, heartedness (measured by the risk inflicted) could make a difference, and success could make a difference to the extent that it is part of our evidence about heartedness.

Now, how do the two alternatives compare?

The penal lottery may have some practical advantages. It gets the case over and done with quickly. It is not a crime school. A prison costs a lot more than a gallows plus a supply of long and short straws.\(^7\)

(Likewise a prison with adequate protection against random brutality by guards and fellow inmates costs more than a prison without. So it

---

\(^7\) This point would disappear if something less cheap and quick than death were the penalty for losers of the lottery.
seems that we have already been attracted by the economy of a system that has at least some covert admixture of lottery.

Like a prison term (or fines, or flogging) and unlike the death penalty *simpliciter*, the penal lottery can be graduated as finely as we like. When we take the crime to be worse, we provide fewer long straws to go with the fatal short straws. In particular, that is how we can provide a more severe punishment for the more wholehearted attempt that subjected the victim to a greater risk.

From the standpoint of dramatizing our abhorrence of wicked and dangerous conduct, a penal lottery seems at least as good as a prison sentence. Making the punishment fit the crime, Mikado-fashion, is *poetic* justice. The point we want to dramatize, both to the criminal and to the public, is that what we think of the crime is just like what the criminal thinks of his punishment. If it's a risk for a risk, how can anybody miss the point?

From the standpoint of deterrence, there is no doubt that we are sometimes well deterred by the prospect of risk. It happens every time we wait to cross the street. It is an empirical question how effective a deterrent the penal lottery might be. Compared with the alternative punishment of a certain harm, such as a moderate prison term, the lottery might give us more deterrence for a given amount of penal harm, or it might give us less. Whether it gives us more or less might depend a lot on the details of how the two systems operate. If the lottery gave us more, that would make it preferable from the standpoint of deterrence.

(We often hear about evidence that certainty is more deterring than severity. But to the extent that this evidence pertains only to the uncertainty of getting caught, getting convicted, and serving the full sentence, it is scarcely relevant. The criminal might think of escaping punishment as a game of skill—his skill, or perhaps his lawyer's. For all we know, a risk of losing a game of chance might be much more deterring than an equal risk of losing a game of skill.)

From the standpoint of defense, the penal lottery gets some dangerous criminals off the streets forever, while others go free at once. Moderate

8. See Thomas C. Schelling, "The Threat That Leaves Something to Chance," in his book *The Strategy of Conflict* (Cambridge: Harvard University Press, 1960). Schelling does not discuss penal lotteries as such, but much of his discussion carries over. What does not carry over, or not much, is his discussion of chancy threats as a way to gain credibility when one has strong reason not to fulfill one's threat.
prison terms would let all go free after a longer time, some of them perhaps reformed and some of them hardened and embittered. It is another empirical question which alternative is the more effective system of defense. Again, the answer may depend greatly on the details of the two systems, and on much else that we cannot easily find out.9

IV

So far we have abundant uncertainties, but no clear-cut case against the penal lottery. If anything, the balance may be tipping in its favor. So let us turn finally to the standpoint of desert. Here it is a bit hard to know what to make of the penal lottery. If the court has done its job correctly, then all who are sentenced to face the same lottery, with the same odds, are equally guilty of equally grave crimes. They deserve equal treatment. Do they get it?—Yes and no.

Yes. We treat them alike because we subject them all to the very same penal lottery, with the very same odds. And when the lots are drawn, we treat them alike again, because we follow the same predetermined contingency plan—death for losers, freedom for winners—for all of them alike.

No. Some of them are put to death, some are set free, and what could be more unequal than that?

Yes. Their fates are unequal, of course. But that is not our doing. They are treated unequally by Fortune, not by us.

No. But it is we who hand them over to the inequity of Fortune. We are Fortune’s accomplices.

Yes. Everyone is exposed to the inequity of Fortune, in ever so many ways. However nice it may be to undo some of these inequities, we do not ordinarily think of this as part of what is required for equal treatment.

No. It’s one thing not to go out of our way to undo the inequities of Fortune; it’s another thing to go out of our way to increase them.

9. This question would have to be reconsidered if something other than death were the maximum penalty, and so the penalty for losers of the lottery. It would remain an empirical question, and probably a difficult one, which is the more effective system of defense.
Yes. We do that too, and think it not at all contrary to equal treatment. When we hire astronauts, or soldiers or sailors or firemen or police, we knowingly subject these people to more of the inequities of Fortune than are found in ordinary life.

No. But the astronauts are volunteers . . .

Yes. . . . and so are the criminals, when they commit the crimes for which they know they must face the lottery. The soldiers, however, sometimes are not.

No. Start over. We agreed that the winners and losers deserve equal punishment. That is because they are equally guilty. Then they deserve to suffer equally. But they do not.

Yes. They do not suffer equally; but if they deserve to, that is not our affair. We seldom think that equal punishment means making sure of equal suffering. Does the cheerful man get a longer prison sentence than the equally guilty morose man, to make sure of equal suffering? If one convict gets lung cancer in prison, do we see to it that the rest who are equally guilty suffer equally? When we punish equally, what we equalize is not the suffering itself. What we equalize is our contribution to expected suffering.

No. This all seems like grim sophistry. Surely, equal treatment has to mean more than just treating people so that some common description of what we are doing will apply to them all alike.

Yes. True. But we have made up our minds already, in other connections, that lotteries count as equal treatment, or near enough. When we have an indivisible benefit or burden to hand out (or even one that is divisible at a significant cost) we are very well content to resort to a lottery. We are satisfied that all who have equal chances are getting equal treatment—and not in some queer philosophers' sense, but in the sense that matters to justice.

It seems to me that "Yes" is winning this argument, but that truth and justice are still somehow on the side of "No." The next move, dear reader, is up to you. I shall leave it unsettled whether a penal lottery would be just. I shall move on to my second task, which is to show that our present practice amounts to a covert penal lottery. If the penal lottery is just, so is our present practice. If not, not.
To show that they do not matter, I shall introduce the differences between an overt penal lottery and our present practice one at a time, by running through a sequence of cases. I claim that at no step is there any significant difference of justice between one case and the next. Such differences as there are will be practical advantages only, and will come out much in favor of our present practice.

*Case 1* is the overt penal lottery as we have imagined it already, with one added stipulation, as follows. We will proportion the punishment to the heartedness of the attempt, as measured by the risk of death\(^{10}\) the criminal knowingly and wrongfully inflicted on the victim. We will do this by sentencing the criminal to a risk equal to the one he inflicted on the victim. If the criminal subjected his victim to an 80 percent risk of death, he shall draw his straw from a bundle of eight short and two long; whereas if he halfheartedly subjected the victim to a mere 40 percent risk, he shall draw from four short and six long; and in this way his punishment shall fit his crime. Therefore the court's task is not limited to ascertaining whether the defendant did knowingly and wrongfully subject the victim to a risk of death; also the court must ascertain how much of a risk it was.

*Case 2* is like *Case 1*, except that we skip the dramatic ceremony on the steps of the gallows and draw straws ahead of time. In fact, we have the drawing even before the trial. It is not the defendant himself who draws, but the Public Drawer. The Drawer is sworn to secrecy; he reveals the outcome only when and if the defendant has been found guilty and sentenced to the lottery. If the defendant is acquitted and the drawing turns out to have been idle, no harm done. Since it is not known ahead of time whether the sentence will be eight and two, four and six, or what, the Drawer must make not one but many drawings ahead of time. He reveals the one, if any, that turns out to be called for.

*Case 3* is like *Case 2*, except without the secrecy. The Drawer announces at once whether the defendant will win or lose in case he is

---

10. I note a complication once and for all, but I shall ignore it in what follows. The relevant risk is not really the victim's risk of death, but rather the risk of being killed—that is, of dying a death which is caused, perhaps probabilistically, and in the appropriate insensitive fashion, by the criminal's act. Likewise for the criminal's risk in the penal lottery. (On probabilistic and insensitive causation, see my *Philosophical Papers*, vol. II [New York: Oxford University Press, 1986], pp. 175–88.)
found guilty and sentenced. (Or rather, whether he will win or lose if he is sentenced to eight and two, whether he will win or lose if he is sentenced to four and six, and so on.) This means that the suspense in the courtroom is higher on some occasions than others. But that need not matter, provided that the court can stick conscientiously to the task of ascertaining whether the defendant did knowingly and wrongfully subject the victim to risk, and if so how much risk. It is by declaring that a criminal deserves the lottery that the court expresses society's abhorrence of the crime. So the court's task is still worth doing, even when it is a foregone conclusion that the defendant will win the lottery if sentenced (as might happen if he had won all the alternative draws). But the trial may seem idle, and the expression of abhorrence may fall flat, when it is known all along that, come what may, the defendant will never face the lottery and lose.

Case 4 is like Case 3, except that we make the penal lottery less pure. Losers of the penal lottery get death, as before; winners get a short prison sentence. Therefore it is certain that every criminal who is sentenced to the lottery will suffer at least some penal harm. Thus we make sure that the trial and the sentence will be taken seriously even when it is a foregone conclusion that the defendant, if sentenced, will win the lottery.

Case 1 also was significantly impure. If the draw is held at the last minute, on the very steps of the gallows, then every criminal who is sentenced to face the lottery must spend a period of time—days? weeks? years?—in fear and trembling, and imprisoned, waiting to learn whether he will win or lose. This period of terror is a certain harm that falls on winners and losers alike. Case 2 very nearly eliminates the impurity, since there is no reason why the Drawer should not reveal the outcome very soon after the criminal is sentenced. Case 3 eliminates it entirely. (In every case, a defendant must spend a period in fear as he waits to learn whether he will be convicted. But this harm cannot count as penal, because it falls equally on the guilty and the innocent, on those who will be convicted and those who will be acquitted.) Case 4 restores impurity, to whatever extent we see fit, but in a different form.

Case 5 is like Case 4, except that the straws are replaced by a different chance device for determining the outcome of the lottery. The Public Drawer conducts an exact reenactment of the crime. If the victim in the reenactment dies, then the criminal loses the lottery. If it is a good reen-
The Punishment
That Leaves
Something to Chance

...the risk to the original victim equals the risk to the new victim in the reenactment, which in turn equals the risk that the criminal will lose the lottery; and so, as desired, we punish a risk by an equal risk.

If the outcome of the lottery is to be settled before the trial, as in Cases 2, 3, and 4, then it will be necessary for the Drawer to conduct not just one but several reenactments. He will entertain all reasonable alternative hypotheses about exactly how the crime might have happened—exactly what the defendant might have done by way of knowingly and wrongfully inflicting risk on the victim. He will conduct one reenactment for each hypothesis. The court’s task changes. If the court finds the defendant guilty of knowingly and wrongfully inflicting a risk of death, it is no longer required also to measure the amount of risk. Nobody need ever figure out whether it was 80 percent, 40 percent, or what. Instead, the court is required to ascertain which hypothesis about exactly how the crime happened is correct. Thereby the court chooses which of all the hypothetical reenactments is the one that determines whether the criminal wins or loses his lottery. If the court finds that the criminal took careful aim, then the chosen reenactment will be one in which the criminal’s stand-in also took careful aim, whereas if the court finds that the criminal halfheartedly fired in the victim’s general direction, the chosen reenactment will be one in which the stand-in did likewise. So the criminal will be more likely to lose his lottery in the first case than in the second.

The drawbacks of a lottery by reenactment are plain to see. Soon we shall find the remedy. But first, let us look at the advantages of a lottery by reenactment over a lottery by drawing straws. We have already noted that with straws, the court had to measure how much risk the criminal inflicted, whereas with reenactments, the court has only to ascertain exactly how the crime happened. Both tasks look well-nigh impossible. But the second must be easier, because the first task consists of the second plus more besides. The only way for the court to measure the risk would be to ascertain just what happened, and then find out just how much risk results from such happenings.

Another advantage of reenactments over straws appears when we try to be more careful about what we mean by “amount of risk.” Is it (1) an “objective chance”? Or is it (2) a reasonable degree of belief for a hypothetical observer who knows the situation in as much minute detail as feasible instruments could permit? Or is it (3) a reasonable degree of...
belief for someone who knows just as much about the details of the situation as the criminal did? Or is it (4) the criminal’s actual degree of belief, however unreasonable that might have been? It would be nice not to have to decide. But if we want to match the criminal’s risk in a lottery by straws to the victim’s risk, then we must decide. Not so for a lottery by reenactment. If the reenactment is perfect, we automatically match the amount of risk in all four senses. Even if the reenactment is imperfect, at least we can assure ourselves of match in senses (3) and (4). It may or may not be feasible to get assured match in senses (1) and (2), depending on the details of what happened. (If it turns out that the criminal left a bomb hooked up to a quantum randomizer, it will be comparatively easy. If he committed his crime in a more commonplace way, it will be much harder.) But whenever it is hard to get assured match in senses (1) and (2), it will be harder still to measure the risk and get assured match in a lottery by straws. So however the crime happened, and whatever sense of match we want, we do at least as well by reenactment as by straws, and sometimes we do better.

Case 6 is like Case 5, except that enactment replaces reenactment. We use the original crime, so to speak, as its own perfect reenactment. If the criminal is sentenced to face the lottery, then if his victim dies, he loses his lottery and he dies too, whereas if the victim lives, the criminal wins, and he gets only the short prison sentence. It does not matter when the lottery takes place, provided only that it is not settled so soon that the criminal may know its outcome before he decides whether to commit his crime.

The advantages are many: we need no Drawer to do the work; we need not find volunteers to be the stand-in victims in all the hypothetical reenactments; the “reenactment” is automatically perfect, matching the risk in all four senses; we spare the court the difficult task of ascertaining exactly how the crime happened. If we want to give a risk for a risk, and if we want to match risks in any but a very approximate and uncertain fashion, the lottery by enactment is not only the easy way, it is the only remotely feasible way.

The drawback is confusion. When a criminal is sentenced to face the lottery by straws, nobody will think him more guilty or more wicked just because his straw is short. And when a criminal is sentenced to face the lottery by reenactment, nobody will think him more guilty just because
the stand-in victim dies. But if he is sentenced to the lottery by enactment, then one and the same event plays a double role: if his victim dies, that death is at once the main harm done by his crime and also the way of losing his lottery. If we are not careful, we are apt to misunderstand. We may think that the successful attempter suffers a worse fate because he is more guilty when he does a worse harm, rather than because he loses his lottery. But it is not so: his success is irrelevant to his guilt, his wickedness, his desert, and his sentence to face the lottery—exactly as the shortness of his straw would have been, had he been sentenced to the lottery by straws.

VI

I submit that our present practice is exactly Case 6: punishment for attempts regardless of success, a penal lottery by enactment, impurity to help us take the affair seriously even when the lottery is won, and the inevitable confusion. We may not understand our practice as a penal lottery—confused as we are, we have trouble understanding it at all—but, so understood, it does make a good deal of sense. It is another question whether it is really just. Most likely it isn't, but I don't understand why not.

11. If it were known that the victim's risk was fifty-fifty, or if we did not care about matching risks, we could just as well reverse the lottery by enactment: the criminal loses if the victim lives, wins if the victim dies. Certainly nobody will think the criminal is more guilty if the victim lives.
Roughly, values are what we are disposed to value. Less roughly, we have this schematic definition: Something of the appropriate category is a value if and only if we would be disposed, under ideal conditions, to value it. It raises five questions. (1) What is the favourable attitude of ‘valuing’? (2) What is the ‘appropriate category’ of things? (3) What conditions are ‘ideal’ for valuing? (4) Who are ‘we’? (5) What is the modal status of the equivalence?

By answering these questions, I shall advance a version of the dispositional theory of value. I begin by classifying the theory that is going to emerge. First, it is naturalistic: it advances an analytic definition of value. It is naturalistic in another sense too: it fits into a naturalistic metaphysics. It invokes only such entities and distinctions as we need to believe in anyway, and needs nothing extra before it can deliver the values. It reduces facts about value to facts about our psychology.

The theory is subjective: it analyses value in terms of our attitudes. But it is not subjective in the narrower sense of implying that value is a topic on which whatever we may think is automatically true, or on which there is no truth at all. Nor does it imply that if we had been differently disposed, different things would have been values. Not quite—but it comes too close for comfort.

The theory is internalist: it makes a conceptual connection between value and motivation. But it offers no guarantee that everyone must be motivated to pursue whatever is of value; still less, whatever he judges to be of value. The connection is defeasible, in more ways than one.

The theory is cognitive: it allows us to seek and to gain knowledge about what is valuable. This knowledge is a posteriori knowledge of contingent matters of fact. It could in principle be gained by psychological experimentation. But it is more likely to be gained by difficult exercises of imagination, carried out perhaps in a philosopher’s or a novelist’s armchair.
The theory is conditionally relativist: it does not exclude the possibility that there may be no such thing as value *simply*, just value for this or that population. But it does not imply relativity, not even when taken together with what we know about the diversity of what people actually value. It leaves the question open.

Is it a form of realism about value?—That question is hard. I leave it for the end.

What is ‘valuing’? It is some sort of mental state, directed toward that which is valued. It might be a feeling, or a belief, or a desire. (Or a combination of these; or something that is two or three of them at once; or some fourth thing. But let us set these hypotheses aside, and hope to get by with something simpler.1)

A feeling?—Evidently not, because the feelings we have when we value things are too diverse.

A belief? What belief? You might say that one values something just by believing it to be a value. That is circular. We might hide the circularity by maneuvering between near-synonyms, but it is better to face it at once. If so, we have that being a value is some property such that something has it iff we are disposed, under ideal conditions, to believe that the thing has it. In other words, such that we are disposed, under ideal

---

1 The most interesting of the hypotheses here set aside is that an attitude of valuing might be a ‘desire’: a special kind of attitude that is both a belief and a desire and that motivates us, without benefit of other desires, in just the way that ordinary desires do. (Or it might be an attitude that is not identical with, but rather is necessarily connected with, a belief and a desire; or an attitude that is not strictly speaking either a belief or a desire, but is just like each apart from also being like the other.) Valuing X might be the desire that is at once a belief that X is good and a desire for X; where *goodness* just means that property, whatever it may be, such that a belief that X has it may double as a desire for X.

But we should hesitate to believe in desires, because integrating them into the folk psychology of belief and desire turns out to be no easy thing. On the difficulty with instrumental desires, see my ‘Desire as Belief’ and John Collins, ‘Belief, Desire and Revision’, *Mind* 97 (1988), pp. 323-342: when a system of attitudes changes under the impact of new information, beliefs evolve in one way and (instrumental) desires in another. A desire, trying to go both ways at once, would be torn apart. Intrinsic desires—a better candidate for the attitude of valuing—face a different difficulty. At least in miniature examples, they turn out to be altogether impervious to change under the impact of experience. Not bad, you might think—why should experience change our mind about what’s intrinsically good? The trouble is that the result applies not only to perceptual experience but also to experience of moral reflection, ‘intuiting’, and the like.
conditions, to be right about whether something has it. That is not empty; but it tells us little, since doubtless there are many properties about which we are disposed to be right.

Further, if valuing something just meant having a certain belief about it, then it seems that there would be no conceptual reason why valuing is a favourable attitude. We might not have favoured the things we value. We might have opposed them, or been entirely indifferent.

So we turn to desires. But we'd better not say that valuing something is just the same as desiring it. That may do for some of us: those who manage, by strength of will or by good luck, to desire exactly as they desire to desire. But not all of us are so fortunate. The thoughtful addict may desire his euphoric daze, but not value it. Even apart from all the costs and risks, he may hate himself for desiring something he values not at all. It is a desire he wants very much to be rid of. He desires his high, but he does not desire to desire it, and in fact he desires not to desire it. He does not desire an unaltered, mundane state of consciousness, but he does desire to desire it. We conclude that he does not value what he desires, but rather he values what he desires to desire.

Can we do better by climbing the ladder to desires of ever-higher order? What someone desires to desire to desire might conceivably differ from what he does desire to desire. Or... Should we perhaps say that what a person really values is given by his highest order of desire, whatever order that is? It is hard to tell whether this would really be better, because it is hard to imagine proper test cases. Further, if we go for the highest

---

2 Often in decision theory and economics, 'value' does just mean a measure of desiredness, and all desires count equally. But it's not the sense we want here.


4 It is comparatively easy to imagine instrumental third-order desires. Maybe our addict wishes he could like himself better than he does; and not by doing away with his addiction, which he takes to be impossible, but by becoming reconciled to it and accepting himself as he is. Or maybe he just fears that his second-order desire not to be addicted will someday lead him to suffer the pains of withdrawal. Either way, he wants to be rid of his second-order desire not to be addicted, but he wants it not for itself but as a means to some end. This is irrelevant: presumably it is intrinsic, not instrumental, desiring that is relevant to what someone values.
order, we automatically rule out the case of someone who desires
to \textit{value} differently than he does, yet this case is not obviously
impossible. I hesitantly conclude we do better to stop on the
second rung: valuing is just desiring to desire.

Recall G. E. Moore: ‘To take, for instance, one of the more
plausible, because one of the more complicated, of such
proposed definitions, it may easily be thought, at first sight, that
to be good may mean to be that which we desire to desire’.\(^5\) Of
course he does not endorse the definition, but at least he does it
the honour of choosing it for his target to display the open
question argument. I don’t say that everything we value is good;
but I do echo Moore to this extent. I say that to be \textit{valued} by us
means to be that which we desire to desire. Then to be a
value—to be good, near enough—means to be that which we
are disposed, under ideal conditions, to desire to desire. Still
more complicated, still more plausible. It allows, as it should,
that under less-than-ideal conditions we may wrongly value
what is not really good. As for Moore’s open question, we shall
face that later.

We have this much of an ‘internalist’ conceptual connection
between value and motivation. If something is a value, and if
someone is one of the appropriate ‘we’, and if he is in ideal
conditions, then it follows that he will value it. And if he values
it, and if he desires as he desires to desire, then he will desire it.
And if he desires it, and if this desire is not outweighed by other
conflicting desires, and if he has the instrumental rationality to
do what serves his desires according to his beliefs, then he will
pursue it. And if the relevant beliefs are near enough true, then
he will pursue it as effectively as possible. A conceptual
connection between value and motivation, sure enough—but a
multifariously iffy connection. Nothing less iffy would be
credible. But still less is it credible that there is no connection at
all.

In general, to find out whether something is disposed to give
response \(R\) under conditions \(C\), you can put it in \(C\) and find out
whether you get \(R\). That is a canonical way to learn whether
the disposition is present, though surely not the only possible

way. If a dispositional theory of value is true, then we have a canonical way to find out whether something is a value. To find out whether we would be disposed, under ideal conditions, to value it, put yourself in ideal conditions, if you can, making sure you can tell when you have succeeded in doing so. Then find out whether you value the thing in question, i.e. whether you desire to desire it. If you do, that confirms that it is a value. (I assume you are one of the appropriate 'we' and you know it.) Now we have this much of an 'internalist' conceptual connection between value judgements and motivation. It is even iffier than the connection between value itself and motivation; and again I say that if it were less iffy, it would be less credible. If someone believes that something is a value, and if he has come to this belief by the canonical method, and if he has remained in ideal conditions afterward or else retained the desire to desire that he had when in ideal conditions, then it follows that he values that thing. And if he desires as he desires to desire, then he desires that thing; and so on as before.

The connection is not with the judgement of value per se, but with the canonical way of coming to it. If someone reached the same judgement in some non-canonical way—as he might—that would imply nothing about his valuing or desiring or pursuing.

*What is the 'appropriate category'?* If values are what we are disposed to desire to desire, then the things that can be values must be among the things that can be desired. Those fall into two classes.

Sometimes, what one desires is that the world should be a

---

6 It is a fallible way; for it may be that you cannot put the thing in C without making the disposition disappear. Imagine that a surface now has just the molecular structure that disposes things to reflect light; but that exposing it to light would catalyze a swift chemical change and turn it into something unreflective. So long as it's kept in the dark, is it reflective?—I think so; but its reflectivity is what Ian Hunt once called a 'finkish' disposition, one that would vanish if put to the test. (So a simple counterfactual analysis of dispositions fails.) Could a disposition to value, or to disvalue, be finkish? Yes; here is an example due to Michael Tooley. Suppose, as I shall claim, that 'ideal conditions' include imaginative acquaintance; suppose there is no way to imagine direct electrical stimulation of the pleasure centre of the brain except by trying it out; and suppose that one brief trial would enslave you to the electrode and erase all other desires. Then I think you might well have a finkish disposition to disvalue the experience. If, per impossibile, you could manage to imagine it without at the same time having your present system of desires erased by the current, you would desire not to desire it.
certain way: that it should realise one of a certain class of
(maximally specific, qualitatively delineated) possibilities for the
whole world. This class—a ‘proposition’, in one sense of that
word—gives the content of the desire. To desire that the world
realise some possibility within the class is to desire that the
proposition be true. Call this ‘desire de dicto’.

But sometimes, what one desires concerns not just the world
but oneself: one simply desires to be a certain way. For instance,
Fred might want to be healthy, or wealthy, or wise. Then what
he wants is that he himself should realise one of a certain class of
(maximally specific, qualitatively delineated) possibilities for
an individual—or better, for an individual-in-a-world-at-a-
time. This class—a ‘property’ in one sense of that word, or an
‘egocentric proposition’—gives the content of the desire. To
desire to realise some possibility in the class is to desire to have
the property, or to desire that the egocentric proposition be true
of one. Call this ‘desire de se’, or ‘egocentric’ or ‘essentially
indexical’ desire.7

You might think to reduce desire de se to desire de dicto, saying
that if Arthur desires to be happy, what he desires is that the
world be such that Arthur is happy. (You might doubt that such
worlds comprise a qualitatively delineated class, so you might
consider dropping that requirement.) But no. That is not
exactly the same thing, though the difference shows up only
when we imagine someone who is wrong or unsure about who in
the world he is. Suppose Arthur thinks he is Martha. If Arthur is
self-centred he may desire to be happy, desire that the world be
one wherein Martha is happy, but not desire that the world is
one wherein Arthur is happy. If instead Arthur is selflessly
benevolent he may not desire to be happy, yet he may desire that
the world be such that Arthur is happy. If Arthur is so befuddled
as not to know whether he is Arthur or Martha, but hopes he is
Arthur, he does not just desire that the world be such that

7See Peter Geach, ‘On Beliefs about Oneself’, Analysis 18 (1957), pp. 23-24;
Hector-Neri Castañeda ‘On the Logic of Attributions of Self-Knowledge to Others’,
Philosophical Review 86 (1977), pp. 474-497; and ‘The Problem of the Essential
Indexical’, Noûs 14 (1979), pp. 3-21; my ‘Attitudes De Dicto and De Se’, Philosophical Review
88 (1979), pp. 513-543; Roderick Chisholm, The First Person: An Essay on Reference and
Arthur is self-identical! In all these cases, Arthur’s desire is, at least in part, irreducibly de se.\(^8\)\(^9\)

When we acknowledge desires de se, we must distinguish two senses of ‘desiring the same thing’. If Jack Sprat and his wife both prefer fat meat, they desire alike. They are psychological duplicates, on this matter at least. But they do not agree in their desires, because no possible arrangement could satisfy them both. Whereas if Jack prefers the fat and his wife prefers the lean, then they differ psychologically, they do not desire alike. But they do agree, because if he eats no fat and she eats no lean, that would satisfy them both. In general, they desire alike iff they desire de se to have exactly the same properties and they desire de dicto that exactly the same propositions hold. They agree in desires iff exactly the same world would satisfy the desires of both; and a world that satisfies someone’s desires is one wherein he has all the properties that he desires de se and wherein all the propositions hold that he desires de dicto. Agreement in desire makes for harmony; desiring alike may well make for strife.

As we can desire de dicto or de se, so we can desire to desire de dicto or de se. If desiring to desire is valuing, and if values are what we are disposed to value, then we must distinguish values de dicto and de se. A value de dicto is a proposition such that we are disposed to desire to desire de dicto that it hold. A value de se is a property such that we are disposed to desire to desire de se to have it.

It is essential to distinguish. Consider egoism: roughly, the thesis that one’s own happiness is the only value. Egoism is meant to be general. It is not the thesis that the happiness of a certain special person, say Thrasymachus, is the only value. Egoism de dicto says that for each person X, the proposition that

---

\(^8\) What we can do is to go the other way, subsuming desire de dicto under desire de se. To desire that the world be a certain way is to desire that one have the property of living in a world that is that way—a property that belongs to all or none of the inhabitants of the world, depending on the way the world is. This subsumption, artificial though it be, is legitimate given a suitably broad notion of property. But for present purposes we need distinction, not unification. So let us henceforth ignore those desires de se that are equivalent to desires de dicto, and reserve the term ‘de se’ for those that are not.

\(^9\) If you like, you can put the egocentricity not in the content of desire itself but in an egocentric mode of presentation of that content. The choice matters little, save to simplicity. See Jeremy Butterfield, ‘Content and Context’ in Butterfield, ed., Language Mind and Logic (Cambridge University Press, 1986).
X is happy is the only value. That is inconsistent, as Moore observed. It says that there are as many different values as there are people, and each of them is the only value. Egoism de se says that the property of happiness—in other words, the egocentric proposition that one is happy—is the only value. Moore did not confute that. He ignored it. False and ugly though it be, egoism de se is at least a consistent doctrine. What it alleges to be the only value would indeed be just one value de se, not a multitude of values de dicto.

Insofar as values are de se, the wholehearted pursuit by everyone of the same genuine value will not necessarily result in harmony. All might value alike, valuing de se the same properties and valuing de dicto the same propositions. Insofar as they succeed in desiring as they desire to desire, they will desire alike. But that does not ensure that they will agree in desire. If egoism de se were true, and if happiness could best be pursued by doing others down and winning extra shares, then the pursuit by all of the very same single value would be the war of all against all.

Because egoism is false and ugly, we might be glad of a theoretical framework that allowed us to confute it a priori. And some of us might welcome a framework that promises us harmony, if only we can all manage to pursue the same genuine values. Was it right, then, to make a place for values de se? Should we have stipulated, instead, that something we are disposed to desire to desire shall count as a value only when it is a proposition that we are disposed to desire to desire de dicto?

No. Probably it is already wrong to reject egoism a priori but, be that as it may, there are other doctrines of value de se, more plausible and more attractive. Self-improvement and self-sacrifice are no less egocentric than self-aggrandizement and

---

10 Principia Ethica, Section 59.

11 Someone who said that happiness was the only value might mean something else, which is not a form of egoism at all. He might mean that the proposition that happiness is maximized is the only value—a single value de dicto. Or he might mean that for each person X, the proposition that X is happy is a value de dicto, and that these many values of parallel form are the only values. Mean what you please—I take these to be legitimate, but derivative, senses in which a property may be called a value. I only say they should not be confused with, or drive out, the sense in which a property may be a value de se.
self-indulgence. Surely we should make a place for putative values de se of altruism, of honour, and of loyalty to family, friends, and country. We may entertain the substantive thesis that none of these putative values de se is genuine, and that all genuine values are de dicto. But even if we believed this—myself, I think it wildly unlikely—we should not beg the question in its favour by building it into our theoretical framework.

What conditions are ‘ideal’? If someone has little notion what it would be like to live as a free spirit unbound by law, custom, loyalty, or love; or what a world of complete harmony and constant agreement would be like; then whether or not he blindly values these things must have little to do with whether or not they are truly values. What he lacks is imaginative acquaintance. If only he would think harder, and imagine vividly and thoroughly how it would be if these putative values were realised (and perhaps also how it would be if they were not) that would make his valuing a more reliable indicator of genuine value. And if he could gain the fullest imaginative acquaintance that is humanly possible, then, I suggest, his valuing would be an infallible indicator. Something is a value iff we are disposed, under conditions of the fullest possible imaginative acquaintance, to value it.

Compare a version of Intuitionism: by hard thought, one becomes imaginatively well acquainted with X; in consequence, but not as the conclusion of any sort of inference, one intuits that X has a certain unanalysable, non-natural property; and in consequence of that, one comes to value X. My story begins and ends the same. Only the middle is missing. Again, an exercise of imaginative reason plays a crucial role. Again, its relation to what follows is causal, and in no way inferential. But in my story, the consequent valuing is caused more directly, not via the detection of a peculiar property of X.

Can we say that the valuing ensued because X was a value?—Maybe so, but if we do, we are not saying much: it

---


13 Without in the process having his dispositions to value altered—see Footnote 6.
ensues because there is something about imaginative acquaintance with X that causes valuing.\textsuperscript{14}

The canonical way to find out whether something is a value requires a difficult imaginative exercise. And if you are to be sure of your answer, you need to be sure that you have gained the fullest imaginative acquaintance that is humanly possible. A tall order! You had better settle for less. Approximate the canonical test. Try hard to imagine how it would be if the putative value were (or were not) realised. Hope that your acquaintance comes close enough to the fullest possible that getting closer would not change your response. Then you may take your valuing as fallible evidence that you were acquainted with a genuine value, or your indifference as fallible evidence that you were not. You cannot be perfectly certain of your answer, but you can take it as sure enough to be going on with, subject to reconsideration in the light of new evidence. How sure is that?—Well, as always when we acknowledge fallibility, some of us will be bolder than others.

New evidence might be a more adequate imaginative exercise of your own. It might be the testimony of others. It might in principle be a result of scientific psychology—though it is far from likely that any such results will come to hand soon!

A trajectory toward fuller imaginative acquaintance with putative value X is not just a sequence of changes in your imaginative state. It has a direction to it. And that is so independently of my claim that it leads, after a point, to ever-surer knowledge about whether X is a value. For in learning how to imagine X, you gain abilities; later you have all the relevant imaginative abilities you had before, and more besides. And you notice, \textit{a priori}, relationships of coherence or incoherence between attitudes that might figure in the realisation of X; later you are aware of all that you had noticed before, and more

\textsuperscript{14}How does imaginative acquaintance cause valuing, when it does? How does imagination render values attractive? Does it happen the same way for all values?—For our purposes, it is enough to say that it happens. We needn't know how. But we may guess. Maybe imaginative acquaintance shows us how new desires would be seamless extensions of desires we have already. Or maybe we gravitate toward what we understand, lest we baffle ourselves—see J. David Velleman, \textit{Practical Reflection} (Princeton University Press, forthcoming). But that cannot be the whole story, because some easily understood lives—say a life of lethargy, ruled by a principle of least action—remain repellent.
besides. And you think of new questions to explore in your imagining—what might the life of the free spirit become, long years after its novelty had worn off?—and later you have in mind all the questions you had thought of before, and more besides. Forgetting is possible, of course. But by and large, the process resists reversal.¹⁵

Our theory makes a place for truth, and in principle for certain knowledge, and in practice for less-than-certain knowledge, about value. But also it makes a place for ignorance and error, for hesitant opinion and modesty, for trying to learn more and hoping to succeed. That is all to the good. One fault of some subjective and prescriptive theories is that they leave no room for modesty: just decide where you stand, then you may judge of value with the utmost confidence!

There is a long history of theories that analyse value in terms of hypothetical response under ideal conditions, with various suggestions about what conditions are ideal. Imaginative acquaintance often gets a mention. But much else does too. I think imaginative acquaintance is all we need—the rest should be in part subsumed, in part rejected.

First, the responder is often called an ideal spectator. That is tantamount to saying that conditions are ideal only when he is observing a sample of the putative value in question (or of its absence). If the putative value is de se, a property, then a sample can just be an instance. If it is de dicto, a proposition, it is hard to say in general what an observable sample could be. But if it is the proposition that a certain property is instantiated sometimes, or often, or as often as possible, or in all cases of a certain kind, then again a sample can just be an instance of the property. Anyone happy may serve as a sample of the proposition that total happiness is maximised.

Observable samples can sometimes prompt the imagination and thereby help us to advance imaginative acquaintance. But they are of limited use. For one thing, observation does not include mind-reading. Also, it does best with short, dramatic episodes. A lifelong pattern of stagnation, exemplifying the absence of various values, goes on too long to be easily

observed. Samples are dispensable as aids to imagination, and sometimes they are comparatively ineffective. A novel might be better.

The notion of an ideal spectator is part of a longstanding attempt to make dispositional theories of value and of colour run in parallel. But the analogy is none too good, and I doubt that it improves our understanding either of colour or of value. Drop it, and I think we have no further reason to say that a disposition to value is a disposition to respond to observed samples.¹⁶

Second, the ideal responder is often supposed to be well informed. If any item of empirical knowledge would affect his response, he knows it.—But some sorts of knowledge would not help to make your valuing a more reliable indicator of genuine value. Instead they would distract. If you knew too well how costly or how difficult it was to pursue some value, you might reject the grapes as sour, even when imaginative acquaintance with the value itself would have caused you to value it. Genuine values might be unattainable, or unattainable without undue sacrifice of other values. An ideal balancer of values needs thorough knowledge of the terms of trade. An ideal valuer may be better off without it. Our present business is not with the balancing, but with the prior question of what values there are to balance."²

Another unhelpful sort of knowledge is a vivid awareness that we are small and the cosmos is large; or a vivid awareness of the mortality of mankind, and of the cosmos itself. If such knowledge tends to extinguish all desire, and therefore all

¹⁶ If we had demanded samples, we would have had a choice about where to locate the disposition. Is it within us or without? Is it a disposition in the samples to evoke a response from spectators?—that is what best fits the supposed parallel with a dispositional theory of colour. See Robert Parfit and John Campbell, ‘Goodness and Fragility’, American Philosophical Quarterly 23 (1986), pp. 155-166, for an analysis of this kind. Or is it a disposition in the spectators to respond to samples? Or is it a disposition of the sample-cum-spectator system to respond to having its parts brought together? For us there is no choice. The propositions and properties that are the values cannot harbour any causal bases for dispositions. Samples could, but there needn’t be any samples. Imaginative experiences could, but those are within us, and are not themselves samples of values. So the disposition must reside in us, the responders. Being a value comes out as a dispositionaly analysed property, but not as a disposition of the things that have it. Values themselves are not disposed to do anything.

¹⁷ Previous theories of hypothetical response may indeed have been concerned as much with the analysis of right balancing as with value itself. If so, they cannot be faulted for trying to characterise an ideal balancer. However my present analysandum is different.
DISPOSITIONAL THEORIES OF VALUE

valuing, it will not help us to value just what is valuable. Likewise it will be unhelpful to dwell too much on the lowly causal origins of things. If some feature of our lives originated by kin selection, or Pavlovian conditioning, or sublimation of infantile sexuality, that is irrelevant to what it is like in itself. Unless he can overcome the illusion of relevance, a valuer will be more reliable if he remains ignorant of such matters.

However, I grant one case—a common one—in which one does need empirical knowledge in order to gain imaginative acquaintance with a given putative value. It may be 'given' in a way that underspecifies it, with the rest of the specification left to be filled in by reference to the actual ways of the world. For instance when I mentioned the life of a free spirit as a putative value, what I meant—and what you surely took me to mean—was the life of a free spirit in a world like ours. In such cases, a valuer must complete the specification by drawing on his knowledge of the world, else he will not know what he is supposed to imagine. To that extent—and only to that extent, I think—being well-informed is indeed a qualification for his job.18

Third, it may be said that the ideal responder should not only imagine having (or lacking) a putative value, but also imagine the effect on other people of someone's having (or lacking) it. Thinking what it would be like to live as a free spirit is not enough. You must also think what it would be like to encounter the free spirit and be ill-used. But again, I think the requirement is misplaced. It is appropriate not to an ideal valuer, but to an ideal balancer who must think through the cost to some values of the realisation of others. In addressing the prior question of what values there are, counting the cost is a distraction to be resisted.

Often, however, realising a putative value *de se* would itself involve imagining the impact of one's conduct on other people. When that is so, imagining realising the value involves

---

18 Imaginative acquaintance is sometimes thought to consist in the possession of a special kind of 'phenomenal' information. If that is so, of course my own candidate for 'ideal conditions' comes down to a special case of being well-informed. But it is not so—not even in the most favourable case, that of imaginative acquaintance with a kind of sense-experience. See my 'What Experience Teaches' in William Lycan, ed., *Mind and Cognition: A Reader* (Blackwell, 1989).
imagining the impact; and that cannot be done without simply imagining the impact. In such cases, imagining the impact does fit in; for it is already subsumed as part of imaginative acquaintance with the value itself.

Fourth, the ideal responder is often said to be dispassionate and impartial, like a good judge. Once more, the requirement is appropriate not to an ideal valuer but to an ideal balancer. The valuer is not a judge. He is more like an advocate under the adversarial system. He is a specialist, passionate and partial perhaps, in some one of all the values there are. On the present theory, when I say that X is a value iff we are disposed to value X under ideal conditions, I do not mean conditions that are ideal simpliciter, but rather conditions that are ideal for X. We should not assume that there is any such thing as a condition of imaginative acquaintance with all values at once. (Still less, all putative values.) Imagination involves simulation—getting into the skin of the part. How many skins can you get into all at once? Tranquillity and vigorous activity might both be values; but a full imaginative acquaintance with one might preclude a full imaginative acquaintance with the other. (The incompatibility might even be conceptual, not just psychological.) Then if we value both, as surely many of us do, it is not because of acquaintance with both at once. It might be a lasting effect of past imaginative acquaintance at some times with one and at other times with the other.

A further speculation: it might happen that there were values that could not even be valued all at once. If so, then conflict of values would go deeper than is ever seen in hard choices; because what makes a choice hard is that conflicting values are valued together by the unfortunate chooser. An alarming prospect! — or exhilarating, to those of us who delight in the rich variety of life.

Who are 'we'? An absolute version of the dispositional theory says that the 'we' refers to all mankind. To call something a value is to call it a value simpliciter, which means that everyone, always and everywhere, is disposed under ideal conditions to value it. Then there are values only insofar as all mankind are alike in their dispositions.

Maybe all mankind are alike. The manifest diversity of
valuing between different cultures—or for that matter within a culture, say between colleagues in the same philosophy department—is no counterevidence. In the first place, people may not be valuing as they would be disposed to value under ideal conditions. In the second place, remember that conditions of imaginative acquaintance are ideal for particular values, not *simpliciter*. So even if all are disposed alike, and all value as they would under ideal conditions, that may mean that some people value X as they would under conditions ideal for X, while others, who are no differently disposed, value Y as they would under conditions ideal for Y. If no conditions are ideal at once for X and for Y (still more if X and Y cannot both be valued at once), there could be diversity of valuing even in a population of psychological clones, if different ones had been led into different imaginative exercises.

We saw that it would be no easy job to find out for sure whether a particular person would be disposed to value something under ideal conditions of imaginative acquaintance with it. It would be harder still to find out all about one person’s dispositions. And not just because one hard job would have to be done many times over. It might happen that imaginative acquaintance with X would leave traces, in one’s valuing or otherwise, that got in the way of afterward imagining Y. To the extent that there was such interference, each new imaginative experiment would be harder than the ones before.

The fallback, if we are wary of presupposing that all mankind are alike in their dispositions to value, is tacit relativity. A *relative* version says that the ‘we’ in the analysis is indexical, and refers to a population consisting of the speaker and those somehow like him. If the analysis is indexical, so is the analysandum. Then for speaker S to call something a value is to call it a value for the population of S and those like him; which means that S and those like him are all disposed, under ideal conditions, to value it.

The relative version is not just one version, but a spectrum. What analysis you get depends on how stringent a standard of similarity you apply to the phrase ‘the speaker and those somehow like him’. At one end of the spectrum stands the absolute version: common humanity is likeness enough, so whoever speaks, all mankind are ‘we’. At the other end, ‘we’
means: 'you and I, and I'm none too sure about you'. (Or it might be 'I, and those who think as I do', which reduces to 'I'.) In between, 'we' means: 'I, and all those who are of a common culture with me'. Since mankind even at one moment is not made up of isolated and homogeneous tribes, and since we should not limit ourselves to the part of mankind located at one moment, we may haggle endlessly over how much cultural affiliation is meant.

(We have a piece of unfinished business: if someone is to find out about values by the canonical method, he must somehow know that he is one of the appropriate 'we'. All our versions, absolute or relative, make this knowledge automatic. Not so for elitist versions, on which 'we' means 'the best-qualified of us' or maybe 'the most normal of us'. But elitist versions are pointless. We're already considering dispositions under extravagantly ideal conditions; we needn't idealise all over again by being selective about who counts as one of the 'we'.)

If some relative version were the correct analysis, wouldn't that be manifest whenever people talk about value? Wouldn't you hear them saying 'value for me and my mates' or 'value for the likes of you'? Wouldn't you think they'd stop arguing after one speaker says X is a value and the other says it isn't?—Not necessarily. They might always presuppose, with more or less confidence (well-founded or otherwise), that whatever relativity there is won't matter in this conversation. Even if they accept in principle that people sometimes just differ in their dispositions to value, they may be very reluctant to think the present deadlocked conversation is a case of such difference. However intractable the disagreement may be, they may go on thinking it really is a disagreement: a case in which two people are disposed alike, but one of them is wrong about what is a value relative to their shared dispositions, because he is not valuing as he would under ideal conditions. So long as they think that—and they might think it very persistently—they can hold the language of explicit relativity in reserve. It is there as a last resort, if ever they meet with a proven case of ultimate difference. But it will not be much heard, since it is a practical impossibility to prove a case. If the language of absolutism prevails, that is not strong evidence against relativity.

(Those who have heard of the relativity of simultaneity do not
manifest this knowledge all the time. They speak as the ignorant do, and no harm done. They’ll resort to the language of relativity when it matters, say in discussing the exploits of the interstellar navy.

*Does* the language of absolutism prevail? Not really. With some of us it does. Others of us resort to the language of relativity at the drop of a hat. Yet this too is poor evidence. The eager relativists may have been confused by philosophy. For who can escape it?

So what version should we prefer, absolute or relative?—Neither; instead, I commend a *wait-and-see* version. In making a judgement of value, one makes many claims at once, some stronger than others, some less confidently than others, and waits to see which can be made to stick. I say X is a value; I mean that all mankind are disposed to value X; or anyway all nowadays are; or anyway all nowadays are except maybe some peculiar people on distant islands; or anyway . . . ; or anyway you and I, talking here and now, are; or anyway I am.19 How much am I claiming?—as much as I can get away with. If my stronger claims were proven false—though how that could be proven is hard to guess—I still mean to stand by the weaker ones. So long as I’m not challenged, there’s no need to back down in advance; and there’s no need to decide how far I’d back down if pressed. What I mean to commit myself to is conditionally relative: relative if need be, but absolute otherwise.

*What is the modal status of the equivalence?* The equivalence between value and what we are disposed to value is meant to be a piece of philosophical analysis, therefore analytic. But of course it is not obviously analytic; it is not even obviously true.

It is a philosophical problem how there can ever be unobvious analyticity. We need not solve that problem; suffice it to say that it is everybody’s problem, and it is not to be solved by denying the phenomenon. There are perfectly clear examples of it: the epsilon-delta analysis of an instantaneous rate of change, for one. Whenever it is analytic that all A’s are B’s, but not obviously analytic, the Moorean open question—whether all A’s are indeed B’s—is intelligible. And not only is it intelligible

in the sense that we can parse and interpret it (that much is true even of the question whether all A’s are A’s) but also in the sense that it makes sense as something to say in a serious discussion, as an expression of genuine doubt.

Besides unobvious analyticity, there is equivocal analyticity. Something may be analytic under one disambiguation but not another, or under one precisification but not another. Examples abound. Quine was wrong that analyticity was unintelligible, right to doubt that we have many clearcut cases of it. If differing versions of a concept (or, if you like, different but very similar concepts) are in circulation under the same name, we will get equivocal analyticity. It is analytic under one disambiguation of ‘dog’ that all dogs are male; under one disambiguation of ‘bitch’ that all bitches are canine. It is analytic under some precisifications of ‘mountain’ that no mountain is less than one kilometre high. When analyticity is equivocal, open questions make good conversational sense: they are invitations to proceed under a disambiguation or precisification that makes the answer to the question not be analytic. By asking whether there are mountains less than one kilometre high, you invite your conversational partners to join you in considering the question under a precisification of ‘mountain’ broad enough to make it interesting; yet it was analytic under another precisification that the answer was ‘no’. So even if all is obvious, open questions show at worst that the alleged analyticity is equivocal.

I suggest that the dispositional theory of value, in the version I have put forward, is equivocally as well as unobviously analytic. I do not claim to have captured the one precise sense that the word ‘value’ bears in the pure speech, uncorrupted by philosophy, that is heard on the Clapham omnibus. So far as this matter goes, I doubt that speakers untouched by philosophy are found in Clapham or anywhere else. And if they were, I doubt if they’d have made up their minds exactly what to mean any more than the rest of us have. I take it, rather, that the word ‘value’, like many others, exhibits both semantic variation and semantic indecision. The best I can hope for is that my dispositional theory lands somewhere near the middle of the

---


---

This content downloaded from 90.230.11.108 on Thu, 22 Sep 2016 18:52:01 UTC
All use subject to http://about.jstor.org/terms
range of variation and indecision—and also gives something that I, and many more besides, could be content to adopt as our official definition of the word 'value', in the unlikely event that we needed an official definition.

I've left some questions less than conclusively settled: the matter of absolute versus relative versus wait-and-see versions, the details of 'ideal conditions', the question of admitting values _de se_, the definition of valuing as second-order versus highest-order intrinsic desiring. It would not surprise or disturb me to think that my answers to those questions are only equivocally analytic—but somewhere fairly central within the range of variation and indecision—and that the same could be said of rival answers. Even if no version of the dispositional theory is unequivocally analytic, still it's fair to hope that some not-too-miscellaneous disjunction of versions comes out analytic under most reasonable resolutions of indeterminacy (under some reasonable precisification of 'most' and 'reasonable'.)

If the dispositional theory is only unobviously and equivocally analytic, why think that it's analytic at all?—Because that hypothesis fits our practice. (The practice of many of us, much of the time.) It does seem that if we try to find out whether something is a genuine value, we do try to follow—or rather, approximate—the canonical method. We gain the best imaginative acquaintance we can, and see if we then desire to desire it. In investigating values by the canonical method, we ignore any alleged possibility that values differ from what we're disposed to value. The dispositional theory explains nicely why we ignore it: no such possibility exists.

Now this should sound an alarm. Phenomenalism, behaviourism, and the like might be supported in exactly the same way: we ignore the possibility that our method of investigation deceives us radically, and the alleged explanation is that no such possibility exists. But in those cases, we know better. We know how systematic hallucination might deceive its victim about the world around him, and how a clever actor might deceive everyone he meets about his inner life (and, in both cases, how it might be that experience or behaviour would remain deceptive throughout the appropriate range of counterfactual suppositions). And it doesn't just strike us that such deception is possible somehow. Rather, we can imagine just how it might happen. We
can give a story of deception all the detail it takes to make it convincing. So we must confess that our method of gaining knowledge of the outer world and the inner lives does consist in part of ignoring genuine possibilities—possibilities that cannot credibly be denied.

The case of value is different, because the convincing detail cannot be supplied. Yes, you might think that perhaps the genuine values somehow differ from what we are disposed to value, even under ideal conditions. (Charles Pigden has noted that a misanthrope might think it because he thinks mankind is irremediably depraved.) The conjecture is not unthinkable; the dispositional theory is not obviously analytic; counterexamples are not obviously impossible. That is not yet much evidence of possibility. Better evidence would be a detailed story of just how it might happen that something—something specific—is after all a value that we are not disposed to value, or a non-value that we are disposed to value. But I have no idea how to flesh out the story. Without 'corroborative detail', insistence that there exist such possibilities is 'bald and unconvincing'. This time, nothing outweighs the niceness of explaining the ignoring by denying the possibilities allegedly ignored.

But is it realism? Psychology is contingent. Our dispositions to value things might have been otherwise than they actually are. We might have been disposed, under ideal conditions, to value seasickness and petty sleaze above all else. Does the dispositional theory imply that, had we been thus disposed, those things would have been values? That seems wrong.

No: we can take the reference to our dispositions as rigidified. Even speaking within the scope of a counterfactual supposition, the things that count as values are those that we are actually disposed to value, not those we would have valued in the counterfactual situation. No worries—unless seasickness actually is a value, it still wouldn’t have been a value even if we’d been disposed to value it.

This is too swift. The trick of rigidifying seems more to hinder the expression of our worry than to make it go away. It can still be expressed as follows. We might have been disposed to value seasickness and petty sleaze, and yet we might have been no different in how we used the word ‘value’. The reference of ‘our
actual dispositions’ would have been fixed on different dispositions, of course, but our way of fixing the reference would have been no different. In one good sense—though not the only sense—we would have meant by ‘value’ just what we actually do. And it would have been true for us to say ‘seasickness and petty sleaze are values’.

The contingency of value has not gone away after all; and it may well disturb us. I think it is the only disturbing aspect of the dispositional theory. Conditional relativity may well disturb us too, but that is no separate problem. What comfort would it be if all mankind just happened to be disposed alike? Say, because some strange course of cultural evolution happened to be cut short by famine, or because some mutation of the brain never took place? Since our dispositions to value are contingent, they certainly vary when we take all of mankind into account, all the inhabitants of all the possible worlds. Given the dispositional theory, trans-world relativity is inevitable. The spectre of relativity within our own world is just a vivid reminder of the contingency of value.

If wishes were horses, how would we choose to ride? What would it take to satisfy us? Maybe this new version of the dispositional theory would suit us better: values are what we’re necessarily disposed to value. Then no contingent ‘value’ would deserve the name; and there would be no question of something being a value for some people and not for others, since presumably what’s necessary is a fortiori uniform (unless different dispositions to value are built into different people’s individual essences, an unlikely story).

What kind of necessity should it be? Not mere deontic necessity—values are what we’re disposed to value on pain of being at fault, where the fault in question turns out to consist in failing to be disposed to value the genuine values. That dispositional theory is empty. Its near relatives are nearly empty. And it won’t help to juggle terms; as it might be, by calling it ‘rational necessity’ and then classifying the disposition to value genuine values as a department of ‘rationality’. Probably not nomological necessity either—small comfort to think that we were disposed to disvalue seasickness only because, luckily, our neurons are not subject to a certain fifth force of nature that would distort their workings in just the wrong way. It had better
be necessity *simpliciter*, so-called 'metaphysical' necessity.

If we amend the dispositional theory by inserting 'necessarily', we can be much more confident that the 'values' it defines would fully deserve the name—if there were any of them. But it is hard to see how there possibly could be. If a value, strictly speaking, must be something we are necessarily disposed to value, and if our dispositions to value are in fact contingent, then, strictly speaking, there are no values. If Mackie is right that a value (his term is 'objective good') would have to be

sought by anyone who was acquainted with it, not because of any contingent fact that this person, or every person, is so constituted that he desires this end, but just because the end has to-be-pursuedness somehow built into it,

then he is also right to call values 'queer' and to repudiate the error of believing in them.21 (Replacing 'sought' by 'valued' would not change that.) If we amend the dispositional theory, requiring values to be all that we might wish them to be, we bring on the error theory. The fire is worse than the frying pan.

Is it, after all, out of the question that our dispositions to value might be necessary? If the theory of mind I favour is true, then the platitudes of folk psychology do have a certain necessity—albeit conditional necessity—to them.22 There are states that play the functional roles specified in those platitudes, and it is in virtue of doing so that they deserve their folk-psychological names. It is not necessary that there should be any states in us that deserve such names as 'pain', 'belief', or 'desire'. But it is necessary that if any states do deserve those names, then they conform to the platitudes. Or rather, they conform well enough. Now suppose that some of the platitudes of folk psychology specified exactly what we were disposed, under ideal conditions, to desire to desire. And suppose those platitudes were non-negotiable: if a system of states did not satisfy them, that would settle that those states did not conform well enough to folk

---

134 II—DAVID LEWIS

---

21 J. L. Mackie, *Ethics: Inventing Right and Wrong* (Penguin, 1977), p. 40. But note that the queerness Mackie has in mind covers more than just the to-know-it-is-to-love-it queerness described in this passage.

psychology to deserve the mental names it implicitly defines. Then there would be things we were necessarily disposed to value—on condition that we had mental lives at all!

The suggestion is intelligible and interesting, but too good to be true. For one thing, it only spreads the trouble. Instead of losing the risk that nothing deserves the name of value, we gain the added risk that nothing deserves commonplace folk-psychological names. Pace the Churchlands, it's not really credible that there might turn out to be no beliefs, no desires, no pains, . . . For another thing, it proves too much. It denies outright that it's possible for someone to differ from others in his dispositions to value. Yet this does seem possible; and we can flesh out the story with plenty of 'corroborative detail'. This cunning and subtle villain once was as others are; he gained excellent imaginative acquaintance with many values, and valued them accordingly. Now he has gone wrong, and cares not a fig for what he once valued; and yet he has forgotten nothing. (He certainly has not stopped having any mental life deserving of the name.) He hates those who are as he once was, and outwits them all the better because of his superb empathetic understanding of what they hold dear. Could it not happen?—not if the present suggestion were true. So the present suggestion is false. Yet it was the only hope, or the only one I know, for explaining how there might be things we are necessarily disposed to value. The dispositions are contingent, then. And, at least in some tacit way, we know it. If the story of the subtle villain strikes you as a possible story, that knowledge thereby reveals itself.

But if we know better, it is odd that we are disturbed—as I think many of us will be—by a dispositional theory of value, unamended, according to which values are contingent. It feels wrong. Why might that be?—Perhaps because a large and memorable part of our discussion of values consists of browbeating and being browbeaten. The rhetoric would fall flat if we kept in mind, all the while, that it is contingent how we are disposed to value. So a theory which acknowledges that contingency

---

23 As argued in Frank Jackson and Philip Pettit, 'In Defence of Folk Psychology', forthcoming in Philosophical Studies.

cannot feel quite right. You might say that it is unfaithful to the
distinctive phenomenological character of lived evaluative
thought. Yet even if it feels not right, it may still be right, or as
near right as we can get. It feels not quite right to remember that
your friends are big swarms of little particles—it is inadequate to
the phenomenology of friendship—but still they are.

I suggested earlier that my version of the dispositional theory
of value might be equivocally analytic. So might the amended
version, on which values are what we are necessarily disposed to
value. Between these two versions, not to mention others, there
might be both semantic variation and semantic indecision. If so,
it is part of a familiar pattern. One way to create indeterminacy
and equivocal analyticity is to define names implicitly in terms
of a theory (folk or scientific), and later find out that the theory is
wrong enough that nothing perfectly deserves the names so
introduced, but right enough that some things, perhaps several
rival candidates, deserve the names imperfectly. Nothing
perfectly deserves the name ‘simultaneity’, since nothing quite
fits the whole of our old conception. So the name will have to go
to some imperfect deserver of it, or to nothing. What it takes to
deserve this name, not perfectly but well enough, was never
officially settled. One resolution of the indeterminacy makes it
analytic that simultaneity must be frame-independent; another,
that it must be an equivalence relation; a third, that it must be
both at once. The third brings with it an error theory of
simultaneity.25

I suggest that (for some of us, or some of us sometimes) the
amended dispositional theory best captures what it would take
for something to perfectly deserve the name ‘value’. There are
no perfect deservers of the name to be had. But there are plenty
of imperfect deservers of the name, and my original version is
meant to capture what it takes to be one of the best of them. (But
I do not say mine is the only version that can claim to do so.
Doubtless there are more dimensions of semantic variation and
indeterminacy than just our degree of tolerance for imperfection.)
Strictly speaking, nothing shall get the name without deserving
it perfectly. Strictly speaking, Mackie is right: genuine values

25 See Hartry Field, ‘Theory Change and the Indeterminacy of Reference’, Journal of
would have to meet an impossible condition, so it is an error to think there are any. Loosely speaking, the name may go to a claimant that deserves it imperfectly. Loosely speaking, common sense is right. There are values, lots of them, and they are what we are disposed de facto to value.

Then is my position a form of realism about values?—Irrealism about values strictly speaking, realism about values loosely speaking. The former do not exist. The latter do.

What to make of the situation is mainly a matter of temperament. You can bang the drum about how philosophy has uncovered a terrible secret: there are no values! (Shock horror: no such thing as simultaneity! Nobody ever whistled while he worked!) You can shout it from the housetops—browbeating is oppression, the truth shall make you free.26 Or you can think it better for public safety to keep quiet and hope people will go on as before. Or you can declare that there are no values, but that nevertheless it is legitimate—and not just expedient—for us to carry on with value-talk, since we can make it all go smoothly if we just give the name of value to claimants that don’t quite deserve it. This would be a sort of quasi-realism, not the same as Blackburn’s quasi-realism.27 Or you can think it an empty question whether there are values: say what you please, speak strictly or loosely. When it comes to deserving a name, there’s better and worse but who’s to say how good is good enough? Or you can think it clear that the imperfect deservers of the name are good enough, but only just, and say that although there are values we are still terribly wrong about them. Or you can calmly say that value (like simultaneity) is not quite as some of us sometimes thought. Myself, I prefer the calm and conservative responses. But so far as the analysis of value goes, they’re all much of a muchness.

26 See Hinckfuss, op. cit.
ARE WE FREE TO BREAK THE LAWS?

DAVID LEWIS

Soft determinism seems to have an incredible consequence. It seems to imply, given certain acceptable further premises, that sometimes we are able to act in such a way that the laws of nature are broken. But if we distinguish a strong and a weak version of this incredible consequence, I think we shall find that it is the strong version that is incredible and the weak version that is the consequence.

Soft determinism is the doctrine that sometimes one freely does what one is predetermined to do; and that in such a case one is able to act otherwise though past history and the laws of nature determine that one will not act otherwise.

Compatibilism is the doctrine that soft determinism may be true. A compatibilist might well doubt soft determinism because he doubts on physical grounds that we are ever predetermined to act as we do, or perhaps because he doubts on psychoanalytic grounds that we ever act freely. I myself am a compatibilist but no determinist, hence I am obliged to rebut some objections against soft determinism but not others. But for the sake of the argument, let me feign to uphold soft determinism, and indeed a particular instance thereof.

I have just put my hand down on my desk. That, let me claim, was a free but predetermined act. I was able to act otherwise, for instance to raise my hand. But there is a true historical proposition \( H \) about the intrinsic state of the world long ago, and there is a true proposition \( L \) specifying the laws of nature that govern our world, such that \( H \) and \( L \) jointly determine what I did. They jointly imply the proposition that I put my hand down. They jointly contradict the proposition that I raised my hand. Yet I was free; I was able to raise my hand. The way in which I was determined not to was not the sort of way that counts as inability.

What if I had raised my hand? Then at least one of three things would have been true. Contradictions would have been true together; or the historical proposition \( H \) would not have been true; or the law proposition \( L \) would not have been true. Which? Here we need auxiliary premises; but since I accept the premises my opponent requires to make his case, we may proceed. Of our three alternatives, we may dismiss the first; for if I had raised my hand, there would still have been no true contradictions. Likewise we may dismiss the second; for if I had raised my hand, the intrinsic state of the world long ago would have been no different. That leaves the third alternative. If I had raised my hand, the law proposition \( L \) would not have been true. That follows by a principle of the logic of counterfactuals which is almost uncontroversial:

\[
A \rightarrow (B \lor C) \lor D, A \rightarrow \neg B, A \rightarrow \neg C, A \rightarrow D.
\]

If \( L \) had not been true, that implies that some law of nature would have been broken, for \( L \) is a specification of the laws. That is not to say that anything would have been both a law and broken—that is a contradiction in terms if, as I suppose, any genuine law is at least an absolutely unbroken regularity. Rather, if \( L \) had not been true, something that is in fact a law, and unbroken, would have been broken, and no law. It would at best have been an almost-law.

In short, as a (feigned) soft determinist, who accepts the requisite auxiliary premises and principle of counterfactual logic, I am committed to the consequence that if I had done what I was able to do—raise my hand—then some law would have been broken.

"That is to say," my opponent paraphrases, "you claim to be able to break the very laws of nature. And with so little effort! A marvelous power indeed! Can you also bend spoons?"

*Distinguo.* My opponent's paraphrase is not quite right. He has replaced the weak thesis that I accept with a stronger thesis that I join him in rejecting. The strong thesis is utterly incredible, but it is no part of soft determinism. The weak thesis is controversial, to be sure, but a soft determinist should not mind being committed to it. The two theses are as follows.

(Weak Thesis) I am able to do something such that, if I did it, a law would be broken.

(Strong Thesis) I am able to break a law.

To see the difference, consider not a marvelous ability to break a law but a commonplace ability to break a window. Perhaps I am able to throw a stone...
in a certain direction; and perhaps if I did, the stone would hit a certain window and the window would break. Then I am able to break a window. For starters: I am able to do something such that, if I did it, a window would be broken. But there is more to be said. I am able to do something such that, if I did it, my act would cause a window-breaking event.

Or consider a commonplace ability to break a promise. Perhaps I am able to throw a stone; and perhaps if I did, I would break my promise never to throw a stone. Then I am able to break a promise. For starters: I am able to do something such that, if I did it, a promise would be broken. But there is more to be said. I am able to do something such that, if I did it, my act would itself be a promise-breaking event.

Next, consider what really would be a marvelous ability to break a law—an ability I could not credibly claim. Suppose that I were able to throw a stone very, very hard. And suppose that if I did, the stone would fly faster than light, an event contrary to law. Then I really would be able to break a law. For starters: I would be able to do something such that, if I did it, a law would be broken. But there is more to be said. I would be able to do something such that, if I did it, my act would cause a law-breaking event.

Or suppose that I were able to throw a stone so hard that in the course of the throw my own hand would move faster than light. Then again I would be able to break a law, regardless of what my act might cause. For starters: I would be able to do something such that, if I did it, a law would be broken. But there is more to be said. I would be able to do something such that, if I did it, my act would itself be a law-breaking event.

If no act of mine either caused or was a window-, promise-, or law-breaking event, then I think it could not be true that I broke a window, a promise, or a law. Therefore I am able to break a window, a promise, or a law only if I am able to do something such that, if I did it, my act either would cause or would be a window-, promise-, or law-breaking event.

Maybe my opponent will contend that according to soft determinism, there is another way of being able to break a law. But I see no reason to grant this contention.

Now consider the disputed case. I am able to raise my hand, although it is predetermined that I will not. If I raised my hand, some law would be broken. I even grant that a law-breaking event would take place. (Here I use the present tense neutrally. I mean to imply nothing about my act or even what would happen.) But is it so that my act of raising my hand would cause any law-breaking event? Is it so that my act of raising my hand would itself be a law-breaking event? Is it so that any other act of mine would cause or would be a law-breaking event? If not, then my ability to raise my hand confers no marvelous ability to break a law, even though a law would be broken if I did it.

Had I raised my hand, a law would have been broken beforehand. The course of events would have diverged from the actual course of events a little while before I raised my hand, and at the point of divergence there would have been a law-breaking event—a divergence miracle, as I have called it (4). But this divergence miracle would not have been caused by my raising my hand. If anything, the causation would have been the other way around. Nor would the divergence miracle have been my act of raising my hand. That act was altogether absent from the actual course of events, so it cannot get under way until there is already some divergence. Nor would it have been caused by any other act of mine, earlier or later. Nor would it have been any other act of mine. Nor is there any reason to say that if I had raised my hand there would have been some other law-breaking event besides the divergence miracle; still less, that some other law-breaking event would have been caused by, or would have been, my act of raising my hand. To accommodate my hypothetical raising of my hand while holding fixed all that can and should be held fixed, it is necessary to suppose one divergence miracle, gratuitous to suppose any further law-breaking.

Thus I insist that I was able to raise my hand, and I acknowledge that a law would have been broken had I done so, but I deny that I am therefore able to break a law. To uphold my instance of soft determinism, I need not claim any incredible powers. To uphold the compatibilism that I actually believe, I need not claim that such powers are even possible.

I said that if I had raised my hand, the divergence miracle beforehand would not have been caused by my raising my hand. That seems right. But my opponent might argue ad hominem that according to my own analysis of causation (3), my raising my hand does turn out to cause the divergence miracle. The effect would precede the cause, but I do not object to that. We seem to have the right pattern of counterfactual dependence between

---

1 Up to a point, my strategy here resembles that of Keith Lehrer ([2], p. 199). Lehrer grants a weak thesis: the agent could have done something such that, if he had done it, there would have been a difference in either laws or history. He rejects, as I would, the step from that to a stronger thesis: the agent could have brought about a difference in laws or history. So far, so good. But Lehrer's reason for rejecting the stronger thesis is one I cannot accept. His reason is this: it is false that if the agent had preferred that there be a difference in laws or history, there would have been a difference in laws or history. I say, first, that this conditional may not be false. Suppose the agent is predetermined to prefer that there be no difference; had he preferred otherwise, there would have been a difference. (Had anything been otherwise than it was predetermined to be, there would have been a difference in either laws or history.) And second, if this conditional is not false, that is not enough to make the stronger thesis true. There must be some other reason, different from the one Lehrer gives, why the stronger thesis is false.
distinct events: (1) if I had raised my hand, the divergence miracle would have occurred, but (2) if I had not raised my hand, it would not have occurred.

I reply that we do not have this required pattern, nor would we have had it if I had raised my hand. Therefore I am safe in denying that the miracle would have been caused by my act.

We do not have the pattern because (1) is false. What is true is only that if I had raised my hand, then some or other divergence miracle would have occurred. There is no particular divergence miracle that definitely would have occurred, since the divergence might have happened in various ways.4

If I had raised my hand, (1) would have been true. But we still would not have had the right pattern, because in that case (2) would have been false. Consider a counterfactual situation in which a divergence miracle beforehand has allowed me to raise my hand. Is it so, from the standpoint of that situation, that if I had not raised my hand, the miracle would not have taken place? No; the miracle might have taken place, only to have its work undone straightway by a second miracle. (Even in this doubly counterfactual context, when I speak of a miracle I mean a violation of the actual laws.) What is true, at most, is that if I had not raised my hand, then the first miracle might not have taken place.

My incompatibilist opponent is a creature of fiction, but he has his prototypes in real life. He is modelled partly on Peter van Inwagen ([5], [6], [7]) and partly on myself when I first worried about van Inwagen’s argument against compatibilism. He definitely is not van Inwagen; he does not choose his words so carefully. Still I think that for all his care, van Inwagen is in the same boat with my fictitious opponent.

Van Inwagen’s argument runs as follows, near enough. (I recast it as a reductio against the instance of soft determinism that I feign to uphold.) I did not raise my hand; suppose for reductio that I could have raised my hand, although determinism is true. Then it follows, given four premises that I cannot question, that I could have rendered false the conjunction HL of a certain historical proposition H and a certain law proposition L. If so, then I could have rendered L false. (Premise 5.) But I could not have rendered L false. (Premise 6.) This refutes our supposition.

To this I reply that Premise 5 and Premise 6 are not both true. Which one is true depends on what van Inwagen means by “could have rendered false”.

It does not matter what “could have rendered false” means in ordinary language; van Inwagen introduced the phrase as a term of art. It does not even matter what meaning van Inwagen gave it. What matters is whether we can give it any meaning that would meet his needs—any meaning that would make all his premises defensible without circularity. I shall consider two meanings. I think there is nothing in van Inwagen’s text to suggest any third meaning that might work better than these two.5

First, a preliminary definition. Let us say that an event would falsify a proposition iff, if that event occurs then that proposition is false. For instance, an event consisting of a stone’s flying faster than light would falsify a law. So would an act of throwing in which my hand moves faster than light. So would a divergence miracle. But my act of throwing a stone would not itself falsify the proposition that the window in the line of fire remains intact; all that is true is that my act would cause another event that would falsify that proposition. My act of raising my hand would falsify any sufficiently inclusive conjunction of history and law. But it would not itself falsify any law—not if all the requisite law-breaking were over and done with beforehand. All that is true is that my act would be preceded by another event—the divergence miracle—that would falsify a law.

Let us say that I could have rendered a proposition false in the weak sense iff I was able to do something such that, if I did it, the proposition would have been falsified (though not necessarily by my act, or by any event caused by my act). And let us say that I could have rendered a proposition false in the strong sense iff I was able to do something such that, if I did it, the proposition would have been falsified either by my act itself or by some event caused by my act.

The Weak Thesis, which as a soft determinist I accept, is the thesis that I could have rendered a law false in the weak sense. The Strong Thesis, which I reject, is the thesis that I could have rendered a law false in the strong sense.

The first part of van Inwagen’s argument succeeds whichever sense we take. If I could have raised my hand despite the fact that determinism is true

4 Cf. [4], p. 463. At this point I am relying on contingent features of the world as we suppose it to be; as Allen Hazen has pointed out to me, we can imagine a world of discrete processes at which one divergent history in which I raise my hand clearly takes less of a miracle than any of its rivals. I think this matters little, since the task of compatibilism is to show how freedom and determinism might coexist at a world that might, for all we know, be ours.

5 Van Inwagen has indicated (personal communication, 1981) that he would adopt a third meaning for “could have rendered false”, different from both of the meanings that I discuss here. His definition is roughly as follows: an agent could have rendered a proposition false iff he could have arranged things in a certain way, such that his doing so, plus the whole truth about the past, together strictly imply the falsehood of the proposition. On this definition, Premise 6 simply says that I could not have arranged things in any way such that I was predetermined not to arrange things in that way. It is uninformative to learn that the soft determinist is committed to denying Premise 6 thus understood.
and I did not raise it, then indeed it is true both in the weak sense and in the strong sense that I could have rendered false the conjunction $HL$ of history and law. But I could have rendered false the law proposition $L$ in the weak sense, though I could not have rendered $L$ false in the strong sense. So if we take the weak sense throughout the argument, then I deny Premise 6. If instead we take the strong sense, then I deny Premise 5.

Van Inwagen supports both premises by considering analogous cases. I think the supporting arguments fail because the cases produced are not analogous: they are cases in which the weak and strong senses do not diverge. In support of Premise 6, he invites us to reject the supposition that a physicist could render a law false by building and operating a machine that would accelerate protons to twice the speed of light. Reject that supposition by all means; but that does nothing to support Premise 6 taken in the weak sense, for the rejected supposition is that the physicist could render a law false in the strong sense. In support of Premise 5, he invites us to reject the supposition that a traveler could render false a conjunction of a historical proposition and a proposition about his future travels otherwise than by rendering false the nonhistorical conjunct. Reject that supposition by all means, but that does nothing to support Premise 5 taken in the strong sense. Given that one could render false, in the strong sense, a conjunction of historical and nonhistorical propositions (and given that, as in the cases under consideration, there is no question of rendering the historical conjunct false by means of time travel or the like), what follows? Does it follow that one could render the nonhistorical conjunct false in the strong sense? That is what would support Premise 5 in the strong sense. Or does it only follow, as I think, that one could render the nonhistorical conjunct false in at least the weak sense? The case of the traveler is useless in answering that question, since if the traveler could render the proposition about his future travels false in the weak sense, he could also render it false in the strong sense.

**References**


Argle. I believe in nothing but concrete material objects.

Bargle. There are many of your opinions I applaud; but one of your less pleasing characteristics is your fondness for the doctrines of nominalism and materialism. Every time you get started on any such topic, I know we are in for a long argument. Where shall we start this time: numbers, colors, lengths, sets, force-fields, sensations, or what?

Argle. Fictions all! I've thought hard about every one of them.

Bargle. A long evening's work. Before we start, let me find you a snack. Will you have some crackers and cheese?

Argle. Thank you. What splendid Gruyère!

Bargle. You know, there are remarkably many holes in this piece.

Argle. There are.

Bargle. Got you!

* * *

Bargle. You admit there are many holes in that piece of cheese. Therefore, there are some holes in it. Therefore, there are some holes. In other words, holes exist. But holes are not made of matter; to the contrary, they result from the absence of matter.

Argle. I did say that there are holes in the cheese; but that is not to imply that there are holes.

Bargle. However not? If you say that there are A's that are B's, you are committed logically to the conclusion that there are A's.

Argle. When I say that there are holes in something, I mean nothing more nor less than that it is perforated. The synonymous shape-predicates '... is perforated' and 'there are holes in ...'—just like any other shape-predicate, say '... is a dodecahedron'—may truly be predicated of pieces of cheese, without any implication that perforation is due to the presence of occult, immaterial entities. I am sorry my innocent predicate confuses you by sounding like an idiom of existential quantification, so that you think that inferences involving it are valid when they are not. But I have my reasons. You, given a perforated piece of cheese and believing as you do that it is perforated because it contains immaterial entities called holes, employ an idiom of existential quantification to say falsely 'There are holes in it'. Agreeable fellow that I am, I wish to have a sentence that sounds like yours...
and that is true exactly when you falsely suppose your existential quantification over immaterial things to be true. That way we could talk about the cheese without philosophizing, if only you'd let me. You and I would understand our sentences differently, but the difference wouldn't interfere with our conversation until you start drawing conclusions which follow from your false sentence but not from my homonymous true sentence.¹

*Bargle.* Oh, very well. But behold: there are as many holes in my piece of cheese as in yours. Do you agree?

*Argle.* I'll take your word for it without even counting: there are as many holes in mine as in yours. But what I mean by that is that either both pieces are singly-perforated, or both are doubly-perforated, or both are triply-perforated, and so on.

*Bargle.* What a lot of different shape-predicates you know! How ever did you find time to learn them all? And what does 'and so on' mean?²

*Argle.* Let me just say that the two pieces are equally-perforated. Now I have used only one two-place predicate.

*Bargle.* Unless I singly-perforate each of these crackers, how will you say that there are as many holes in my cheese as crackers on my plate? Be so kind as not to invent another predicate on the spot. I am quite prepared to go on until you have told me about all the predicates you have up your sleeve. I have a good imagination, and plenty of time.

*Argle.* Oh, dear... (ponders)

*Bargle.* I was wrong. There are holes.

*Argle.* You recant?

*Bargle.* I expected that sooner. You are thinking, doubtless, that every hole is filled with matter: silver amalgam, air, interstellar gas, luminiferous ether or whatever it may be.

*Argle.* No. Perhaps there are no truly empty holes; but I cannot deny that there might be.

*Bargle.* How can something utterly devoid of matter be made of matter?

*Argle.* You're looking for the matter in the wrong place. (I mean to say, that's what you would be doing if there were any such things as places, which there aren't.) The matter isn't inside the hole. It would be absurd to say it was: nobody wants to say that holes are inside themselves. The matter surrounds the hole. The lining of a hole, you agree, is a material object. For every hole there is a hole-lining; for every hole-lining there is a hole. I say the hole-lining is the hole.

David and Stephanie Lewis

Bargle. Didn't you say that the hole-lining surrounds the hole? Things don't surround themselves.

Argle. Holes do. In my language, 'surrounds' said of a hole (described as such) means 'is identical with'. 'Surrounds' said of other things means just what you think it means.

Bargle. Doesn't it bother you that your dictionary must have two entries under 'surrounds' where mine has only one?

Argle. A little, but not much. I'm used to putting up with such things.

Bargle. Such what?

Argle. Such dictionary entries. They're made of dried ink, you recall.

Bargle. Oh. I suppose you'll also say that '... is in ...' or '... is through ...' said of a hole means '... is part of ...'.

Argle. Exactly so, Bargle.

Bargle. Then do you still say that 'There are holes in the cheese' contains an unanalyzed shape-predicate synonymous with '... is perforated'?

Argle. No; it is an existential quantification, as you think it is. It means that there exist material objects such that they are holes and they are parts of the piece of cheese.

* * *

Bargle. But we wouldn't say, would we, that a hole is made out of cheese?

Argle. No; but the fact that we wouldn't say it doesn't mean it isn't true. We wouldn't have occasion to say, unless philosophizing, that these walls are perpendicular to the floor; but they are. Anyhow we do say that caves are holes in the ground and that some of them are made out of limestone.

* * *

Bargle. Take this paper-towel roller. Spin it on a lathe. The hole-lining spins. Surely you'd never say the hole spins?

Argle. Why not?

Bargle. Even though the hole might continue to be entirely filled with a dowel that didn't spin or move at all?

Argle. What difference does that make?

Bargle. None, really. But now I have you: take a toilet-paper roller, put it inside the paper-towel roller, and spin it the other way. The big hole spins clockwise. The little hole spins counter-clockwise. But the little hole is part of the big hole, so it spins clockwise along with the rest of the big hole. So if holes can spin, as you think, the little hole turns out to be spinning in both directions at once, which is absurd.

Argle. I see why you might think that the little hole is part of the big hole, but you can't expect me to agree. The little hole is inside the big hole, but that's all. Hence I have no reason to say that the little hole is spinning clockwise.

* * *

Bargle. Consider a thin-walled hole with a gallon of water inside. The volume of the hole is at least a gallon, whereas the volume of the hole-
Holes

lining is much less. If the hole is the hole-lining, then whatever was true of one would have to be true of the other. They could not differ in volume.

Argle. For 'hole' read 'bottle'; for 'hole-lining' also read 'bottle'. You have the same paradox. Holes, like bottles, have volume—or, as I'd rather say, are voluminous or equi-voluminous with other things—in two different senses. There's the volume of the hole or bottle itself, and there's the volume of the largest chunk of fluid which could be put inside the hole or bottle without compression. For holes, as for bottles, contextual clues permit us to keep track of which we mean.

*   *   *

Bargle. What is the volume of the hole itself? How much of the cheese do you include as part of one of these holes? And how do you decide? Arbitrarily, that's how. Don't try saying you include as little of the cheese as possible, for however much you include, you could have included less.

Argle. What we call a single hole is really many hole-linings. Some include more of the cheese, some include less. Therefore I need not decide, arbitrarily or otherwise, how much cheese is part of the hole. Many different decisions are equally correct.

Bargle. How can a single hole be identical with many hole-linings that are not identical with one another?

Argle. Really there are many different holes, and each is identical with a different hole-lining. But all these different holes are the same hole.

Bargle. You contradict yourself. Don't you mean to say that they all surround the same hole—where by 'surround' I mean 'surround', not 'be identical with'?

Argle. Not at all. I would contradict myself if I said that two different holes were identical. But I didn't; what I said was that they were the same hole. Two holes are the same hole when they have a common part that is itself a hole.

Bargle. You agreed before that there were as many holes in my cheese as crackers on my plate. Are there still?

Argle. Yes; there are two of each left.

Bargle. Two crackers, to be sure, but how can you say there are two holes?

Argle. Thus: there is a hole, and there is another hole that is not the same hole, and every hole in the cheese is the same hole as one or the other.

Bargle. Be so kind as to say 'co-perforated', not 'same', and stop pretending to talk about identity when you are not. I understand you now: co-perforation is supposed to be an equivalence relation among hole-linings, and when you say there are two holes you are trying to say that there are two non-identical co-perforation-classes of hole-linings. Really you identify holes not with hole-linings but with classes of hole-linings.

Argle. I would if I could, but I can't. No; holes are hole-linings; but when I speak of them as holes, I find it convenient to use 'same' meaning...
'co-perforated' wherever a man of your persuasion would use 'same' meaning 'identical'. You know my reason for this trickery: my sentences about sameness of holes will be true just when you wrongly suppose your like-sounding sentences to be. The same goes for sentences about number of holes, since we both analyse these in terms of sameness.  

\* \* \*  

Bargle. You still haven't told me how you say there are as many holes in my cheese as crackers on my plate, without also saying how many there are.  

Argle. Here goes. There exist three things X, Y, and Z. X is part of the sum of the crackers, Y is part of the cheese, and Z is part of Y. Every maximal connected part of Y is a hole, and every hole in the cheese is the same hole as some maximal connected part of Y. X overlaps each of the crackers and Z overlaps each maximal connected part of Y. Everything which is either the intersection of X and a cracker or the intersection of Z and some maximal connected part of Y is the same size as any other such thing. X is the same size as Z.

\* \* \*  

Bargle. Your devices won't work because co-perforation is not an equivalence relation. Any two overlapping parts of my cheese have a common part that is a hole-lining, though in most cases the hole-lining is entirely filled with cheese. To be co-perforated is therefore nothing more than to overlap, and overlapping is no equivalence relation. The result is that although, as you say, you can find two hole-linings in this cheese that are not co-perforated, you can find another one that is co-perforated with both of them.

Argle. If you were right that a hole made of cheese could be entirely filled with the same kind of cheese, you could find far more than two non-co-perforated hole-linings; and there would be no such thing as cheese without holes in it. But you are wrong. A hole is a hole not just by virtue of its own shape but also by virtue of the way it contrasts with the matter inside it and around it. The same is true of other shape-predicates; I wouldn't say that any part of the cheese is a dodecahedron, though I admit that there are parts—parts that do not contrast with their surroundings—that are shaped like dodecahedra.

Bargle. Consider the paper-towel roller. How many holes?

Argle. One. You know what I mean: many, but they're all the same.

Bargle. I think you must say there are at least two. The left half and the right half are not the same hole. They have no common part, so no common part that is a hole.

Argle. They're not holes, they're two parts of a hole.

---


**Holes**

*Bargle.* Why aren't they holes themselves? They are singly-perforated and they are made of matter unlike the matter inside them. If I cut them apart you'd have to say they were holes?

*Argle.* Yes.

*Bargle.* You admit that a hole can be a proper part of a bigger—say, thicker-skinned—hole?

*Argle.* Yes.

*Bargle.* You admit that they are shaped like holes?

*Argle.* Yes, but they aren't holes. I can't say why they aren't. I know which things are holes, but I can't give you a definition. But why should I? You already know what hole-linings are. I say the two halves of the roller are only parts of a hole because I—like you—would say they are only parts of a hole-lining. What isn't a hole-lining isn't a hole.

*Bargle.* In that case, I admit that co-perforation may be an equivalence relation at least among singly-perforated hole-linings.

*Argle.* All holes are singly-perforated. A doubly-perforated thing has two holes in it that are not the same hole.

*Bargle.* Are you sure? Take the paper-towel roller and punch a little hole in its side. Now you have a hole in a hole-lining. You'd have to say you have a hole in a hole. You have a little hole which is part of a big hole; the big hole is not singly-perforated; and the little hole and the big hole are the same hole, since the little hole is a common part of each.

*Argle.* I think not. You speak of the big hole; but what we have are two big holes, not the same, laid end to end. There is also the little hole, not the same as either big hole, which overlaps them both. Of course we sometimes call something a hole, in a derivative sense, if it is a connected sum of holes. Any decent cave consists of many holes that are not the same hole, so I must have been speaking in this derivative sense when I said that caves are holes.

*Bargle.* What peculiar things you are driven to say when philosophy corrupts your mind! Tell me the truth: would you have dreamt for a moment of saying there were two big holes rather than one if you were not suffering under the influence of a philosophical theory?

*Argle.* No; I fear I would have remained ignorant.

*Bargle.* I see that I can never hope to refute you, since I no sooner reduce your position to absurdity than you embrace the absurdity.

*Argle.* Not absurdity; disagreement with common opinion.

*Bargle.* Very well. But I, for one, have more trust in common opinions than I do in any philosophical reasoning whatever. In so far as you disagree with them, you must pay a great price in the plausibility of your theories.

*Argle.* Agreed. We have been measuring that price. I have shown that it is not so great as you thought; I am prepared to pay it. My theories can earn credence by their clarity and economy; and if they disagree a little with
common opinion, then common opinion may be corrected even by a philosopher.

Bargle. The price is still too high.

Argle. We agree in principle; we're only haggling.

Bargle. We do. And the same is true of our other debates over ontic parsimony. Indeed, this argument has served us as an illustration—novel, simple, and self-contained—of the nature of our customary disputes.

Argle. And yet the illustration has interest in its own right. Your holes, had I been less successful, would have punctured my nominalistic materialism with the greatest of ease.

Bargle. Rehearsed and refreshed, let us return to—say—the question of classes.\(^5\)

Received September 1969

University of California
Los Angeles

\(^5\) There would be little truth to the guess that Argle is one of the authors and Bargle is the other. We thank Charles Chastain, who also is neither Argle nor Bargle, for many helpful comments.

Argie. I’ve said it before and I’ll say it again: all things are material. Either holes are somehow material, or else there are no such things. Maybe a hole is the material hole-lining that, as we so misleadingly say, “surrounds” the hole; or else whatever ostensible reference we make to holes is secretly some other sort of language-game altogether, or it’s fictitious reference, or it’s just plain mistaken.

Bargle. You’re ready to say anything, aren’t you, so long as it isn’t plain common sense. Of course what’s true is that holes are immaterial entities.—But what do these fellows think?

Casati and Varzi. Exactly so, Bargle: holes are immaterial entities. As Tucholsky put it: a hole is, where something isn’t.

Argie. If there were no matter at all, there’d be one big hole?

Casati and Varzi. No; a hole is always a hole in something: a cavity in the cheese, a hollow in a glass bottle, a tunnel through rock. A hole requires a host; and these hosts are material. (Normally, anyway; a hole in the electromagnetic field might be a region where the field vector is uniformly zero. Let’s ignore these special cases here.) No matter, no hosts; no hosts, no holes. Holes are dependent entities: they exist in virtue of the arrangement of matter.

Bargle. Right! And the hole is immaterial through and through, and the host is material; and besides, the hole is where the host isn’t. So the host and the hole are entirely distinct—the clearest case ever of necessary connection between distinct existences. And there’s another of Argle’s prejudices punctured!

Casati and Varzi. Steady on. These are deep matters, and there’s a lot we can say without becoming embroiled in them. Whether or not the dependence of holes upon matter means what either one of you thinks it means, at least it does mean that hole-statements are true or false in virtue of the

Argie. Right! The hole is redundant: it supervenes upon the arrangement of matter. So it is no genuine addition to reality; it is nothing over and above the matter it supervenes on. Its being—if being it be—must not be taken with ontological seriousness!

arrangement of matter. While we do have a view about the nature of holes—we endorse Bargle’s common sense, we reject the alternatives Argle has on offer—it scarcely matters. Take some hole-statements: that a hole is present in a certain material host; or that it is one or another kind of hole, cavity or tunnel or hollow or whatever; or that one hole is part of another; or that the same hole that was here yesterday is here today, though expanded or even shifted; or instead that there’s a new and different hole; or ... It’s the arrangement of matter that makes these statements true or false. You, Bargle, will agree with us that the hole-statements depend for their truth upon the holes, which in turn depend for their existence upon the arrangement of matter; whereas Argle will think that the truth of the hole-statements somehow depends more directly upon the arrangement of matter; but, either way, our job is to say how the truth of the hole-statements depends upon the arrangement of matter. And about that, we have quite a lot to say.

[Days pass]

Argle and Bargle. You do indeed. We’re impressed. But we do have a few questions for you. Take your taxonomy of holes, and stick to the simplest cases. There are cavities, with no entrance from the outside; hollows, with one entrance; tunnels, with two or more entrances ...

Casati and Varzi. And depressions: holes that are like hollows except that their entrances begin gradually, without a sharp edge to demarcate the surface inside the hole from the surface outside.

Bargle. Yes—for instance a test tube with a lip hosts a depression, whereas one without a lip hosts a hollow. That’s our problem. On page 6, you “distinguish three basic kinds of holes”; but on page 40 it’s four. We think you had it right the first time. Why do you divide hollows from depressions by whether there’s a demarcating edge, when you don’t divide other holes on the same principle? Some tunnels have entrances like the entrances of depressions, whereas others have entrances like the entrances of hollows ...

Argle. ... And even a single tunnel might have one kind of entrance at one end and the other at the other; or, if it’s a branching tunnel, two entrances of one kind and six of the other.

Bargle. And even a cavity might have a sharp edge dividing part of its inner surface from another part.

Argle. What happens when you fill a hole? No hole any more, right?

Casati and Varzi. That depends. In the first place, you might fill it incompletely, leaving a cavity. In the second place, you might leave a crack where
the surface of the hole used to be—a thin hole, maybe even a mere two-dimensional hole, but still a hole. But most important, you might fill it with the wrong sort of matter. A hole in the ground, filled with air or water, is still a hole. A hole in wood, filled with hardened putty, is still a hole.

Argle. For short: a hole filled seamlessly goes away, a hole filled not seamlessly is still there but it has stuff in it. But what if the host itself is none too seamless? Consider a hollow in a pile of miscellaneous rubbish. Fill the hollow with more miscellaneous rubbish. Plenty of inhomogeneity, plenty of cracks, plenty of cavities and hollows and tunnels; but the big hole you filled is gone without a trace.

Bargle. That’s part of a bigger problem. You have foremost in mind the case of a hole hosted by homogeneous and cohesive matter. But not all holes are hosted by homogeneous and cohesive matter. In fact, when you think of it microscopically, there’s no such thing as homogeneous matter—not in our world, anyway. And cohesiveness is very much a matter of degree.

Argle. Can you have a hole in a big dense swarm of bees?

Bargle. Can you have a hole in a stream of oncoming traffic? If not, how can you ever turn onto a busy road?

Argle. Perhaps some holes are strictly and literally holes, whereas others are holes by courtesy: holes by some sort of metaphorical extension from the case where the host is homogeneous and cohesive. But where do you draw the line?

Bargle. Last question. Why did you call hole-linings Ludovician holes? That’s what Argle thinks holes are. But who appointed him the spokesman for our authors?

DAVID LEWIS
Princeton University

STEPHANIE LEWIS
Municipal Capital Management, Inc.


Beyond Formalism is Jay Rosenberg’s attempt to articulate his dissatisfactions with the Kripkean “revolution” in the philosophy of language and to propose an alternative to it. According to Rosenberg, even though a “surprisingly large number of philosophers simply adopted the Kripkean ideas, images, and idioms root and branch” (xii), he has been “inarticulately irritated by Kripke’s views for almost twenty years” (xiii). Rosenberg claims that Kripke’s semantics for proper names and natural kind terms is a mis-
Many, but almost one

THE PROBLEM OF THE MANY

Think of a cloud — just one cloud, and around it clear blue sky. Seen from the ground, the cloud may seem to have a sharp boundary. Not so. The cloud is a swarm of water droplets. At the outskirts of the cloud the density of the droplets falls off. Eventually they are so few and far between that we may hesitate to say that the outlying droplets are still part of the cloud at all; perhaps we might better say only that they are near the cloud. But the transition is gradual. Many surfaces are equally good candidates to be the boundary of the cloud. Therefore many aggregates of droplets, some more inclusive and some less inclusive (and some inclusive in different ways than others), are equally good candidates to be the cloud. Since they have equal claim, how can we say that the cloud is one of these aggregates rather than another? But if all of them count as clouds, then we have many clouds rather than one. And if none of them count, each one being ruled out because of the competition from the others, then we have no cloud. How is it, then, that we have just one cloud? And yet we do.

This is Unger’s (1980) ‘problem of the many’. Once noticed, we can see that it is everywhere, for all things are swarms of particles.

---

There are always outlying particles, questionably parts of the thing, not definitely included and not definitely not included. So there are always many aggregates, differing by a little bit here and a little bit there, with equal claim to be the thing. We have many things or we have none, but anyway not the one thing we thought we had. That is absurd.

Think of a rusty nail, and the gradual transition from steel, to steel with bits of rust scattered through, to rust adhering to the nail, to rust merely resting on the nail. Or think of a cathode, and its departing electrons. Or think of anything that undergoes evaporation or erosion or abrasion. Or think of yourself, or any organism, with parts that gradually come loose in metabolism or excretion or perspiration or shedding of dead skin. In each case, a thing has questionable parts, and therefore is subject to the problem of the many.

If, as I think, things perdure through time by having temporal parts, then questionable temporal parts add to the problem of the many. If a person comes into existence gradually (whether over weeks or over years or over nanoseconds doesn’t matter for our present purpose) then there are questionable temporal parts at the beginning of every human life. Likewise at the end, even in the most sudden death imaginable. Do you think you are one person? – No, there are many aggregates of temporal parts, differing just a little at the ends, with equal claim to count as persons, and equal claim to count as you. Are all those equally good claims good enough? If so, you are many. If not, you are none. Either way we get the wrong answer. For undeniably you are one.

If, as some think but I do not,¹ ordinary things extend through other possible worlds, then the problem of the many takes on still another dimension. Here in this world we have a ship, the Enigma; there in another world is a ship built at about the same time, to plans that are nearly the same but not quite, using many of the same planks and some that are not the same. It is questionable whether the ship in that other world is Enigma herself, or just a substitute. If Enigma is a thing that extends through worlds, then the question is whether Enigma includes as a part what’s in that other world. We have two versions of Enigma, one that includes this questionable other-worldly

¹ See Lewis (1986a, pp. 210–20).
part and one that excludes it. They have equal claim to count as ships, and equal claim to count as \textit{Enigma}. We have two ships, coinciding in this world but differing in their full extent. Or else we have none; but anyway not the one ship we thought we had.

\textbf{THE PARADOX OF 1001 CATS}

Cat Tibbies is alone on the mat. Tibbies has hairs $h_1, h_2, \ldots, h_{1000}$. Let $c$ be Tibbles including all these hairs; let $c_1$ be all of Tibbles except for $h_1$; and similarly for $c_2, \ldots, c_{1000}$. Each of these $c$'s is a cat. So instead of one cat on the mat, Tibbles, we have at least 1001 cats — which is absurd. This is P. T. Geach's (1980, pp. 215–16) paradox of 1001 cats.

Why should we think that each $c_n$ is a cat? Because, says Geach, `$c_n$ would clearly be a cat were the hair $h_n$ plucked out, and we cannot reasonably suppose that plucking out a hair \textit{generates} a cat, so $c_n$ must already have been a cat' (p. 215). This need not convince us. We can reply that plucking out $h_n$ turns $c_n$ from a mere proper part of cat Tibbles into the whole of a cat. No new cat is generated, since the cat that $c_n$ becomes the whole of is none other than Tibbles. Nor do $c_n$ and Tibbles ever become identical \textit{simpliciter} — of course not, since what’s true about $c_n$'s past still differs from what’s true about Tibbles's past. Rather, $c_n$ becomes the whole of cat Tibbles in the sense that $c_n$'s post-plucking temporal part is identical with Tibbles’s post-plucking temporal part. So far, so good; except for those, like Geach, who reject the idea of temporal parts. The rest of us have no paradox yet.

But suppose it is spring, and Tibbles is shedding. When a cat sheds, the hairs do not come popping off; they become gradually looser, until finally they are held in place only by the hairs around them. By the end of this gradual process, the loose hairs are no longer parts of the cat. Sometime before the end, they are questionable parts: not definitely still parts of the cat, not definitely not. Suppose each of $h_1, h_2, \ldots, h_{1000}$ is at this questionable stage. Now indeed all of $c_1, c_2, \ldots, c_{1000}$, and also $c$ which includes all the questionable hairs, have equal claim to be a cat, and equal claim to be Tibbles. So now we have 1001 cats. (Indeed, we have many more than that. For instance there is the
cat that includes all but the four hairs \(h_6, h_{408}, h_{882},\) and \(h_{907}\).) The paradox of 1001 cats, insofar as it is a real paradox, is another instance of Unger's problem of the many.

To deny that there are many cats on the mat, we must either deny that the many are cats, or else deny that the cats are many. We may solve the paradox by finding a way to disqualify candidates for cathood: there are the many, sure enough, but the many are not all cats. At most one of them is. Perhaps the true cat is one of the many; or perhaps it is something else altogether, and none of the many are cats. Or else, if we grant that all the candidates are truly cats, we must find a way to say that these cats are not truly different from one another. I think both alternatives lead to successful solutions, but we shall see some unsuccessful solutions as well.

**TWO SOLUTIONS BY DISQUALIFICATION: NONE OF THE MANY ARE CATS**

We could try saying that not one of the \(c\)'s is a cat; they are many, sure enough, but not many cats. Tibbles, the only genuine cat on the mat, is something else, different from all of them.

One way to disqualify the many is to invoke the alleged distinction between things and the parcels of matter that constitute them. We could try saying that the \(c\)'s are not cats. Rather, they are cat-constituting parcels of matter. Tibbles is the cat that each of them constitutes.\(^2\)

This dualism of things and their constituters is unparsimonious and unnecessary. It was invented to solve a certain problem, but a better solution to that problem lies elsewhere, as follows. We know that the matter of a thing may exist before and after the thing does; and we know that a thing may gain and lose matter while it still exists, as a cat does, or a wave or a flame. The dualists conclude that the matter is not the thing; constitution is not identity; there are things, there are the parcels of matter that temporarily constitute those things; these are items of two different categories, related by the special relation of

\(^2\) This is the solution advanced in Lowe (1982).
constitution. We must agree, at least, that the temporally extended thing is not the temporally extended parcel of matter that temporarily constitutes that thing. But constitution may be identity, all the same, if it is identity between temporal parts. If some matter constitutes a cat for one minute, then a minute-long temporal segment of the cat is identical to a minute-long temporal segment of the matter. The cat consists entirely of the matter that constitutes it, in this sense: The whole of the cat, throughout the time it lives, consists entirely of temporal segments of various parcels of matter. At any moment, if we disregard everything not located at that moment, the cat and the matter that then constitutes it are identical. So only those who reject the notion of temporal parts have any need for the dualism of things and constituters. But suppose we accept it all the same. At best, this just transforms the paradox of 1001 cats into the paradox of 1001 cat-constituters. Is that an improvement? We all thought there was only one cat on the mat. After distinguishing Tibbies from her constituter, would we not still want to think there was only one cat-constituter on the mat?

Further, even granted that Tibbles has many constituters, I still question whether Tibbles is the only cat present. The constituters are cat-like in size, shape, weight, inner structure, and motion. They vibrate and set the air in motion — in short, they purr (especially when you pat them). Any way a cat can be at a moment, cat-constituters also can be; anything a cat can do at a moment, cat-constituters also can do. They are all too cat-like not to be cats. Indeed, they may have unfeline pasts and futures, but that doesn’t show that they are never cats; it only shows that they do not remain cats for very long. Now we have the paradox of 1002 cats: Tibbles the constituted cat, and also the 1001 all-too-feline cat-constituters. Nothing has been gained.

3 The dualism of things and their constituters is also meant to solve a modal problem: Even at one moment, the thing might have been made of different matter, so what might have been true of it differs from what might have been true of its matter, so constitution cannot be identity. This problem too has a better solution. We should allow that what is true of a given thing at a given world is a vague and inconstant matter. Conflicting answers, equally correct, may be evoked by different ways of referring to the same thing, e.g., as cat or as cat-constituter. My counterpart theory affords this desirable inconstancy; many rival theories do also. See Lewis (1986a, pp. 248–63).
I conclude that invoking the dualism of cats and cat-constituters to solve the paradox of 1001 cats does not succeed.

A different way to disqualify the many appeals to a doctrine of vagueness in nature. We could try saying that cat Tibbles is a vague object, and that the c's are not cats but rather alternative precisifications of a cat.

In one way, at least, this solution works better than the one before. This time, I cannot complain that at best we only transform the paradox of 1001 cats into the paradox of 1001 cat-precisifications, because that is no paradox. If indeed there are vague objects and precisifications, it is only to be expected that one vague object will have many precisifications.

If the proposal is meant to solve our paradox, it must be meant as serious metaphysics. It cannot just be a way of saying 'in the material mode' that the words ‘Tibbles’ and ‘cat’ are vague, and that this vagueness makes it indefinite just which hairs are part of the cat Tibbles. Rather, the idea must be that material objects come in two varieties, vague and precise; cats are vague, the c’s are precise, and that is why none of the c’s is a cat.

This new dualism of vague objects and their precisifications is, again, unparsimonious and unnecessary. The problem it was made to solve might better be solved another way. It is absurd to think that we have decided to apply the name ‘Tibbles’ to a certain precisely delimited object; or that we have decided to apply the term ‘cat’ to each of certain precisely delimited objects. But we needn’t conclude that these words must rather apply to certain imprecisely delimited, vague objects. Instead we should conclude that we never quite made up our minds just what these words apply to. We have made up our minds that ‘Tibbles’ is to name one or another Tibbles-precisification, but we never decided just which one; we decided that ‘cat’ was to apply to some and only some cat-precisifications, but again we never decided just which ones. (Nor did we ever decide just which things our new-found terms ‘Tibbles-precisification’ and ‘cat-precisification’ were to apply to.) It was very sensible of us not to decide. We probably couldn’t have done it if we’d tried; and even if we could have, doing it would have been useless folly.
Semantic indecision will suffice to explain the phenomenon of vagueness. We need no vague objects.

Further, I doubt that I have any correct conception of a vague object. How, for instance, shall I think of an object that is vague in its spatial extent? The closest I can come is to superimpose three pictures. There is the multiplicity picture, in which the vague object gives way to its many precisifications, and the vagueness of the object gives way to differences between precisifications. There is the ignorance picture, in which the object has some definite but secret extent. And there is the fadeaway picture, in which the presence of the object admits of degree, in much the way that the presence of a spot of illumination admits of degree, and the degree diminishes as a function of the distance from the region where the object is most intensely present. None of the three pictures is right. Each one in its own way replaces the alleged vagueness of the object by precision. But if I cannot think of a vague object except by juggling these mistaken pictures, I have no correct conception.

I can complain as before that we end up with a paradox of 1002 cats: Tibbles the vague cat, and also the 1001 precise cats. Once again,

---

4 Provided that there exist the many precisifications for us to be undecided between. If you deny this, you will indeed have need of vague objects. See van Inwagen (1990, pp. 213–83).

5 I grant that the hypothesis of vague objects, for all its faults, can at least be made consistent. If there are vague objects, no doubt they sometimes stand in relations of ‘vague identity’ to one another. We might think that when \( a \) and \( b \) are vaguely identical vague objects, the identity statement \( a = b \) suffers a truth-value gap; but in fact this conception of vague identity belongs to the theory of vagueness as semantic indecision. As Gareth Evans showed, it doesn’t mix with the idea that vague identity is due to vagueness in nature. For if \( a \) and \( b \) are vaguely identical, they differ in respect of vague identity to \( a \); but nothing, however peculiar it may be, differs in any way from itself; so the identity \( a = b \) is definitely false. See Evans (1978). (Evans’s too-concise paper invites misunderstanding, but his own testimony confirms my interpretation. See Lewis 1988.) To get a consistent theory of vague objects, different from the bastard theory that is Evans’s target, we must disconnect ‘vague identity’ from truth-value gaps in identity statements. Even if \( a = b \) is definitely false, \( a \) and \( b \) can still be ‘vaguely identical’ in the sense of sharing some but not all of their precisifications.
the cat-precisifications are all too cat-like. More so than the cat-
constituters, in fact: The precisifications are cat-like not just in what
they can do and how they can be at a moment, but also over time.
They would make good pets — especially since 1001 of them will not
eat you out of house and home!

Don’t say that the precisifications cannot be cats because cats cannot
be precise objects. Surely there could be cats in a world where nature
is so much less gradual that the problem of the many goes away. It
could happen that cats have no questionable parts at all, neither spatial
nor temporal. (In this world, when cats shed in the spring, the hairs
do come popping off.) So it is at least possible that cat-like precise
objects are genuine cats. If so, how can the presence of one vague cat
spoil their cathood?

I conclude that invoking the dualism of vague objects and their
precisifications to solve the paradox of 1001 cats does not succeed.

A BETTER SOLUTION BY DISQUALIFICATION:
ONE OF THE MANY IS A CAT

Since all of the many are so cat-like, there is only one credible way to
deny that all of them are cats. When is something very cat-like, yet
not a cat? — When it is just a little less than a whole cat, almost all of
a cat with just one little bit left out. Or when it is just a little more
than a cat, a cat plus a little something extra. Or when it is both a little
more and a little less.

Suppose we say that one of our many is exactly a cat, no more and
no less; and that each of the rest is disqualified because it is a little less
than a cat, or a little more, or both more and less. This invokes no
unparsimonious and unnecessary dualisms; it disqualifies all but one of
the many without denying that they are very cat-like; it leaves us with
just one cat. All very satisfactory.

The trouble, so it seems, is that there is no saying which one is a
cat. That is left altogether arbitrary. Settling it takes a semantic decision,
and that is the decision we never made (and shouldn’t have made, and
maybe couldn’t have made). No secret fact could answer the question,
for we never decided how the answer would depend on secret facts.
Which one deserves the name ‘cat’ is up to us. If we decline to settle the question, nothing else will settle it for us.\(^6\)

We cannot deny the arbitrariness. What we can deny, though, is that it is trouble. What shall we do, if semantic indecision is inescapable, and yet we wish to carry on talking? The answer, surely, is to exploit the fact that very often our unmade semantic decisions don’t matter. Often, what you want to say will be true under all different ways of making the unmade decision. Then if you say it, even if by choice or by necessity you leave the decision forever unmade, you still speak truthfully. It makes no difference just what you meant, what you say is true regardless. And if it makes no difference just what you meant, likewise it makes no difference that you never made up your mind just what to mean. You say that a famous architect designed Fred’s house; it never crossed your mind to think whether by ‘house’ you meant something that did or that didn’t include the attached garage; neither does some established convention or secret fact decide the issue; no matter, you knew that what you said was true either way.

This plan for coping with semantic indecision is van Fraassen’s (1966) method of *supervaluations*. Call a sentence *super-true* if and only if it is true under all ways of making the unmade semantic decisions; *super-false* if and only if it is false under all ways of making those decisions; and if it is true under some ways and false under others, then it suffers a super-truth-value gap. Super-truth, with respect to a language interpreted in an imperfectly decisive way, replaces truth *simpliciter* as the goal of a cooperative speaker attempting to impart information. We can put it another way: Whatever it is that we do to determine the ‘intended’ interpretation of our language determines not one interpretation but a range of interpretations. (The range depends on context, and is itself somewhat indeterminate.) What we try for, in imparting information, is truth of what we say under all the intended interpretations.

\(^6\) I do not think reference is entirely up to our choice. Some things are by their nature more eligible than others to be referents or objects of thought, and when we do nothing to settle the contest in favour of the less eligible, then the more eligible wins by default; see Lewis (1984). That’s no help here: nature is gradual, no handy joint in nature picks out one of the c’s from all the rest.
Each intended interpretation of our language puts one of the cat-candidates on the mat into the extension of the word ‘cat’, and excludes all the rest. Likewise each intended interpretation picks out one cat-candidate, the same one, as the referent of ‘Tibbles’. Therefore it is super-true that there is just one cat, Tibbles, on the mat. Because it is super-true, you are entitled to affirm it. And so you may say what you want to say: there is one cat. That is how the method of super-valuations solves the paradox of 1001 cats.

**Objection.** Just one of the candidates is a cat, no more and no less. But don’t try to say which one it is. Nothing you might say would be super-true. For it is exactly this semantic decision that remains unmade; it is exactly in this respect that the intended interpretations differ. Although it is super-true that something is a cat on the mat, there is nothing such that it is super-true of it that *it* is a cat on the mat. (It’s like the old puzzle: I owe you a horse, but there’s no horse such that I owe you that horse.) This is peculiar.

**Reply.** So it is. But once you know the reason why, you can learn to accept it.

**Objection.** Supervaluationism works too well: it stops us from ever stating the problem in the first place. The problem supposedly was that all the many candidates had equal claim to cathood. But under the supervaluationist rule, that may not be said. For under any one way of making the unmade decision, one candidate is picked as a cat. So under any one way of making the decision, the candidates do *not* have equal claim. What’s true under all ways of making the decision is super-true. So what’s super-true, and what we should have said, is that the candidates do *not* have equal claim. Then what’s the problem? And yet the problem was stated. So supervaluationism is mistaken.

**Reply.** What’s mistaken is a fanatical supervaluationism, which automatically applies the supervaluationist rule to any statement whatever, never mind that the statement makes no sense that way. The rule

---

7 Here I’m indebted to remarks of Saul Kripke many years ago. At his request, I note that what I have written here may not correspond exactly to the whole of what he said on that occasion.
should instead be taken as a defeasible presumption. What defeats it, sometimes, is the cardinal principle of pragmatics: The right way to take what is said, if at all possible, is the way that makes sense of the message. Since the supervaluationist rule would have made hash of our statement of the problem, straightway the rule was suspended. We are good at making these accommodations; we don't even notice when we do it. Under the supervaluationist rule, it's right to say that there's only one cat, and so the candidates have unequal claim. Suspending the rule, it's right to say that the candidates have equal claim, and that all of them alike are not definitely not cats. Suspending the rule, it's even right to say that they are all cats! Is this capitulation to the paradox? — No; it's no harm to admit that in some sense there are many cats. What's intolerable is to be without any good and natural sense in which there is only one cat.

Objection. The supervaluationist's notion of indeterminate reference is conceptually derivative from the prior notion of reference simpliciter. But if the problem of the many is everywhere, and semantic indecision is inescapable, then reference simpliciter never happens. To the extent that we gain concepts by 'fixing the reference' on actual examples, we are in no position to have the concept of reference. Then neither are we in a position to have the derivative concept of indeterminate reference due to semantic indecision.

Reply. We don't need actual examples to have the concept. We have plenty of imaginary examples of reference simpliciter, uncomplicated by semantic indecision. These examples are set in sharper worlds than ours: worlds where clouds have no outlying droplets, where cats shed their hairs instantaneously, and so on. When we picked up the concept of reference, in childhood, we probably took for granted that our own world was sharp in just that way. (When not puzzling over the problem of the many, maybe we half-believe it still.) We fixed the reference of 'reference' on these imaginary examples in the sharp world we thought we lived in — and if any theory of reference says that cannot be done, so much the worse for it.

8 Here I'm indebted to Andrew Strauss (personal communication, 1989).
I conclude that the supervaluationist solution to the paradox of 1001 cats, and to the problem of the many generally, is successful. But is it the only successful solution? — I think not. I turn now to the other sort of solution: the kind which concedes that the many are cats, but seeks to deny that the cats are really many.

**RELATIVE IDENTITY: THE MANY ARE NOT DIFFERENT CATS**

Geach himself favours one such solution. The paradox of 1001 cats serves as a showcase for his doctrine of relative identity.

Everything falls into place if we realize that the number of cats on the mat is the number of different cats on the mat; and \( c_{13}, c_{279}, \) and \( c \) are not three different cats, they are one and the same cat. Though none of these 1001 lumps of feline tissue is the same lump of feline tissue as another, each is the same cat as any other: each of them, then, is a cat, but there is only one cat on the mat, and our original story stands. . . . The price to pay is that we must regard '——is the same cat as——' as expressing only a certain equivalence relation, not an absolute identity restricted to cats; but this price, I have elsewhere argued, must be paid anyhow, for there is no such absolute identity as logicians have assumed. (1980, p. 216)

'Same cat' is a relation of partial indiscernibility, restricted to respects of comparison somehow associated with the term 'cat', and discernibility by just a few hairs doesn't count. 'Same lump of feline tissue' is a different relation of partial indiscernibility, and a more discerning one.

I agree that sometimes we say 'same', and mean by it not 'absolute identity' but just some relation of partial indiscernibility. I also agree that sometimes we count by relations of partial indiscernibility. As I once wrote:

If an infirm man wishes to know how many roads he must cross to reach his destination, I will count by identity-along-his-path rather than by identity. By crossing the Chester A. Arthur Parkway and Route 137 at the brief stretch where they have merged, he can cross both by crossing only one road. (1976, p. 27)
I'll happily add that for that brief stretch, the two roads are the same. But though I don't object to this positive part of Geach's view, it doesn't ring true to apply it as he does to the case of the cats.

If you ask me to say whether \(e_{13}, e_{279},\) and \(e\) are the same or different, I may indeed be of two minds about how to answer. I might say they're different—after all, I know how they differ! Or I might say they're the same, because the difference is negligible, so I duly ignore it. (Not easy to do while attending to the example as I now am; if I attend to my ignoring of something, \textit{ipso facto} I no longer ignore it.) But if you add the noun phrase, either 'same cat' or 'same lump of feline tissue', it seems to me that I am no less hesitant than before. Just as I was of two minds about 'same', so I am still of two minds about 'same cat' and 'same lump of feline tissue'.

Other cases are different. If you ask me 'same or different?' when you hold Monday's \textit{Melbourne Age} in one hand and Tuesday's \textit{Age} in the other, or when you hold one Monday \textit{Age} in each hand, again I won't know how to answer. But if you ask me 'same or different newspaper?' or 'same or different issue?' or 'same or different copy?' then I'll know just what to say. We can dispute his explanation of what happens, but at least the phenomenon happens exactly as Geach says it does. Not so, I think, for the case of 'same cat' versus 'same lump'.

Something else is lacking in Geach's solution. In other cases where it comes natural to count by a relation other than identity, it seems that identity itself—'absolute identity'—is not far away. Local identity, as between the Arthur Parkway and Route 137 for the stretch where they have merged, is identity \textit{simpliciter} of spatial parts. Likewise temporal identity, as between a thing and the matter that temporarily constitutes it, is identity \textit{simpliciter} of temporal parts. Qualitative identity is identity \textit{simpliciter} of qualitative character. The newspaper that Monday's \textit{Age} is an issue of and the newspaper that Tuesday's \textit{Age} is an issue of are identical \textit{simpliciter}; likewise my copy and your copy of Monday's \textit{Age} are copies of the identical issue. But Geach never tells us what the 'same cat' relation has to do with identity \textit{simpliciter}.

He wouldn't, of course, because he thinks 'there is no such absolute identity as logicians have assumed'. (Nor would he accept all my examples above; certainly not the one about temporary identity and identity of temporal parts.) But Geach's case against absolute identity
is unconvincing. It seems to come down to a challenge: If Geach is
determined to construe all that I say in terms of relations of partial
indiscernibility, is there any way I can stop him? Can I force him to
understand? (What's more, can I do it with one hand tied behind my
back? Can I do it, for instance, without ever using the second-order
quantification that Geach (1967) also challenges?) I suppose not. But
I don't see why that should make me doubt that I know the difference
between identity and indiscernibility.

We have the concept of identity, pace Geach; and if we are to justify
denying that the cats are many, we need to show that they are inter-
related by a relation closely akin to identity itself. Geach has not shown
this, and wouldn't wish to show it. Nevertheless it can be shown, as
we shall soon see. But at that point we shall have a solution that by-
passes Geach's doctrine of relative identity altogether.

PARTIAL IDENTITY: THE MANY ARE ALMOST ONE

What is the opposite of identity? Non-identity, we'd offhand say. Any-
thing is identical to itself; otherwise we have two 'different' things,
two 'distinct' things; that is, two non-identical things. Of course it's
true that things are either identical or non-identical, and never both.
But the real opposite of identity is distinctness: not distinctness in the
sense of non-identity, but rather distinctness in the sense of non-
overlap (what is called 'disjointness' in the jargon of those who reserve
'distinct' to mean 'non-identical'). We have a spectrum of cases. At
one end we find the complete identity of a thing with itself: it and
itself are entirely identical, not at all distinct. At the opposite end we
find the case of two things that are entirely distinct: They have no part
in common. In between we find all the cases of partial overlap: things
with parts in common and other parts not in common. (Sometimes
one of the overlappers is part of the other, sometimes not.) The things
are not entirely identical, not entirely distinct, but some of each. They
are partially identical, partially distinct. There may be more overlap or
less. Some cases are close to the distinctness end of the spectrum: Si-
amese twins who share only a finger are almost completely distinct,
but not quite. Other cases are close to the identity end. For instance,
any two of our cat candidates overlap almost completely. They differ
by only a few hairs. They are not quite completely identical, but they are almost completely identical and very far from completely distinct.

It's strange how philosophers have fixed their attention on one end of the spectrum and forgotten how we ordinarily think of identity and distinctness. You'd think the philosophers of common sense and ordinary language would have set us right long ago, but in fact it was Armstrong (1978, Vol. 2, pp. 37–8) who did the job. Overshadowed though it is by Armstrong's still more noteworthy accomplishments, this service still deserves our attention and gratitude.

Assume our cat-candidates are genuine cats. (Set aside, for now, the supervaluationist solution.) Then, strictly speaking, the cats are many. No two of them are completely identical. But any two of them are almost completely identical; their differences are negligible, as I said before. We have many cats, each one almost identical to all the rest.

Remember how we translate statements of number into the language of identity and quantification. 'There is one cat on the mat' becomes 'For some \( x \), \( x \) is a cat on the mat, and every cat on the mat is identical to \( x \}'. That's false, if we take 'identical' to express the complete and strict identity that lies at the end of the spectrum. But the very extensive overlap of the cats does approximate to complete identity. So what's true is that for some \( x \), \( x \) is a cat on the mat, and every cat on the mat is almost identical to \( x \). In this way, the statement that there is one cat on the mat is almost true. The cats are many, but almost one. By a blameless approximation, we may say simply that there is one cat on the mat. Is that true? – Sometimes we’ll insist on stricter standards, sometimes we’ll be ambivalent, but for most contexts it’s true enough. Thus the idea of partial and approximate identity affords another solution to the paradox of 1001 cats.

The added noun phrase has nothing to do with it. Because of their extensive overlap, the many are almost the same cat; they are almost the same lump of feline tissue; and so on for any other noun phrase that applies to them all. Further, the relation of almost-identity, closely akin to the complete identity that we call identity simpliciter, is not a relation of partial indiscernibility. Of course we can expect almost-identical things to be very similar in a great many ways: size, shape, location, weight, purring, behaviour, not to mention relational properties like location and ownership. But it is hard to think of any very
salient respect in which almost-identical things are guaranteed to be entirely indiscernible. Finally, the relation of almost-identity, in other words extensive overlap, is not in general an equivalence relation. Many steps of almost-identity can take us from one thing to another thing that is entirely distinct from the first. We may hope that almost-identity, when restricted to the many cats as they actually are, will be an equivalence relation; but even that is not entirely guaranteed. It depends on the extent to which the cats differ, and on the threshold for almost-identity (and both of these are matters that we will, very sensibly, leave undecided). What this solution has in common with Geach’s is just that we count the cats by a relation other than strict, ‘absolute’ identity. Beyond that, the theories differ greatly.9

ONE SOLUTION TOO MANY?

We find ourselves with two solutions, and that is one more than we needed. Shall we now choose between the way of supervaluation and the way of partial identity? I think not. We might better combine them. We shall see how each can assist the other.

Here is how to combine them. In the first place, there are two kinds of intended interpretations of our language. Given many almost-identical cat-candidates, some will put every (good enough) candidate into the extension of ‘cat’; others will put exactly one. Context will favour one sort of interpretation or the other, though not every context will settle the matter. Sometimes, especially in our offhand and unphilosophical moments, context will favour the second, one-cat sort of interpretation; and then the supervaluation rule, with nothing to defeat it, will entitle us to say that there is only one cat. But sometimes,

9 There is another way we sometimes count by a relation other than strict identity.

You draw two diagonals in a square; you ask me how many triangles; I say there are four; you deride me for ignoring the four large triangles and counting only the small ones. But the joke is on you. For I was within my rights as a speaker of ordinary language, and you couldn’t see it because you insisted on counting by strict identity. I meant that, for some \( w, x, y, z \), (1) \( w, x, y, z \) are triangles; (2) \( w \) and \( x \) are distinct, and \( x \) and \( y \) and \( z \) (six clauses); and (3) for any triangle \( t \), either \( t \) and \( w \) are not distinct, or \( t \) and \( x \) are distinct (four clauses). And by ‘distinct’ I meant non-overlap rather than non-identity, so what I said was true.
for instance when we have been explicitly attending to the many candidates and noting that they are equally cat-like, context will favour the first, many-cat sort of interpretation. (If we start with one-cat interpretations, and we say things that the supervaluation rule would make hash of, not only is the rule suspended but also the many-cat interpretations come into play.) But even then, we still want some good sense in which there is just one cat (though we may want a way to say the opposite as well). That is what almost-identity offers.

This is one way that almost-identity helps a combined solution. It is still there even when we discuss the paradox of 1001 cats, and we explicitly choose to say that the many are all cats, and we thereby make the supervaluation solution go away.

Perhaps it helps in another way too. The supervaluation rule is more natural in some applications than in others. For instance it seems artificial to apply it to a case of unrelated homonyms. ‘You said you were going to the bank. Is that true? No worries, you bank at the ANZ, it’s right down by the river, so what you said was true either way!’ — I don’t think such a response is utterly forbidden, but it’s peculiar in a way that other applications of the supervaluation rule are not. The two interpretations of ‘bank’ are so different that presumably you did make up your mind which one you meant. So the means for coping with semantic indecision are out of place. The supervaluation rule comes natural only when the alternative interpretations don’t differ too much. If they are one-cat interpretations that differ only by picking almost-identical cats, that’s one way for them not to differ much.

How, on the other hand, do supervaluations help the combined solution? Why not let almost-identity do the whole job?

For one thing, not every case of the problem of the many is like the paradox of 1001 cats. The almost-identity solution won’t always work well.10 We’ve touched on one atypical case already: if not a problem of the many, at least a problem of two. Fred’s house taken as including the garage, and taken as not including the garage, have equal claim to be his house. The claim had better be good enough, else he has no house. So Fred has two houses. No! We’ve already seen how

10 Here I’m indebted to Phillip Bricker (personal communication, 1990).
to solve this problem by the method of supervaluations. (If that seemed
good to you, it shows that the difference between the interpretations
was not yet enough to make the supervaluation rule artificial.) But
although the two house-candidates overlap very substantially, having
all but the garage in common, they do not overlap nearly as exten-
sively as the cats do. Though they are closer to the identity end of the
spectrum than the distinctness end, we cannot really say they’re al-
most identical. So likewise we cannot say that the two houses are al-
most one.

For another thing, take a statement different from the statements of
identity and number that have concerned us so far. Introduce a definite
description: ‘The cat on the mat includes hair \( h_{17} \).’ The obvious re-
sponse to this statement, I suppose, is that it is gappy. It has no definite
truth-value, or no definite super-truth-value, as the case may be. But
how can we get that answer if we decide that all the cat-candidates
are cats, forsake supervaluations, and ask almost-identity to do the
whole job? We might subject the definite description to Russellian
translation:

\[(R1) \text{ There is something that is identical to all and only cats on the mat, and that includes } h_{17}.\]

Or equivalently:

\[(R2) \text{ Something is identical to all and only cats on the mat, and every cat on the mat includes } h_{17}.\]

Both these translations come out false, because nothing is strictly iden-
tical to all and only cats on the mat. That’s not the answer we wanted.
So we might relax ‘identical’ to ‘almost identical’. When we do, the
translations are no longer equivalent: \((R1)\)-relaxed is true, \((R2)\)-relaxed is false. Maybe we’re in a state of semantic indecision between
\((R1)\)-relaxed and \((R2)\)-relaxed; if so, we could apply the supervalua-
tion rule to get the desired gappiness. Or we might apply the supervalua-
tion rule more directly. Different one-cat interpretations pick out
different things as the cat, some that include \( h_{17} \) and some that don’t.
Under any particular one-cat interpretation the Russellian translations
are again equivalent, and different one-cat interpretations give them
different truth values; so the translations, and likewise the original sen-
tence, suffer super-truth-value gaps. Or more simply, different one-cat interpretations differ in the referent of 'the cat'; some of these referents satisfy 'includes $h_{17}$' and some don't, so again we get a super-truth-value gap. Whichever way we go, supervaluations give us the gappiness we want. It's hard to see how else to get it.

REFERENCES


