Causal Efficacy and Externalist Mental Content

by

Anthony E. Newman

B.A. Philosophy
University of California, Berkeley, 1993

B.Phil. Philosophy

SUBMITTED TO THE DEPARTMENT OF LINGUISTICS AND PHILOSOPHY
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY IN PHILOSOPHY
AT THE
MASSACHUSETTS INSTITUTE OF TECHNOLOGY

SEPTEMBER 2002

© 2002 Anthony E. Newman. All rights reserved.

The author hereby grants to MIT permission to reproduce and distribute publicly
paper and electronic copies of this thesis document in whole or in part.

Signature of Author: ...........................................

Department of Linguistics and Philosophy
September 6, 2002

Certified by: ...........................................

Alex Byrne
Professor of Philosophy
Thesis Supervisor

Accepted by: ...........................................

Vann McGee
Professor of Philosophy
Chairman, Committee for Graduate Students
Causal Efficacy and Externalist Mental Content

By

Anthony E. Newman

Submitted to the Department of Linguistics and Philosophy
on September 6, 2002 in Partial Fulfillment of the Requirements
for the Degree of Doctor of Philosophy in Philosophy

ABSTRACT

Internalism about mental content is the view that microphysical duplicates must be mental duplicates as well. This dissertation develops and defends the idea that only a strong version of internalism is compatible with our commonsense commitment to mental causation.

Chapter one defends a novel necessary condition on a property's being causally efficacious — viz., that any property F that is efficacious with respect to event E cannot be instantiated in virtue of any property G that is itself ceteris paribus sufficient for E -- and shows that that necessary condition vindicates the idea that externalism is incompatible with our commonsense commitment to mental causation.

The internalist's core intuition is that only intrinsic properties can be causally efficacious. Chapter two defends that intuition from the common externalist response that extrinsic properties abound.

A popular “Middle Way” between externalism and internalism holds that although ordinary, “folk-psychological” contents of prepositional attitudes are extrinsic, there exists some other non-folk-psychological kind of content that is intrinsic. Chapter three argues that Jerry Fodor’s influential argument for the Middle Way is incoherent.

Chapter four identifies a weak but popular grade of internalism, endorsed by John Searle among others, and argues that it is untenable.

The preceding defense of internalism can be believed only if there is something wrong with the canonical arguments for externalism developed by Hillary Putnam, Tyler Burge, and Saul Kripke. My postscript says what I think is wrong with the canonical externalist arguments: they assume the nonexistence of propositions that are truth-evaluable only relative to particular persons, places, or times; while I argue that our commonsense commitment to mental causation requires at least some such “indexical propositions”.

Thesis Supervisor: Alex Byrne
Title: Professor of Philosophy
Chapter 1:

Externalist Mental Content and the Fight for a Good Cause

O. Introduction

Internalism about mental content holds that any persons who are particle-for-particle indiscernible must be indiscernible in all mental respects as well. For instance: internalists think that anyone molecule-for-molecule indiscernible from someone who wants a drink of water must also want a drink of water -- no matter what linguistic society she belongs to, no matter what her physical surroundings are like, and no matter what prior history she has. Externalism denies this, holding instead that persons can differ from one another mentally thanks solely to occupying relevantly different social, environmental, or historical contexts. A famous externalist claim, for instance, is that only someone who has had sufficient interaction with genuine samples of H₂O can desire water.

Although for centuries internalism was arguably a background presupposition in the commonsense view of the mind, nowadays only a small minority of philosophers believes it. In part this situation is due to the inherent strength of Hilary Putnam’s and Tyler Burge’s groundbreaking “twin” arguments; but I believe it is also due to the fact that one of the best cases for internalism has never been made out as clearly or as convincingly as it could be. I want to fill this gap.

What I will call “the internalist’s causal argument” alleges that externalism is
incompatible with our commonsense commitment to the existence of mental causation: if externalism is true, then many mental properties cannot in fact play the important role in the production of intentional behavior that we believe they often do. Consider, for instance, a person -- let us call her “Thirsty” -- who wants a refreshing drink of water; she then notices a glass full of water in front of her and so she reaches out her arm. A sample of the internalist’s causal argument (which I will soon refine) runs follows:

(1) Thirsty reaches out because she wants to drink some water.
(2) If externalism is true, then (1) is false.
(3) Externalism is false.

This general line of argument has been echoed by many authors;¹ but no acceptable account has ever been given of the causal role that a commonsense claim like premise (1) commits us to seeing mental properties as playing in producing behavior. As a result, externalists have always been able to insist that their view is perfectly compatible with our commitment to mental causation.

In this paper I search for an acceptable account of the causal role that

¹ Jerry Fodor, e.g. in (1987), founds his version of internalism on the premise that science should taxonomize mental states according to their “causal powers”. Stephen Stich (1983) combines the premise that extrinsic “states and processes” are “irrelevant to explaining behavior” (pp. 165-6) with an implicit assumption of externalism, and concludes that scientific psychology should be eliminativist about mental content. Colin McGinn (1989) tips his hat to versions of the argument from Burge, Dennett, Fodor, Loar, and Stich, saying that externalism seems incompatible with the requirement that “mental explanations must cite local causes” (p. 135). Tim Crane (1991), finally, offers a version of the argument that is most like the one I will develop.
commonsense commits us to seeing mental properties as playing in the production of behavior. First I run through two past accounts that are externalist-friendly, and I show that each is deeply flawed. Then I propose a better account and show that it lacks the flaws of its predecessors. Finally, I show, my account turns out to support the internalist’s argument. In the end I hope to have improved our understanding of causation in general and of mental causation in particular; and, as a consequence, to have made clear why externalism must be false if our convictions about the reality of mental causation are true.

But before all of that, I will need to make the internalist’s argument a bit more precise. In the upcoming section I will refine premise (1) by making more explicit our commonsense commitment to mental causation; and in the section after that I will refine premise (2) by making more explicit the content of the externalist’s thesis.

I. Our commitment to mental causation

The gist of the internalist’s causal argument, again, is that externalism is incompatible with our commonsense commitment to mental causation. But what, exactly, are we committed to when we insist, for instance, that Thirsty reached out because she had a desire for water? An obvious -- and undeniably true -- answer is that we believe in psycho-physical causation: mental causes can produce physical effects. On the common assumption that events are the causal relata, this means that mental
events can cause physical ones: the event of Thirsty’s desiring water, for instance, was among the causes of the event of her reaching out.

But commonsense arguably commits us to more than just psycho-physical event-causation. This was first recognized in discussions of Donald Davidson’s “anomalous monism”: in that theory Davidson explains how mental events can cause physical ones, but objectors quickly pointed out that Davidson’s theory is not strong enough to do full justice to our convictions about mental causation. The reason: when we insist that Thirsty’s desiring caused her to reach out, we mean to disallow the joint possibility that (A) the mental event of Thirsty’s desiring was also a physical one, e.g. identical to the instantiating of some particular neural property; and (B) that event caused her reaching-out solely in virtue of having that neural property. We want to believe not just that the event of Thirsty’s desiring water caused her to reach out, but also that it did so at least partly in virtue of being a desiring of water.

Consider an analogy. Nostradamus predicted the King’s demise; so the King’s demise = the demise predicted by Nostradamus. The King’s demise, furthermore, caused the empire to fall, so the demise predicted by Nostradamus caused the empire to fall. Still, we want to say, the demise did not cause the fall in virtue of having been predicted by Nostradamus; rather, that prediction was irrelevant to the fate of the empire. Instead, the demise caused the fall at least partly in virtue of being the demise of a King: had it been only a temporary illness, or the demise of some unknown pauper, the empire would

2 The point was first made by Ted Honderich (1982), p. 64.
have survived. Our convictions about the reality of mental causation require us to see the
mind's role in producing behavior as more like the King and his demise, and less like
Nostradamus and his prediction.

One popular way of expressing this commitment is to endorse one or another
"fine-grained" view of the causal relata -- i.e., any view on which there is no chance of
mistaking any event involving Nostradamus for a cause of the empire's fall and no
chance of mistaking the mental cause of Thirsty's reaching-out for any neural event. For
instance one might distinguish an event that was essentially the demise of a King but only
accidentally predicted by Nostradamus from an event that was only accidentally the
demise of a King but essentially predicted by Nostradamus, and say that the first, but not
the second, caused the empire to fall. Along these lines one would say that an event that
was essentially a desire for water but only accidentally neural was among the causes of
Thirsty's reaching-out. Or alternatively one might identify events with ordered n-tuples
consisting of m objects and one m-place relation, thereby distinguishing <the King,
dying> from <the King, Nostradamus, dying a death that was predicted by>, and say that
the first but not the second caused the empire to fall. Along these lines one would say
that <Thirsty, desiring water> was among the causes of Thirsty's reaching-out. Or as a
third alternative one might take the fine-grained causal relata to be property-instances:
then the property-instance of dying that is located exactly where the King is when he dies,
but not the property-instance of *dying a death predicted by Nostradamus*, caused the property-instance of *falling* that is located exactly where the empire is when it falls.\(^3\)

I do not have any decisive objections to these “fine-grained” views of the causal relata: the internalist’s argument, at least, could be expressed using any of them as a backdrop. Nevertheless I will from hereon set the discussion against a “coarse-grained” backdrop -- in part arbitrarily, but also in part because the fine-grained views all have about them an air of strangeness. The first two versions just mentioned, for instance, sound strange in countenancing more than just one demise in which the King participated: after all, presumably even royalty die only *once*. And the third version sounds strange in positing the existence of something other than the King occupying exactly the same space that he does when he dies: after all, it is natural to think that no two things can be in exactly the same place at the same time.

I will follow much recent literature in relying on a distinction between *causally efficacious* and *causally inefficacious* properties. In these terms, *being the demise of a King* was a causally efficacious property of the King’s demise with respect to the fall of the empire, whereas *having been predicted by Nostradamus* was not. Exactly what it takes for a property to count as causally efficacious will take up the main part of this paper. But for now we can express our commitment to mental causation by saying that even if the mental event of Thirsty’s desiring is also a *physical* event -- e.g., a neural or

\(^3\) For the first version see David Lewis (1986), “Events” and “Causation”; for the second see Jaegwon Kim (1976); for the third see L. A. Paul (2000).
other type of bodily-event -- still its efficacious properties, with respect to Thirsty’s reaching out, include

**THIRST**: *being a wanting to drink some water.*

So the internalist’s causal argument, which I stated in rough form previously, can now be refined as follows:

(1*) THIRST is a causally efficacious property of the event of Thirsty’s wanting water, with respect to her reaching out.

(2*) If externalism is true, then (1*) is false.

(3) Externalism is false.

Our next job is to see why we should believe premise (2*). The basic idea is that the externalist has to identify THIRST with a property that is not causally efficacious with respect to Thirsty’s reaching out; but before we can appreciate this, we have to better understand not only what it takes for a property to be causally efficacious (which task will occupy the main part of this paper) but also what sort of property the externalist takes THIRST to be. The upcoming section will deal with this final bit of stage-setting.
II. The content of externalism

Internalists believe that particle-for-particle indiscernible persons must also be mentally indiscernible, no matter what social, environmental, or historical context either person occupies. One way to report this belief, which will be helpful to us now, is to say that for any event of mental type M, the property being of type M is intrinsic to any event that has it -- in the sense of its instantiation not depending, at least as a matter of metaphysical necessity, on the (intrinsic) properties of anything beyond the spatial boundaries of the event or on the (intrinsic) properties of anything before or after the event's occurrence. Being an ache is plausibly intrinsic in this sense: whether or not a particular neural event is an ache does not seem to depend as a matter of metaphysical necessity on the (intrinsic) properties of anything before that event, after that event, or beyond the spatial boundaries of the brain in which the event occurs. Internalists believe that THIRST, also, is intrinsic in this sense.

A subtle point here, however, must be kept perfectly clear. Namely that, as I will put it, THIRST must be such that to ascribe someone that property amounts to ascribing them a desire for water de dicto rather than de re. To ascribe a desire for water de re is to say that with respect to something that is in fact water, a person desires that thing; while to ascribe a desire for water de dicto is to say that the person desires that he have some water. Thus the subtle point, here, is that THIRST must be such that to ascribe someone that property amounts to ascribing them the desire that they have a drink of water as opposed to ascribing them, with respect to something is in water, the desire that they have
a drink of some. Or in other words: THIRST must be carefully distinguished from the following sort of property:

**DE RE THIRST**: with respect to what is in fact water, being a wanting

to drink some.

The reason we must be careful to distinguish THIRST from DE RE THIRST is that not even internalists think that DE RE THIRST is intrinsic. For it is accepted on all sides that *de re* ascriptions of desires for water might true for one person while false for an intrinsic duplicate: Hilary desires something that is in fact water, for instance, while Twin-Hilary desires something that is in fact XYZ. So Hilary has DE RE THIRST while Twin-Hilary lacks it; though they might nonetheless share all of their intrinsic properties. The issue between internalism and externalism will arise only around *de dicto* ascriptions of desires for water: the internalist believes that Twin-Hilary desires *that he have some water*, just as Hilary desires; and the externalist, as we will next see, believes otherwise.

In fact the externalist may well agree with the internalist that mental events like *aches* are shared among all possible intrinsic duplicates. But she will disagree when it comes to many *contentful* mental events -- such as believings, desirings, and havings of other propositional attitudes. In many of these cases, the externalist believes, the event’s being the sort of mental event it is – i.e., as I will say, its having the *mental-content property* that it does -- is not a purely intrinsic property.
Certainly, the externalist agrees, being an event of some mental type M could never amount to having a purely extrinsic property like being near a burning barn: for an event to have THIRST, for instance, is at least in part for it to be some appropriate type of neural event, functional bodily event, or what-have-you. This much is evident from the fact that events that are not any appropriate type of neural or bodily event -- like explosions of supernovae or fallings of the Dow -- could not be desires for water no matter what social, environmental, or historical context they occurred in. Nevertheless, THIRST cannot be a purely intrinsic property, the externalist believes -- because it is possible for an event in one person’s brain to be a desiring of water even though a second event, occurring in a brain that is molecule-for-molecule indiscernible from the first, is not. An event’s having THIRST at least partly involves its being an appropriate type of neural event, but it also involves occurring in the appropriate social, environmental, or historical context.

For these reasons we might say that THIRST, in the externalist’s eyes, is a “mixed” property, a property containing both intrinsic and extrinsic components -- along the lines of being an ache near a burning barn. So if externalism is true, having THIRST amounts to having the following sort of property --

**EXTERNALIST THIRST (E-THIRST):** intrinsic property $P$ in context $C$

-- where P is a stand-in for whatever intrinsic property an externalist believes events must
possess in order that they would be desirings of water if they occurred in the right sort of context, and C is a stand-in for whatever sort of context an externalist believes is the "right" sort. So P might be a neural property, or a "syntactic" property of some neural code, or a phenomenological property, or a functional property of the brain or even of the entire body; and as for C, Tyler Burge would say that no person can experience a water-desiring if the words she would use to express her desire are not the words so used by the rest of her socio-linguistic community, while Hilary Putnam would say that no person can experience a water-desiring unless she has had sufficient interaction with genuine samples of H2O. The details of these various views will not matter in what follows.

Now not all externalists would be happy identify THIRST with E-THIRST. Timothy Williamson (1998 & 2000), in particular, insists that most mental properties are not, like E-THIRST, "composites" of intrinsic and extrinsic components but instead are "prime" — in that different intrinsic constraints are satisfiable in conjunction with different extrinsic ones. So instead of E-THIRST Williamson would prefer something like

\[ \text{PRIME-THIRST (P-THIRST): } (P_1 \text{ in } C_1) \lor (P_2 \text{ in } C_2) \lor \ldots \lor (P_n \text{ in } C_n) \]

where each \( P_i \) is a different possible intrinsic realizer and each \( C_i \) is a different possible extrinsic realizer of a water-wanting. Stephen Yablo and others have rallied behind Williamson's insistence that the primeness of mental properties helps block the
internalist’s argument: even if properties like THIRST could not be causally efficacious as we want mental-content properties to be, they argue, still properties like P-THIRST could well be.

The argument I will develop against externalism, however, will take on all of these versions together: no matter whether an externalist prefers E-THIRST or P-THIRST, and no matter how she wants to fill in either of those schemas, she will end up with a causally inefficacious property. Hence, with all of this stage-setting in place, I can now state the internalist’s causal argument in its final, most perspicuous form:

(1*) THIRST was a causally efficacious property of the event of Thirsty’s wanting water, with respect to her reaching out.

(2*₁) If externalism is true, then THIRST is identical to E-THIRST (or P-THIRST).

(2*₂) E-THIRST was not (nor was P-THIRST) a causally efficacious property of the event of Thirsty’s wanting water, with respect to her reaching out.

(3) Externalism is false.

As I previously explained, I will take (1*) as exemplary of a basic commitment to mental causation and (2*₁) as following from the content of the externalist’s thesis. The internalist’s argument, then, rests on the plausibility of (2*₂). The rest of this paper is
devoted to showing that if we think clearly about the notion of causal efficacy, we can see -- despite what externalists have argued in the past -- that premise (2*2) is true.

III. The intuitive idea of causal efficacy

The intuitive idea of a causally efficacious property is the idea of a property that plays an important role in the production of some effect: being the demise of a King, for instance, plays an important role in the fall of the empire and so is among the efficacious properties of the King's demise; its having been predicted by Nostradamus plays no such important role and so is not causally efficacious.

Following a long tradition in the philosophy of causation, most analyses of the intuitive idea of causal efficacy just described agree that it can be made more precise in modal terms. In other words: to tell which of an event's properties are causally efficacious with respect to a particular effect, and which of its properties are not causally efficacious, we need only look at what would happen if things were different in certain crucial ways from how they actually are. Some philosophers disagree with this broadly modal approach to analyses of the central notions of causality; but I agree with the long tradition that it seems to be the most promising approach.

---


5 Namely "singularism" about causation, traceable to Ducasse; and the regularity theory, made famous by Hume. (Cf. Causation, Sosa & Tooley (eds.), Introduction, p. 17).
In particular, most analyses of the intuitive idea of causal efficacy described above agree that it can be harmlessly, if not entirely helpfully, captured in something like the following way.\textsuperscript{6}

\textbf{INTUITIVE ANALYSIS:} A property $F$ is causally efficacious with respect to event $E$ only if $F$ is instantiated by a cause of $E$ and

\textbf{I-1.} F's being so instantiated is \textit{ceteris paribus necessary} for $E$'s occurrence

\textbf{I-2.} F's being so instantiated is \textit{ceteris paribus sufficient} for $E$'s occurrence.

For instance: it is in a natural sense \textit{c.p. necessary} for the fall of the empire that \textit{being the demise of a King} be instantiated; since the empire would not fall, other things equal, if the King did not die. Likewise, it is in a natural sense \textit{c.p. sufficient} that that property be instantiated; since the King's dying is \textit{all it takes}, given the relevant circumstances, to cause the empire to fall. In contrast, it is in no sense either c.p. necessary or c.p. sufficient that \textit{being a demise predicted by Nostradamus} be instantiated: on the one hand the empire would fall even if Nostradamus did not predict the King's demise.

demise; and on the other hand Nostradamus’ prediction could easily turn out false.

The intuitive analysis, however, is still not precise enough for our purposes. The reason is that it contains “ceteris paribus” qualifiers: it does not say that efficacious properties are *strictly* necessary and sufficient for their effects, but rather only necessary and sufficient “other things equal” or “given the relevant circumstances”. This is a good thing: for the empire could still fall even if no event occurred that were the demise of a King, e.g. if the plague broke out or if aliens invaded. And the King *could* die without causing the empire to fall, e.g. if no one ever found out or if the King’s twin brother immediately took over. But the intuitive analysis does not tell us precisely why those possibilities do not count against the causal efficacy of *being the demise of a King*.

In the next two sections I will consider two influential proposals of how to distinguish the possibilities that count against the causal efficacy of a property and the possibilities that do not -- i.e., proposals about the precise content of the “c.p.”-qualifiers in the intuitive analysis of causal efficacy. As we will see, both proposals have been claimed by externalists to undermine the internalist’s argument.

### IV. The counterfactual analysis

Thanks no doubt to the wide influence of David Lewis’s counterfactual analysis of causation, the most common way of fleshing out the idea of “c.p. necessity”, and

---

7 Without its “c.p.” qualifiers, MICE would be subject to the sorts of objection considered by Jean Kazez (1995, p. 87) and Martha Klein (1996, p. 169; cf. also her mention of John Heil & Alfred Mele (1991)).
COUNTERFACTUAL ANALYSIS: Property F is causally efficacious with respect to event E if and only if F is instantiated by a cause of E and

C-1. E fails to occur in the closest possible world in which F was not so instantiated.

For instance: being the demise of a King is a causally efficacious property of the King’s demise, with respect to the fall of the empire, because the closest possible world in no event is the demise of a King – e.g., a world in which the King takes only temporarily ill – is a world in which the empire survives.

On the other hand having been predicted by Nostradamus is not an efficacious property of the King’s demise, with respect to the fall of the empire, because at least one possible world in which the King’s demise is not predicted by Nostradamus but in which the empire still falls is less of a departure from actuality than any world that lacks both Nostradamus’ prediction and the fall of the empire. One such world, for instance, is the world which differs from actuality only in lacking Nostradamus’ prediction.

If the counterfactual analysis adequately captures our intuitive idea of causal efficacy, then as many externalists have pointed out, the internalist’s causal argument is

---

largely unconvincing.\footnote{E.g. Lepore & Loewer (1987); Heil & Mele (1991).}

Internalists, remember, want to deny that the instantiation of E-THIRST (\textit{viz.} $P$ in $C$) by Thirsty’s desiring-event is c.p. necessary for her reaching-out; and their reason is that Thirsty’s desiring does not need to have E-THIRST’s second component (\textit{viz. occurring in} $C$): if Thirsty were in a relevantly similar situation on Twin-Earth, and so saw XYZ instead of water, she would instantiate all of the same intrinsic properties that she actually does and so would still reach out. Externalists who endorse a counterfactual analysis of causal efficacy, however, point out that such a Twin-Earth world is not the closest world in which Thirsty’s desiring-event fails to occur in context C: on the contrary, the closest such world, they claim, is one in which there is nothing around that resembles water \textit{at all}, and so is one in which Thirsty sees no reason to reach-out. So the counterfactual analysis might not disqualify E-THIRST from being causally efficacious.

\section*{V. The problem with counterfactuals}

I find it doubtful that the counterfactual analysis, even if it were the correct analysis of causal efficacy, would in fact allow E-THIRST to count as causally efficacious.

First of all, I do not see how some world in which Thirsty \textit{neither} has intrinsic property P \textit{nor} is in context C could turn out to be closer to actuality than the world that
differs from actuality *only* by her lacking P. To put it intuitively, any world of the former sort differs in two ways from the actual world; while the latter differs in only one. And by similar reasoning, I do not see how some world in which Thirsty neither has P nor is in C could be closer to actuality than the world that differs from actuality *only* by her not being in C. If these claims are true, then the closest possible world in which Thirsty lacks E-THIRST must either be one that differs from actuality *only* by her not having P (call that "W1"), or be one that differs from actuality *only* by her not being in C (call that "W2"). Since, then, Thirsty reaches out in W2 but not in W1; and since the counterfactual analysis will save externalism only if Thirsty fails to reach out in the closest world in which she lacks E-THIRST, the counterfactual analysis will save externalism only if W1 is closer to actuality than W2 is.

But is it so clear that, for instance, changing all of the relevant facts about Thirsty's brain must make for less of a radical departure from reality than simply changing some facts about Thirsty's past interaction with samples of water? Perhaps only a slight change in the world would remove Thirsty from the relevant context, whereas it would take deep and dramatic changes in her neural structure to remove her from the relevant brain-state. If that is the case, then the closest world in which Thirsty lacks E-THIRST will be one in which she still reaches out; so the counterfactual analysis will not block the internalist's argument.

But I want to let this difficulty pass, in order to focus on a more basic failing of the counterfactual analysis. In particular, the counterfactual analysis is far too weak,
counting many properties as causally efficacious that intuitively are not so. The point can be made with two examples.

First, suppose that Socrates had been committed to doing everything in life with unbounded zeal; so whenever he drank, he insisted on guzzling. In the closest possible world in which Socrates did not guzzle his fatal hemlock, then, he did not drink it either (for a world in which he drank lackadaisically must be a world in which he was psychologically very different); so in that world he does not die such an early death. So the counterfactual analysis does not disqualify

GUZZLE: being a guzzling of hemlock

from among the efficacious properties of Socrates’ act of drinking, with respect to his dying such an early death; but in any reasonable sense, that act’s having

DRINK: being a drinking of hemlock

was all that was c.p. necessary in the way our intuitive idea of causal efficacy requires. After all: if Socrates had merely sipped the hemlock, he would have died just as early a death.

10 The example comes from Stephen Yablo (1992), p. 276.
The counterfactualist might be able to block this particular example: if an ordering of possible worlds can be defended on which it is less of a departure from actuality for Socrates to have drunk lackadaisically than it is for him to have abstained from the hemlock altogether, then the closest non-guzzling world will be one in which Socrates still drank the hemlock; and so it will be one in which he dies and the counterfactual analysis will correctly rule GUZZLE inefficacious. But this would only be an idiosyncrasy of the particular example: consider, for instance, being a drinking of hemlock in a universe containing at least 100 stars. Intuitively, again, this property was not an efficacious property of Socrates’ action, since the number of stars was entirely irrelevant to his death; but equally clearly, if his action had lacked that property then he wouldn’t have died. (The reason: the closest possible world in which Socrates’ action was not a drinking of hemlock in a universe containing at least 100 stars is undoubtedly one in which he did not drink, since we would have to go much further out in the realm of possibility to get rid of a billion-plus stars.) It is easy enough, I think, to see that some problematic example can be found along these lines.

For a second example that brings out the extreme weakness of the counterfactual analysis, consider the property

**DISJUNCTION #1 (D#1):** being either a drinking of hemlock or a handstand.

D#1 was presumably not an efficacious property of Socrates’ drinking-event, with respect
to his early death -- since the instantiation of D#1 was not c.p. sufficient as our intuitive idea of causal efficacy would require. (To see this, note that D#1 would still have been instantiated if Socrates had done a handstand instead of drinking hemlock; and doing a handstand is not ordinarily life-threatening.) But the closest possible world in which D#1 was not instantiated must be a world in which Socrates did not drink hemlock, so if D#1 had not been instantiated Socrates would not have died such an early death; so once again the counterfactual analysis fails to exclude an obviously inefficacious property from the realm of causal efficacy.

This second example capitalizes on the glaring lack of any analogue of our "intuitive" clause I-2 -- which clause requires efficacious properties to be c.p. sufficient for their effects. Accordingly, a natural idea is that the counterfactual analysis could overcome the current problem if supplemented with its own version such a clause. Unfortunately for the externalist, I have seen no way of adding such a clause that is consistent with the basic counterfactual approach.

The most straightforward way of working something like the intuitive analysis’s clause I-2 into the counterfactual analysis is to add the requirement that

C-2. E occurs in the closest possible world in which F is so instantiated.

Nick Zangwill, for instance, tries to capture the idea that efficacious properties should be c.p. sufficient for an effect by requiring what he calls a "factual conditional": "if the
cause had possessed them, then, in the circumstances, the effect would have occurred". And on the standard Lewisian semantics for counterfactuals, this requirement amounts precisely to C-2.\textsuperscript{11}

This straightforward fix adds nothing, unfortunately, to the counterfactual analysis -- because a property F can be efficacious with respect to an effect E only if it is actually instantiated and E actually occurs, in which case the closest possible world in which F is instantiated will always be the actual world. D#1, for instance, is not disqualified from the realm of efficacy by C-2, because the closest world in which Socrates' drinking-event has it is the actual world and Socrates actually dies an early death.

Nor will it help if we modify C-2 to focus on the closest non-actual possible world, because any such world will most likely differ from actuality in a wholly irrelevant and insignificant way: for instance the closest non-actual possible world in which Socrates' drinking-event has D#1 will differ from actuality only with respect to something like the position of a single subatomic particle, which difference is nowhere near great enough to stop him from dying.

Stephen Yablo notices that C-2 is "trivial" for the reasons just given, and so suggests a modification along the following lines:

\textsuperscript{11} Zangwill (1996), p. 74. As becomes clear in Zangwill's discussion, he does not assume the standard semantics for counterfactuals; but neither does he specify any clear alternative.
C-2'.  *Relative to the closest world in which F was not* so instantiated,

E occurs in the closest possible world in which F *is* so instantiated.\(^{12}\)

In other words: out of all the possible worlds in which F is instantiated, E occurs in the one closest to the closest possible world in which F is *not* instantiated.

This condition is not quite as trivial as C-2, as it disqualifies from the realm of causal efficacy some properties that C-2 wrongly lets in;\(^{13}\) but still it is not strong enough to disqualify properties like D#1. For the closest world in which Socrates neither drinks hemlock nor does a handstand is (again) one exactly like the actual world except for the fact that Socrates does not drink hemlock there; and relative to that world, the closest world in which he either drinks hemlock or does a handstand is presumably one in which (again) he drinks hemlock (for a world in which Socrates performs acrobatics seems much more distant than that). In sum: relative to the closest world in which Socrates’ action lacks D#1, the closest world in which it *has* that property is one in which he dies an early death; so Yablo’s fix to the original counterfactual analysis is no more helpful than Zangwill’s.


\(^{13}\) If a bridge that can support only 20 tons collapses beneath a truck that weighs 40 tons, then the truck’s *weighing over 10 tons*, although c.p. necessary, was arguably not efficacious because it was not c.p. sufficient (an 11-ton truck would have crossed without incident). C-2’, but not C-2, bears this out.
The upshot of the Socrates-examples is that the counterfactual analysis has captured, at best, only an extremely unrestrictive necessary condition on causal efficacy. The test is far too easy to pass, so it hardly matters that E-THIRST passes it: we have been given no reason to believe that E-THIRST will remain in the running for causal efficacy after the intuitive analysis of causal efficacy is fully fleshed-out.

VI. The determinable analysis

The above problems for the counterfactual analysis of causal efficacy make obvious two requirements for any adequate analysis. First, to keep properties like being either a drinking of hemlock or a handstand (i.e., D#1) from being counted as causally efficacious, an adequate analysis will have to require that efficacious properties be c.p. sufficient for effects of any cause that instantiates them, in addition to requiring that they be c.p. necessary. And second, to correctly judge D#1 as c.p. insufficient for Socrates’ early death, as well as to correctly judge GUZZLE as c.p. unnecessary, an adequate analysis will have to care about more than just the closest possible worlds in which those properties fail to be instantiated.

The all-important question is whether the internalist’s argument will survive once we find a principle that correctly distinguishes the relevant possible worlds from the irrelevant ones. Several authors have noticed the need for such a principle, but have failed to offer an analysis of causal efficacy that is any improvement in precision over our
original intuitive analysis.\textsuperscript{14} Stephen Yablo, however, has offered an analysis that is precise enough to do the job; furthermore he concludes that the internalist’s argument fails.

Yablo’s discussion makes use of a distinction between “determinates” and “determinables”: \textit{red}, for instance, is a “determinable” of its “determinate” \textit{scarlet}, because being scarlet is a \textit{particular way} of being red; you cannot be scarlet without being red). Following Yablo in putting off any detailed analysis of that relation, his suggestion is as follows:\textsuperscript{15}

\textbf{DETERMINABLE ANALYSIS:} A property $F$ is efficacious with respect to event $E$ if and only if $F$ is instantiated by a cause of $E$ and

\begin{itemize}
\item[D-1.] for every determinable $F_{\text{able}}$ of $F$, $E$ does not occur in the closest possible world in which $F_{\text{able}}$ but not $F$ is so instantiated; and
\item[D-2.] for every determinate $F_{\text{ate}}$ of $F$, $E$ occurs in the closest possible world in which $F$ but not $F_{\text{ate}}$ is so instantiated.
\end{itemize}

\textsuperscript{14} Terry Horgan (1989) and Jean Kazez (1995) both recognize the requirements just mentioned.

\textsuperscript{15} Yablo (1992) and (1997). Again, Yablo’s version is in terms of \textit{causation} instead of \textit{causal efficacy}. 
This analysis is not as complex as it looks. Most importantly, for our purposes, note that a property is not efficacious, with respect to some particular effect, if either the instantiation of some determinable of it would suffice for that effect or else the instantiation of some determinate of it is necessary. If the instantiation of some determinable of the property would suffice, then the property is not c.p. necessary; and if the instantiation of some determinate of it is necessary, then the property is not c.p. sufficient.

Yablo's proposal successfully overcomes the problems I raised against the counterfactual analysis. On the one hand DRINK is a determinable of GUZZLE, since you cannot guzzle without drinking, and if Socrates' action had simply had DRINK he would still have died; so clause D-1 correctly disqualifies GUZZLE from being counted as causally efficacious.

On the other hand DRINK is a determinate of D#1, and Socrates would not have died such an early death if his action had had D#1 but not DRINK (e.g., if it had been a handstand instead of a drinking of hemlock); so clause D-2, this time, correctly disqualifies D#1 from being counted as causally efficacious.

It is significantly trickier to understand why Yablo thinks his determinable analysis blocks the internalist's argument -- so much trickier, in fact, that I will avoid debating the issue here. Instead, let us simply assume that Yablo is right that if the

---

16 Primarily because E-THIRST straightforwardly fails the determinable analysis: P is a determinable of E-THIRST, and its instantiation would presumably have sufficed for Thirsty's reaching-out in the nearest non-watery context. Yablo's response, as I understand it, is to identify Thirsty's thirst with a different
determinable analysis were accepted, then premise \((2^*_2)\) of the internalist’s causal argument would turn out false. As I will next argue, however, the internalist still need not worry -- because the determinable analysis has irreparable problems.

**VII. The problem with determinables**

To bring out the central problem with Yablo’s determinable analysis of causal efficacy, notice that even though Socrates might still have died without drinking hemlock -- for instance if he had instead jumped from the edge of a tall cliff, or had run out of air while scuba-diving to the bottom of the ocean, or (etc.) -- we still want to say that his action’s having DRINK was *in fact* causally efficacious with respect to his early death.

The determinable analysis, unfortunately, cannot support this intuition; for consider the following property:

---

extrinsic property, namely something like what I previously called “P-THIRST”: \((P_1\ in \ C_1) \lor (P_2\ in \ C_2) \lor \ldots \lor (P_n\ in \ C_n)\).

Now that property also fails the determinable analysis, because \(P_1 \lor P_2 \lor \ldots \lor P_n\) is a determinable of it and would have been all Thirsty needed in order to reach out. But Yablo’s next move is to argue that this does not vindicate the internalist because on the one hand that huge disjunctive property is the only intrinsic property of Thirsty’s that passes the determinable analysis (having \(P_1\), for instance, would presumably have sufficed for Thirsty’s reaching-out but is not a determinable of Yablo’s candidate) and hence is the only one internalists can identify with Thirsty’s thirst; and on the other hand, according to Yablo internalists should not be happy about identifying any “genuine” property with such an “overly disjunctive” one.

As this exegesis brings out, Yablo’s defense of externalism is far from straightforward. One might wonder, for instance, why the internalist should fear identifying Thirsty’s thirst with the disjunctive property \(P_1 \lor P_2 \lor \ldots \lor P_n\). The biggest mystery to my mind, however, is how externalism is supposed to benefit even if the determinable analysis tells against internalism: the correct conclusion, it would seem to me, is not that externalism avoids the internalist’s objection but rather that internalism succumbs as well -- in which case the determinable analysis leads directly to epiphenomenalism.
**DISJUNCTION #2 (D#2):** being either a drinking of hemlock or a jumping from a cliff or a running-out-of-air-while-scuba-diving or ... (etc.).

On the one hand D#2 is a determinable of DRINK (you cannot drink hemlock without *either* drinking hemlock or jumping from a cliff or ... (etc.) And on the other hand Socrates would still have died an early death if an action of his had instantiated D#2 instead of DRINK -- i.e. he *does* die an early death in the closest possible world in which he did not drink hemlock but instead either jumped from a cliff or ... (etc.). So clause D-1 disqualifies DRINK from being counted as a causally efficacious property of Socrates’ drinking.

And the problem is not specific to this example: the determinable analysis of causal efficacy always reserves causal efficacy for *disjunctions containing all possible sufficient causes* of an effect, instead of granting it to whichever *particular disjunct* is in fact instantiated. So hardly any familiar properties turn out to be efficacious; and that is unacceptable.

In some works currently in progress, Yablo has been developing a modification of the determinable analysis that is intended to avoid the sort of problem just raised. 17 Perhaps the determinable analysis can indeed be saved; any such modification will

---

17 In particular, Yablo has suggested that we should count a property inefficacious only when a determinable of it that is “at least as natural” would have sufficed for the relevant effect. The claim would then be that D#2 is quite a bit *less* natural than *drinking hemlock* and so does not matter when testing the latter for causal efficacy.
deserve a serious look once it is clearly laid out. But in the meantime, we have good reason to be suspicious of any determinable-based approach and so to look elsewhere in our analysis of causal efficacy. In fact, as I will show in the upcoming section, a necessary condition on causal efficacy that boasts all of the benefits of the determinable analysis can be constructed without relying on the notion of determination and so without succumbing to the above problem in the first place.

**VIII. The reducer condition**

A crucial job performed by the determinable analysis of causal efficacy was that it kept GUZZLE from being counted as a causally efficacious property of Socrates’ drinking, with respect to his early death -- since the event’s having DRINK would have sufficed. However, it did this job by requiring that no determinables of an efficacious property should suffice for its effects (DRINK is a determinable of GUZZLE), and that requirement in turn kept DRINK from being counted as efficacious -- since the instantiation of D#2 would have sufficed, and D#2 is a determinable of DRINK.

There is, however, a relation distinct from determination in which GUZZLE stands to DRINK and D#2 stands to DRINK as well, *instead of the other way around* -- and so which can be relied on to explain our intuition that *neither GUZZLE nor D#2 was a causally efficacious property of Socrates’ act of drinking, with respect to his death, while DRINK was*.
The relation I have in mind is one I will call "reduction": there is a natural sense in which the property *red and round*, for instance, as well as the property *red or round*, is reducible both to *red* and to *round*. Let me cash out this "natural sense" in a number of more-or-less standard ways, then explain how the notion of reduction can help us in the case of Thirsty.

To be *red and round*, first of all, is just to be *red* on the one hand and to be *round* on the other; being red and round is "nothing over and above" being red and being round. Similarly, to be *either red or round* is either to be red or else to be round (or both): being red or round is "nothing over and above" either being red or being round. Or again: God needed only to create two properties -- *redness* on the one hand and *roundness* on the other -- in order for there to exist a conjunctive property containing both as conjuncts and a disjunctive property containing both as disjuncts. Or yet again: in order to understand what it takes to be *red and round* -- and similarly, to understand what it takes to be *red or round* -- one need only understand what it takes to be red and what it takes to be round, along with some basic principles of logic.

Or finally, in terms that I find most illuminating: something that is *red and round* is so *partly in virtue of* being *red* and *partly in virtue of* being *round* -- and in more than just the causal or nomological sense a billiard-ball might move "in virtue of" being hit by me. In other words any such thing is red and round partly *because* it is red and partly *because* it is round, in a metaphysical sense in which it is *not* true to say that a billiard-ball moved "because" I hit it. Similarly something that is *red or round* is so *either* in
virtue of being *red* or in virtue of being *round*, i.e. either *because* it is *red* or *because* it is *round*; and again in some deep sense that is less causal or nomological than metaphysical.

In all of these senses, GUZZLE is reducible to DRINK but DRINK is *not* reducible to D#2 (in fact it is the other way around, D#2 reducible to DRINK): any guzzling-event is such at least partly in virtue of being a drinking (in particular in virtue of being an energetic or enthusiastic drinking); but there is no similar sense in which any drinking-event is such at least partly in virtue of being either a drinking or a jumping from a bridge or ... (etc.)

Note that the reducibility of one property to any others (call those other properties "reducers" of the former) cannot always be read off from the predicates used to denote those properties: *red* & *scarlet*, for instance, has neither *red* nor *scarlet* as reducers, because it is identical to the property *scarlet* and nothing that is scarlet is so either in virtue of being scarlet or in virtue of being red, in the intended sense. Whether one property is reducible to another is not a matter of how we conceive of or denote those properties, but is rather a matter of the objective relations which hold or fail to hold between them. Granted, I have not provided anything like an analysis of this notion of reduction. But it is a fairly natural notion, and should be judged by the theoretical

---

18 The following seems to me a promising start: property F is a reducer of property G iff either
   (1) anything that has G does so at least partly in virtue of having F, or
   (2) anything that has F has G at least partly in virtue of having F, or
   (3) F is a reducer of some reducer of G.

19 Jaegwon Kim (1984, 1990) gives an overview of a similar-sounding notion he calls "dependence".
fruits that it brings.

To that end, let me finally define putting one property "in place of" another as adding the former after removing the latter -- or less metaphorically, define the closest possible world in which F is instantiated "in place of" G as the F-world closest to the closest non-G world. I then propose the following necessary condition on causal efficacy:

**REDUCER CONDITION:** A property F is causally efficacious with respect to event E only if F is instantiated by a cause of E and there is no reducer F_R of F such that E occurs in the closest possible world in which F_R is instantiated in place of F.

In other words: a property is not efficacious with respect to a particular effect if one of its reducers would suffice.

The reducer condition is not a complete analysis of causal efficacy, since it does not include a sufficient condition for a property's being causally efficacious. Nevertheless, it reaps all clear benefits of the analyses so far mentioned, and more.

Unlike the counterfactual analysis, first of all, the reducer condition avoids counting GUZZLE as a causally efficacious property of Socrates' act of drinking, with respect to his early death -- since DRINK is a reducer of GUZZLE and was all Socrates' action needed to instantiated in order to cause his death. (To see this, go to the closest
world in which Socrates *merely sipped* his hemlock, and note that he still dies in that world.)

Also unlike the counterfactual analysis, the reducer condition avoids counting D#1 as efficacious -- this time because *D#1* has DRINK as a reducer, and again DRINK would have sufficed for Socrates' death. (To see this, go first to the closest world in which there was *neither* a drinking of hemlock *nor* a handstand, then go from there to the closest world in which there was once again a drinking of hemlock; and note that in that last world Socrates once again dies.)

And furthermore, unlike the *determinable* analysis the reducer condition allows DRINK to be counted as an efficacious property of Socrates' drinking, by avoiding counting D#2 as efficacious -- in this case, because any reducers of D#2 would have sufficed. (To see this, go first to the closest world in which D#2 failed to be instantiated - i.e., in which there was *neither* a drinking of hemlock, *nor* a jumping from a cliff, *nor* ... (etc.) -- then go from there to the closest world in which any one of those disjuncts was instantiated; and note that Socrates still dies there.)

Finally comes the payoff I've been promising: the reducer condition supports the internalist's premise *(2*₂), the claim that externalists cannot see THIRST as an efficacious property of Thirsty’s thirsting-event, with respect to her reaching out. For replacing E-THIRST with its intrinsic reducer \( P \) would have sufficed. The reason: Thirsty's thirsting-event need only to have intrinsic property \( P \) in order to cause her to reach out; it does not need also to occur in context C.
Furthermore, as I promised earlier, even if those like Williamson and Yablo are right in preferring to identify THIRST with something like P-THIRST instead of with E-THIRST, the reducer condition tells against externalism. For both \((P_1 \lor P_2 \lor \ldots \lor P_n)\) and the \(P_i\) that is actually instantiated are reducers of P-THIRST, and either is all that is needed to cause Thirsty’s reaching out.

Note that both Yablo and Lewisian-style counterfactualists accept these last points, namely that replacing the externalist version of THIRST with P would still lead Thirsty to reach out. That point is uncontroversial: only the counterfactualists are not bothered by it because the closest world in which Thirsty’s thirsting lacks E-THIRST (they claim) is not a world in which it is replaced with P; and Yablo is not bothered because P is not a determinable of P-THIRST. But now that we have seen the flaws in the counterfactual and determinable analyses of causal efficacy, and have furthermore seen that what matters to a property’s status as efficacious or inefficacious is whether the relevant effects occur in the closest possible worlds in which its reducers are instantiated in its place, it is clear that externalists should be bothered by the fact that all Thirsty’s thirsting needed, in order to cause her reaching out, was to have P.

Our best understanding of the notion of causal efficacy vindicates the internalist’s causal argument against externalism.
IX. Problems and prospects

The necessary condition on causal efficacy I proposed above, the "reducer" condition, avoids the problems of past analyses. But it also captures a deep and pervasive intuition: it is common to hear, for instance, that dispositional or functional properties such as being provocative, being a mousetrap, or having dormitive virtue cannot be causally efficacious; and the reducer condition bears this out.

According to the reducer condition it can only be a red cape's redness, and not its provocativeness, that is responsible for its provoking a bull, since the cape is provocative in virtue of being red; it can only be a mousetrap's particular construction, and not its "mousetrapness", that is responsible for catching a mouse, since the mousetrap is a mousetrap in virtue of its particular construction; and it can only be a sleeping-pill's chemical constitution, and not its dormitivity, that is responsible for putting a man to sleep, since the pill has dormitive virtue in virtue of its chemical constitution. In short, the reducer condition shows itself to be anything but ad-hoc by according with the intuition which Cynthia and Graham Macdonald express as follows:

[causally efficacious properties] must be capable of exercising their causal powers through their instances independently of others, even ones upon which they supervene. Their force must not be what the emergentists would have called a mere resultant force, one exhausted by the forces of the base ... properties.

---


Accordance with this intuition does not mean, of course, that the reducer condition is uncontroversial. In fact I want to end by briefly mentioning one controversial aspect -- namely, that the reducer condition disqualified from the realm of causal efficacy many properties that are commonly mentioned in seemingly respectable causal explanations of events. For instance: it would not be strange to hear that a bridge collapsed beneath a passing truck because the truck weighed over 20 tons; but the reducer condition will judge that it was the precise weight of the truck, rather than its weighing over 20 tons, that was causally efficacious.\(^{22}\) If the mention of a property in respectable causal explanations means that that property can be causally efficacious, then much of our ordinary explanatory practice comes into conflict with the reducer condition.

This point must be taken seriously by the defender of the reducer condition, and hence by any internalist who relies on the reducer condition in her "causal" argument. But I also think that the point can be made significantly less threatening by noting -- as several authors have done -- that \textit{perfectly respectable causal explanations can mention properties that are not causally efficacious}.\(^{23}\) In particular, a causally inefficacious property might prove helpful to mention in a causal explanation for either of two


First, mentioning a causally inefficacious property in an explanation of some effect might help delimit the range of properties whose causal efficacy with respect to that effect is epistemically possible, by implicitly ruling certain properties out: pointing out that the bridge collapsed under a truck which weighed over 20 tons, for instance, might implicitly inform the listener that the truck’s color, shape, and so on were entirely unimportant to the causal process. And second, mentioning a causally inefficacious property might provide valuable information about which properties could have been causally efficacious if the situation had been only slightly different: pointing out that the truck weighed over 20 tons, again, might inform the listener that for any n > 20, if the truck had weighed precisely n tons then its weighing n tons would have been efficacious in the bridge’s collapse. So the common mention of properties like weighing over 20 tons in perfectly respectable causal explanations need not entail that such properties are causally efficacious, and hence need not undermine the reducer condition.

This defense of the reducer condition, however, leads to one final problem: namely, that the internalist who relies on the reducer condition in her argument against externalism now has more work to do to show that mental properties of mental events are ever causally efficacious. If perfectly respectable causal explanations may in fact mention causally inefficacious properties, as the above defense of the reducer condition insisted, then the internalist will have to give us good reason for believing in particular that being a desiring of water (i.e. THIRST) was indeed a causally efficacious property of

---

Thirsty’s thirsting (as the internalist’s first premise requires) and *not just helpful to mention* in a causal explanation of Thirsty’s reaching-out.

I will leave this task -- showing that THIRST was indeed causally efficacious, and not just helpful to mention in a causal explanation -- as further work for the internalist to do. But I feel confident that it can be done: for on the one hand causally efficacious properties are ones that play a *metaphysically* important role in causal transactions, while properties can be *helpful to mention in causal explanations* due solely to the *epistemological* role they play in our explanatory practice; and on the other hand the role played by THIRST with respect to Thirsty’s reaching-out was decidedly metaphysical, not merely epistemological. Let me end by clarifying my confidence in this matter.

Explaining to you that Socrates died because he *either drank hemlock or jumped off a cliff*, just like explaining that the building collapsed because of an earthquake that registered over 5 Richter, might prove informative in certain imaginable situations -- for instance if, for all you knew, Socrates died while scuba-diving (my explanation rules that possibility out); or if you did not know that jumping off cliffs is a life-threatening activity (my explanation informs you that it is). But this seems to be a fact *more about you* than about the causal transaction between the glass of hemlock and Socrates: to see this, notice that *if you already* knew that Socrates did not die from anything other than drinking hemlock or jumping off a cliff, and *if you did* know that jumping off cliffs is a life-threatening activity, then my explanation would rightly seem less-than-helpful. In particular you would have every right to ask “Well, which was it? Did Socrates die...
because he drank hemlock, or because he jumped off a cliff?"

In contrast, explaining that Socrates died because he drank hemlock seems informative in more than just those peculiar sorts of circumstance, to more than just that peculiarly-ignorant sort of listener. Mentioning that Socrates' action had DRINK, that is, does not merely rule some possible properties out, leaving a set of properties each of which could have been efficacious if things had been relevantly different, but rather pinpoints a single property of that action which was actually efficacious. (To see that this is the case, note that although my explanation would be "less-than-helpful" to someone who already knew that Socrates died from drinking hemlock, such a person need not feel that my explanation was in any sense underspecific.) The instantiation of THIRST, finally, surely played a role in the production of Thirsty's reaching-out that is more like the role played by DRINK with respect to Socrates' early death and less like the role played by D#1. Mentioning THIRST does not seem to provide information about different possible properties of Thirsty's mental event each of which could have been efficacious with respect to her reaching-out; instead, just like mentioning DRINK when explaining Socrates' death, mentioning the instantiation of THIRST informs us of one of the actual reasons why things happened as they did.

Accordingly, it seems to me, the internalist's premise 1* is completely justified in assuming THIRST to have been causally efficacious in producing -- and so not just helpful to mention in a causal explanation of -- Thirsty's reaching-out. Again, the internalist will need to make out this line of response in greater detail than I have done
here; but the underlying idea, I think, seems clearly promising.

**X. Conclusion**

A well-known "causal" argument for internalism is that mental properties are often causally efficacious in the production of intentional behavior; but the sorts of properties an externalist would have to identify with certain mental properties cannot be causally efficacious; so externalism must be false (§I, §II).

If we follow common opinion and accept an analysis of the intuitive idea of causal efficacy that parallels David Lewis's counterfactual analysis of causation, or if we instead follow Stephen Yablo in preferring one of the few solid alternative analyses, the internalist's causal argument might well fail. Unfortunately for the externalist, however, both the counterfactual analysis and Yablo's determinable analysis give the wrong results for many intuitively obvious cases (§III - §VII).

Problems with past analyses of causal efficacy can be overcome, I suggested, if we make it a requirement that none of the properties in virtue of which an efficacious property comes to be instantiated would suffice for the efficacious property's effects. On this lone acceptable analysis of causal efficacy, mental-content properties as the externalist construes them are seen to be inefficacious -- so the internalist's causal argument goes through (§VIII).

In addition to being less problematic than its competitors, this last analysis of causal efficacy fits in nicely with deep and pervasive intuitions about the nature of
causation. Granted, if the internalist hopes to use that analysis to support her causal argument against externalism then she has work to do to show that mental properties should in fact be seen as causally efficacious. But I see no reason to doubt, and indeed every reason to expect, that the internalist can finish that further bit of business (§IX).

As things stand, then, my conclusion is twofold. First, we must make it a requirement on the notion of causal efficacy that no “reducers” of an efficacious property suffice for that property’s effects. And second, externalism about mental content must be false, given that our thoughts can affect how we behave.
Chapter 2: "The Burning-Barn Fallacy ..."

The Burning-Barn Fallacy in Defenses of Externalism About Mental Content

O. Introduction

It must be stopped. It is an extremely common response to one powerful argument against externalism about mental content, and it contains a simple but insidious fallacy.

By "externalism about mental content" I mean the view that many perfectly ordinary content-properties of mental events consist at least partly in relations to entities in the mental subject's social, environmental, or historical context. For a belief to have the content that Aristotle was wise, for instance, is at least partly for it to occur at the end of a causal chain in which Aristotle himself participated in suitable ways.

What I will call "the internalist's causal argument" contends that externalism is incompatible with our commonsense commitment to mental causation. If externalism is true, the argument runs, then many content-properties of mental events cannot play the important causal role in the production of intentional behavior that we ordinarily think

---

1 The title alludes to a passage, whose source and author both escape me, that defends the causal efficacy of being 5 miles from a burning barn with exactly the line of reasoning I will criticize.
they play.² The reason: if externalism is true then many content-properties of mental events must consist at least partly in relations to entities in the mental subjects’ surrounding social, environmental, or historical context; and no such extrinsic, relational properties can play the important causal role in the production of intentional behavior that we ordinarily think content-properties play.

Most externalists insist that extrinsic, relational properties of events can play whatever causal role content-properties of mental events play in the production of intentional behavior. Their most common justification of their insistence, furthermore, calls attention to one or another ordinary causal process and contends that it is simply obvious that extrinsic, relational properties are playing the desired causal role in that example. Let me dub that causal role “causal efficacy”: the internalist’s causal argument contends that extrinsic properties cannot be causally efficacious; while the common externalist response insists that many ordinary causal processes exhibit extrinsic properties that are obviously causally efficacious.

In this paper I will criticize that externalist response to the internalist’s argument. One might first of all notice that that response can show only that at least some extrinsic properties are causally efficacious, while in order for externalism to be compatible with mental causation externalists need a very particular kind of extrinsic property to be

² The line of argument is defended by Jerry Fodor, e.g. (1987); Stephen Stich (1983; though Stich assumes externalism and so uses the argument to support eliminativism about content); Colin McGinn (1989; though McGinn argues against externalism only with respect to certain types of mental state); and Tim Crane (1991).
causally efficacious -- since they need to see mental-content properties as being of that very particular kind. And furthermore, in the previous chapter I argued that the particular kind of extrinsic property that externalists need to identify with mental-content properties cannot be causally efficacious. But in this paper I want to push a more immediate objection.

In particular, I will argue that the above common externalist response does not even show that some extrinsic properties play the important causal role that we ordinarily think content-properties play. On the one hand, no examples of ordinary causal processes need be taken as showing that extrinsic properties are what I will call "causally efficacious"; and furthermore, the only obvious reason to think that they do is to endorse an unacceptable principle concerning the sufficient conditions for causal efficacy. Internalists might be mistaken in believing that extrinsic relations cannot be causally efficacious; but the most common externalist argument that they are so mistaken should not move us at all.

I. The basic idea

To illustrate the externalist line of reasoning I want to criticize, and also to make clear just how common it is, I have picked from the literature four representative
externalist responses to the idea that externalism is incompatible with our commonsense commitment to mental causation.³

(1)
Let us see if we can show how broad content could be causally efficacious. . . .

It seems to me that . . . worlds in which causal-historical properties obtain without the brain states they actually cause are distant from the actual world. That is, in a nearby world in which a person did not have the causal history they did, their brain would not have been in the state it is. . . . Looking at far-flung worlds in which [brain state] Px obtains without [its causal history] Rx is irrelevant to the question of whether the mental state Mx which is determined by the conjunction of Px and Rx is efficacious with respect to the later behavioural state P'y. . . .

When philosophers argue that the causal history of a brain is causally ineffectual with respect to bodily behaviour, it is like arguing that someone's firing a gun was not causally efficacious in causing a death because all that really caused the death was the bullet entering the body. To parody: . . . the bullet entering the body would have caused the death just the same even if the bullet had not come from the barrel of a gun. Surely then, firing the gun was epiphenomenal with respect to the ensuing death. This would be a novel argument for those who defend the constitutional right to bear arms in the U.S.!

The reply to this argument is that near worlds in which the bullet entered the body are worlds in which the shot was fired from the gun. So firing the gun did cause the death. . . . It is the same with externalist mental content. Just as firing guns . . . cause deaths, so the causal history which partly determines mental content plays a causal role with respect to behaviour. (Nick Zangwill, 1996, pp. 93-4.)

(2)
[S] suppose switch S is wired to light L in such a way that one can turn on the light by closing the switch. . . . I now close the switch, causing the light to go on . . .

Think, now, of explaining . . . the fact that closure of the switch causes the light to go on. . . . An answer to this . . . question will take us backward in time, to the processes and circumstances that shaped the control circuits of the device whose behavior is in question. There may be a variety of different possible explanations of why the switch is (now) controlling the lights . . . . Whatever the correct answer, it seems clear that in trying to understand, in this

particular way, why the light came on, we are seeking to find out something about how the device came to be in its present condition, a condition in which one event causes another.

... Were I able to convince you that this is the correct way of thinking about behavior, I would thereby convince you that an event's extrinsic properties -- properties it has because of the way it is related, historically and otherwise, to other objects and events -- are, or at least may be, relevant to explaining the behavior of the system in which those events occur. If, for example, A is an event in the brain of an animal, and B is a bodily movement resulting (in part) from A, then to explain the animal's behavior ... one can, indeed one may have to, advert to precisely those historical and relational facts about A on which its meaning supervenes. This, in turn, would mean that even though meaning does not supervene on the intrinsic physical properties of an event -- even though, in other words, the psychological states, the beliefs and desires, of an organism do not supervene on the biological stuff of which it is composed -- nonetheless, meaning, and therefore what the organism believes and desires, can figure in the explanation of its behavior. (Fred Dretske, 1990, pp. 10-12)

(3)

A non-proximate cause can be causally relevant to an effect if it is connected to that effect via a chain of intermediate causes. Even if the truth of externalism prevents content from being local, it would not thereby prevent it from being causally relevant. It would do so if externalism implied that no part of what individuated content is spatially located, but it does not imply this. ...

We can put the [internalist's] argument like this: what makes a state contentful cannot be what makes it causal, because what makes it contentful is its relation to the objects it represents, or is about, while what makes it causal is something intrinsic or internal to it. The having of the aboutness property depends on the relation between thinker and object. This relation is causally irrelevant; it is only the features which are intrinsic to the thinker or his states which can be plausibly cited as candidates for causal efficacy.

This is open to the same objection .... Why couldn't the intrinsic causally-productive features depend on the relationship between thinker and object? It is acknowledged that ... a causal relationship between thinker and object is necessary for it to be true that the thinker's state (say, his belief) is about that object. So since a causal relationship is necessary for the aboutness relation, it makes sense to suppose that it is necessary for the existence of the belief state. Had it not been for the existence of the object and its impact on the thinker, there would not have been the belief state in question. Thus far from being causally irrelevant, the relationship is essential. (Martha Klein, 1996, pp. 166-8.)

[S]omeone who actually saw a dagger and someone who only hallucinated one might both say: 'Is that a dagger?' And the [internalist's] objection [is] that it is
not in virtue of the relationship definitive of contentfulness that the behaviour occurs. . . .

The [internalist's objection] relies on . . . dubious assumptions . . . . Even if we were to admit that in the hallucination case, there is not a contentful state, but only the appearance of one . . . this would not rule out the possibility that in the non-hallucinatory case there was a contentful state and that it was in virtue of its content that the thinker said what he did . . . .

Surely, the fact that there are possible worlds in which thinkers hallucinate daggers does not show that in the actual world; that is, in the actual situation where there was a dagger, its presence was not causally responsible for the thinker's belief that it was present, and thus his subsequent behaviour. (ibid, pp. 168-9)

(4)

One is thirsty; how likely is one to be drinking soon? Likely enough, if one sees water. Much less likely, if what one sees is a mirage. Even if drinking is individuated narrowly, its explanation in terms of the earlier state of the system involves the presence of water in the environment, not just the earlier internal physical state of the agent. Concepts of broad mental conditions give us a better understanding of connections between present states and actions in the non-immediate future, because the connections involve interaction with the environment. (Timothy Williamson, 1998, pp. 395-6)

As these passages illustrate, the conflict between externalism and mental causation gets discussed in a wide variety of terminological frameworks: some authors focus on whether externalist mental events can be causes; some on whether an externalistic psychology can endorse adequate causal explanations; some on whether externalist mental properties can be causally relevant; and so on. But all of the above passages share at least one of their central aims: namely, to draw an analogy between mental causation as it must be if externalism is true, on the one hand, and many perfectly ordinary causal processes on the other. Since shootings can cause deaths (Zangwill's example), flippings of switches can cause lights to turn on (Dretske), seeings of daggers can cause askings of "Is that a dagger?" (Klein), and seeings of water can cause drinkings
(Williamson), then -- fill in the missing premises -- the internalist must be wrong in claiming externalism to be incompatible with mental causation.

To evaluate that line of argument, obviously we need to get clear about the "missing premises". And in order to do this, we have to first get on the table a clear version of the argument to which the externalists are responding. I take on those two tasks in the upcoming two sections.

II. The internalist’s causal argument

Imagine a guy named “Saul” who believes that Aristotle was wise, wants to read a wise man’s writings, and so signs up to take a course in ancient Greek philosophy. A sample of the internalist’s causal argument can be put as follows:

1. If externalism is true, then having the content that Aristotle was wise is not an intrinsic property of the event of Saul’s believing that Aristotle was wise.

2. Having the content that Aristotle was wise is a causally efficacious property of Saul’s believing-event, with respect to his signing-up.

3. Only intrinsic properties can be causally efficacious.

4. Externalism is false.
This version of the causal argument relies on two distinctions between types of properties: the distinction between intrinsic and extrinsic properties, and the distinction between causally efficacious and causally inefficacious properties. So let me say a few words about each.

Intrinsic properties are properties that do not depend for their instantiation, in a thing at a time, on the intrinsic properties of anything external to that thing or at other times. More helpfully, though importing a bit of harmless (for our purposes here) physicalism, intrinsic properties can be thought of as properties that supervene on the properties of that thing that are describable in the vocabulary of an idealized microphysics: in a slogan, microphysical duplicates are intrinsic duplicates.

The internalist’s premise 1, then, says that if externalism is true, microphysical duplicates of a mental event that has the content that Aristotle was wise can lack that content. A possible externalist claim, for instance, is that a microphysical duplicate of Saul who did not grow up on earth and hence has never been causally impacted by Aristotle cannot share Saul’s belief that Aristotle was wise. Some believing-event, no doubt, occurs in Twin-Saul’s brain; but that event can at best have the content that some other person was wise, or that a philosopher of such-and-such description was wise; but it cannot share the content of Saul’s belief.

Causally efficacious properties, next, are those properties whose instantiation accounts for a cause’s having at least some of the causal powers it does. In other words, efficacious properties account for a cause’s being capable of causing the effects it does:
given the efficacious properties a cause actually has, it would cause such-and-such effects under such-and-such conditions. Assuming, as I will, that events are the causal relata, causally efficacious properties do not themselves cause anything; rather, efficacious properties are those properties of events the having of which accounts for those events' being capable of causing the events they do. Causally inefficacious properties, on the other hand, are properties whose instantiation does not affect a cause's powers. In other words, they are properties that do not make a difference to which effects an event would cause under any given set of conditions.

For instance: a sonic boom shatters a nearby window. The boom's volume is causally efficacious with respect to the shattering, because it is in virtue of having that volume that the boom is capable of shattering any sufficiently fragile window in its proximity. Given it's actual volume, the boom would shatter a window of the appropriate thickness if sufficiently nearby. And hence, since a window of appropriate thickness is nearby, the boom does shatter it. Being pondered by Kripke, on the other hand, is an inefficacious property of the sonic boom with respect to the shattering; because the boom would be capable of causing the shattering whether or not it were being pondered by Kripke.

The internalist's premise 2, then, says that it is at least partly in virtue of having the content that Aristotle was wise that Saul's belief is capable of causing him to sign up for a philosophy course; or again, that in virtue of having that content Saul's belief is such that it would cause him to sign up for a philosophy course under the right
conditions. In fact we know at least one set of conditions that is “right”. Namely, the conditions that actually obtain: including Saul’s wanting to read a wise man’s writings; there being a philosophy course in the offing; and so forth. Because Saul’s belief has the content that it does, and because those other conditions obtain, Saul’s belief does cause him to sign up.

The internalist’s premise 3, then, says that properties that are not necessarily shared among events that are indiscernible from the perspective of microphysics do not contribute to the causal powers of their owners. Sometimes expressed in the rather misleading slogan “causation is local”, premise 3 is rarely argued for. And if the common externalist response exemplified in the above-quoted passages hopes to block the internalist’s causal argument, it must be seen as attempting to refute premise 3.

III. The externalist’s response

Premise 3 is the most obvious target, in the internalist’s causal argument as I have construed it, of the common externalist response exemplified in the above-quoted passages. In particular, I take that response as attempting to show, by use of ordinary examples of causal processes, that extrinsic, relational properties can play whatever causal role we take mental-content properties to play in the production of intentional behavior.
In this terminology Zangwill's example, for instance, is meant to show that at least some extrinsic properties of the event of the bullet flying through the air are causally efficacious in its causing the man's death. In particular, among that event's causally efficacious properties is presumably an extrinsic property consisting of some relation to the past event of the firing of a gun – e.g., a property like

FIRED: \textit{being such that a gun was fired.}

Since the flying-bullet event would not have caused the man's death if no gun had ever been fired – presumably, since if the bullet had simply been thrown by hand it would not have had adequate velocity to kill the man -- we are supposed to conclude that FIRED is among the causally efficacious properties of the flying-bullet event with respect to the man's death.

Dretske's example, next, is meant to show that among the properties of the switch-flipping that are efficacious in its causing the light to turn on is some extrinsic property like

WIRED: \textit{being such that the control-circuit was designed in so-and-so a way.}

Since the switch-flipping event would not have caused the light to turn on if the control circuit had not been constructed as it was, we are supposed to conclude that WIRED is
among the causally efficacious properties of the switch-flipping event with respect to the light’s turning on.

Klein’s example, similarly, is meant to show that among the properties of the seeing of the dagger that are efficacious in its causing an asking of “Is that a dagger?” is some extrinsic property like

\[
\text{DAGGER: being compresent with a dagger.}
\]

Since Klein’s visual experience does not cause her to ask “Is that a dagger?” in the relevant set of possible worlds in which no dagger is present – for in those worlds, presumably, her visual experience is not as of a dagger -- we are supposed to conclude that DAGGER is among the causally efficacious properties of her visual experience with respect to her asking.

Williamson’s example, finally, is meant to show that among the properties of his visual experience that are efficacious in its causing him to drink is some extrinsic property like

\[
\text{WATER: being compresent with some visible water.}
\]

Since Williamson’s visual experience is far less likely to cause him to drink if no water is present and causing that experience in a reliable and continued way – for in that case he
will most likely have, or soon have, no reason to think that drinking is a live option, and so he will soon stop trying -- we are supposed to conclude that WATER is among the causally efficacious properties of the thirsting-event with respect to the drinking.

All of the above lines of reasoning can be viewed as having a common structure, which I will summarize as follows:

Ext-1. C causes E.

Ext-2. D is required for C’s causing E.

Ext-3. Being such that D occurs is a causally efficacious property of C with respect to E.

For Zangwill, C = the event of the bullet flying through the air; E= the man’s death; D = the firing of the gun; and the notion that I am calling “requirement” is cashed out in terms of counterfactual dependence.

For Dretske, C = the flipping of the light switch; E = the lights’ turning-on; D = the previous construction of the control circuit; and again “requirement” means counterfactual dependence.

For Klein, C = the visual experience as of a dagger; E = asking “Is that a dagger?”; D = the presence of a dagger; and the notion of “requirement” is something like covariation in a relevant set of possible worlds.
For Williamson, finally, $C =$ a visual experience as of water; $E =$ a drinking; $D =$ the nearby presence of some visible water; and the notion of "requirement" is cashed out in terms of conditional probability.

Perhaps some of the authors whose passages I have suggested interpreting as attempting to refute premise 3 of the internalist’s causal argument will claim to be attempting no such thing. Again, terminological variation is rampant in the literature; and accordingly, some authors might claim to be showing only that externalist mental events can be causes of intentional behavior; or that externalist mental content can be causally relevant; or that an externalistic psychology can endorse adequate causal explanations; or (etc.).

Some or all of these claims might be true: perhaps externalist mental events (whatever they are) can be causes of intentional behavior; perhaps externalist mental content can be causally relevant (whatever that means); perhaps an externalistic psychology can endorse adequate causal explanations (whatever those are); and so forth. But unless the falsity of premise 3 of the internalist’s argument as I construed it follows from one of these claims, none of them will suffice to block that argument. So if any of the above-quoted externalist passages hope to block that argument – which I firmly believe their authors would hope them to – then they must be seen as attempting to show that extrinsic, relational properties can be causally efficacious.

To be charitable to the externalists who have responded to the internalist’s causal argument, then, I will continue to speak of a “common externalist response” exemplified
in the above-quoted passages and to see that response as endorsing one or another instance of the argument from EXT-1 to EXT-3. Next I will argue that that line of argument is at best unconvincing and at worst fallacious. At the very least we need not take any properties like FIRED, WIRED, DAGGER, or WATER to be causally efficacious; and worse yet, the only obvious reason we would have do so would be if we endorsed an unacceptable principle concerning the sufficient conditions for causal efficacy.

IV. Objection #1: internalist-friendly redescription

I am not going to question Zangwill’s assumption that the firing of the gun is in some deep sense required for the flying-bullet event to cause the man’s death. Nor will I question Dretske’s assumption that the actual way the control circuit was constructed is required for the switch-flipping to cause the light’s turning-on; or Klein’s assumption that the presence of the dagger is required for her visual experience to cause her to ask “Is that a dagger?”; or Williamson’s assumption that the nearby presence of visible water is required for his visual experience to cause him to eventually drink.

None of that, however, means I must accept that any of FIRED, WIRED, DAGGER, or WATER are causally efficacious; and the first way to see this is to note that all of the causal processes in question can be redescribed in a wholly internalist-friendly way. In other words, I can give a perfectly complete causal explanation of the
man's death, of the light turning on, of the asking of "Is that a dagger?", and of the drinking without mentioning any of the cause's extrinsic properties such as FIRED, WIRED, DAGGER, or WATER.

In the Zangwill example, I point out simply that the firing of the gun causes the bullet to fly through the air, and that the event of the bullet flying through the air causes the man to die; but insist that both of those causes cause their effects solely in virtue of their intrinsic properties. Certainly, the bullet has the velocity it does in virtue of being shot from a gun, for if it had simply been thrown it would have been traveling much slower; but that reflects only the fact that the firing of the gun is a cause of the bullet's flying through the air with the velocity it in fact has. The only properties of the event of the bullet flying through the air that we need countenance as causally efficacious, with respect to the man's death, are intrinsic properties like involving a bullet traveling with such-and-such velocity. Obviously Dretske's, Klein's, Williamson's, and any other such examples can be redescribed in a similar internalist-friendly way.

For Dretske: the construction of the control circuit causes the light switch to be as it is, and the flipping of the switch turns on the light; but both solely in virtue of their intrinsic properties. The construction of the control circuit is a cause of the switch's being as it is and hence is also a cause of the light's turning-on, but WIRED is not among the causally efficacious properties of the switch-flipping with respect to that turning-on.

For Klein: the presence of the dagger causes a visual experience as of a dagger, which in turn causes an asking of "Is that a dagger?"; but both solely in virtue of their
intrinsic properties. The presence of the dagger is a *cause* of the visual experience and hence also of the asking, but DAGGER is not among the causally efficacious properties of the visual experience with respect to the asking.

For Williamson: the nearby presence of visible water ensures that the visual experience as of water will also continue, uninterrupted; the visual experience combines with the thirst to cause a drinking; but both solely in virtue of their intrinsic properties. The presence of visible water is an ongoing *cause* of the ongoing visual experience as of water, and hence also of the drinking, but WATER is not among the causally efficacious properties of the visual experience, with respect to the drinking.

Christopher Peacocke anticipates this type of "redescription" objection and argues that it is no good. According to Peacocke, internalists will encounter the following dilemma when attempting to causally explain any intentional behavior with respect to which externalists claim an extrinsic property is causally efficacious.

On the one hand, if the internalist mentions only the intrinsic properties of the behaving subject’s mental events, then the internalist’s explanation of the subject’s behavior will be crucially incomplete. For instance: any complete explanation of the behavior of Williamson’s water-drinker must mention the nearby presence of water; for the presence of the water is in some suitably deep sense required for the drinking. 4 On the other hand, Peacocke argues, to admit this is to admit that an internalist cannot give a complete causal explanation of such behavior. In his own words:

---

4 Peacocke does not use the Williamson example *per se*, but one relevantly similar.
The first competing suggestion [to the externalist's] would be that if we supplement a suitable specification of an internalist state with a statement of selected features of the circumstances surrounding the subject, we will obtain something which does explain the holding of certain relations in the external world. 

But as we add more and more relational specifications to the original explanation involving only an internalist state, the question arises of whether we are not in fact adding sufficient material to ensure that the subject is, after all, in an externalist state. ... If [the supplementary material plus the internal facts] are sufficient, the internal state plus supplementation is hardly an alternative to the view that externalist states are explanatory. The externalist state to which it was proposed as an alternative explanation actually supervenes on the explanatory resources employed in the alleged alternative. (Peacocke, 1993, pp. 209-9.)

In short: either the internalist cannot give a complete causal explanation of Williamson’s drinking, or else the internalist is, after all, conceding everything the externalist wants.

I am willing to grant the first horn of Peacocke’s dilemma: I am willing to accept that an internalist will have to mention the presence of water in order to give a suitably complete causal explanation of Williamson’s drinking. My internalist-friendly redescription of the example, in fact, mentioned the presence of water as among the causes of Williamson’s drinking.

But as I argued above, that does not mean that the internalist must mention any extrinsic properties of any of the causes of Williamson’s drinking: in particular, the internalist need not mention the fact that Williamson’s visual experience has WATER. So I think the Peacocke poses a false dilemma: if I reject the first horn, I need not accept the second since by mentioning the presence of water the internalist is not conceding everything the externalist wants. In particular, Peacocke’s argument that the internalist is
so conceding is doubly unconvincing: in one part that argument is misleading, and in another part it is straightforwardly false.

Peacocke misleads, first of all, when he says “the question arises of whether we are not in fact adding sufficient material to ensure that the subject is, after all, in an externalist state” -- because this is distinctly not the question at stake in the debate as I've construed it. The relevant “externalist state” in the Williamson example, for instance, is the property WATER, viz. being compresent with some nearby visible water; but the mere fact that Williamson’s visual experience has that property is no vindication of externalism at all. It is ridiculous to think that the internalist must deny that there is any water around: if that were true then internalism would be a lost cause. For it is granted in the description of the example that there is water around.

Peacocke stops misleading, and gets to the genuine issue, when he goes on to say “the internal state plus supplementation is hardly an alternative to the view that externalist states are explanatory”. For what is at issue is not, as we have just seen, whether Williamson’s visual experience has WATER; instead it is whether that property is causally efficacious with respect to his eventual drinking. The internalist claims it is required for the drinking but is not efficacious; Peacocke, instead, claims that saying that WATER is “required” is “hardly an alternative” to saying that it is causally efficacious. According to Peacocke, to admit that more than just the intrinsic properties of Williamson’s visual experience are required for that experience to cause the drinking is
just to admit that some extrinsic properties are among that visual experience’s causally efficacious properties.

But once Peacocke paints the issue correctly, it seems to me that what he says is plainly false: there certainly seems to be a difference between a story according to which WATER is a causally efficacious property of Williamson’s visual experience, with respect to his drinking, and a story according to which no extrinsic properties are causally efficacious although the nearby presence of water is required as a contributing cause of the drinking. At the very least there is on the face of it a plain difference between these two types of story. And I will next argue that it would be a mistake to doubt first appearances.

V. Objection #2: required properties

The above objection is enough to take most of the wind out of the common externalist response to the internalist’s argument I am considering. For that response relies on exhibiting examples of extrinsic properties that are alleged to be obviously causally efficacious; but internalist-friendly redescriptions of those examples sound perfectly natural and, hence, if they are unacceptable their unacceptability is far from obvious. What the externalist needs is a reasoned argument to the effect that we should indeed count the examples in question as examples of efficacious extrinsics; but that has never been done.
Perhaps, though, such an argument has simply seemed too trivial to mention. Perhaps the inference from EXT-1 to EXT-3 licensed by some obvious principle that externalists have simply not wanted to waste paper and time spelling out. In particular, my insistence that none of FIRED, WIRED, DAGGER, or WATER is causally efficacious, and my response to Peacocke's objection, rested on my rejecting the following principle:

Efficacy of Required Properties (ERP): If the occurrence of event D is required for event C to cause event E, then being such that D occurs is among the causally efficacious properties of C, with respect to E.

So perhaps externalists have tacitly assumed something like ERP in endorsing the common response I am considering.

Just here, however, prospects for the externalist's response begin to look even dimmer. For ERP judges as causally efficacious many properties that are inefficacious on any intuitive understanding of that notion.

For instance, ERP entails that one of the causally efficacious properties of Zangwill's firing of the gun, with respect to the man's death, is its being such that a bullet later enters the man's body – for certainly the firing would not have killed the man if no bullet had later entered his body. In fact that same property, being such that a bullet
later enters the man’s body, is among the efficacious properties of both the Big Bang and the birth of the man’s grandmother; for neither the Big Bang nor the birth of the man’s grandmother would have been among the causes of the man’s (actual) death if no bullet had later entered his body. And finally, among the efficacious properties of the flying-bullet event, with respect to the man’s death, is both its *being such that the Big Bang occurred* and its *being such that the man’s grandmother was born*. For plausibly, the man would never have been born if either the Big Bang had not occurred or his grandmother had never been born; in which case the flying bullet could not have killed him.

Besides sounding mildly crazy, there are two small reasons and one big reason that these consequences are unacceptable. The first small reason is that intuitively, efficacious properties of any cause of some particular effect should distinguish that cause from at least some other causes of that same effect. *Involving a bullet traveling with such-and-such velocity*, for instance, is something that the flying-bullet event contributes to the causal process leading up to the man’s death: perhaps some other links in the causal chain leading up to that death might have shared that property (e.g., a shooting-event that killed the gangster who would have killed the man’s great-great-great-grandmother); but most of them don’t, and none of them must.

This reflects our intuition that when we ask about a thing’s causally efficacious properties, we are asking what *it in particular contributes* to the causal process. What properties of the event of the flying bullet, we ask, were responsible for the man’s
eventually dying? If ERP is true, however, then far from being contributed by and distinguishing the events that have them, properties like *being such that the Big Bang occurred*, *being such that the man's grandmother was born*, and *being such that a bullet eventually enters the man's body* are properties that are shared by each and every link in the causal chain leading up to the death.

The second small reason is that if ERP were true, it would be much more difficult to identify an event's efficacious properties than one might have previously thought. For to do so one would have to first identify all other causes of all of that event's effects. One would not be able to know, for instance, which properties of the flying bullet contributed to its killing the man until one gained knowledge of the entire causal chain leading from the Big Bang up to that death. In fact, which efficacious properties an event possessed would depend on what it caused; and these points not only sound strange, but seem to get things dead backwards.

This last point hints at the big reason that ERP is unacceptable: namely that, as I mentioned at the outset, *causally efficacious properties are those properties whose instantiation accounts for their owners' having the causal powers they do*. The causally efficacious properties of the sonic boom, for instance, account for its *being capable* of shattering a pane of glass of such-and-such thickness under such-and-such conditions. By virtue of its loud volume, the boom is such that it *would* shatter a glass pane of that thickness under appropriate conditions. In the actual world those conditions are met, so the glass actually shatters.
Being such that a bullet eventually enters the man’s body, however, does not in anything like this way account for the flying-bullet event’s being capable of causing the man’s death under appropriate conditions. Certainly, being such that a bullet eventually enters the man’s body is a property the flying-bullet event will presumably have in any nearby possible world in which it in fact causes the man’s death; but it need not have that property whenever it is capable of causing the man’s death under appropriate conditions. For even if the man had escaped injury – say, by wearing a Kevlar vest – still the flying-bullet event would have been capable, in an obvious sense, of killing him under appropriate conditions -- which conditions presumably include his not wearing a Kevlar vest. In other words: in any relevantly nearby world in which the man is wearing Kevlar and so the flying-bullet event does not kill him, still that event will be such that it would kill the man in relevantly nearby worlds in which he is not wearing Kevlar.

Do I have a worked-out theory of “appropriate conditions”, of precisely when an F-event is “capable” of causing a G? No. But neither do I need such a worked-out theory to make the simple point I am making: namely that there is an intuitively obvious, if not obviously analyzable, distinction between, on the one hand, properties that are in some undeniable sense required for an event to cause the effects it in fact causes, and on the other hand properties that contribute to the event’s causal powers in virtue of which it is capable of causing those effects in the first place. I have argued that our intuitive notion of causal efficacy goes along with the latter but not the former sort of property, while ERP counts all properties of the former sort as causally efficacious.
The externalist who relies on something like ERP, then, is committed to a number of counterintuitive consequences. That fact is not devastating: maybe our intuitions about causal efficacy are radically mistaken. But a far more reasonable conclusion is that the common externalist response I am considering should be rejected.

VI. Conclusion

For all I’ve said in this paper, externalism about mental content might be true. There might also be good arguments in its favor. The internalist argument I considered, furthermore, might still be flawed. And in fact, it might be flawed for the very reason pointed out time and again in what I take to be the most common externalist response: namely, that examples of causally efficacious extrinsic properties abound. Nevertheless, I hope to have shown that that response is far from clearly right. On the contrary, the most obvious way of fleshing it out is clearly wrong.

Granted, the internalist has work left to do to show that only intrinsic properties can be causally efficacious. Toward that end, an informative analysis of the notion of causal efficacy is certainly the place to look. But contrary to popular opinion, the externalist has just as much work left to do to show that extrinsic properties can be causally efficacious. Just because what happens near a burning barn can cause things to happen that will not likely, or would not, or even could not happen elsewhere, that does not license us to infer that being near a burning barn is ever an efficacious property of the causes that have it.
Chapter 3:  

The Good, the Bad, and the Irrational:  

Three Views About Propositional-Attitude Content  

O. Introduction  

Imagine a guy named “Saul” who, being a well-trained philosopher, believes that Aristotle was wise. Now imagine a microphysical duplicate of Saul (call him “Shmaul”) who lives on a distant planet that is a microphysical duplicate of Earth (call it “Shmearth”). Lazing about in the Shmearthy grass, pondering the virtues of a guy he calls “Aristotle” (namely, Shmaristotle), presumably Shmaul believes something. We can then ask: is the content of Shmaul’s belief the same as the content of Saul’s belief?  

There are three possible answers to this question, corresponding to three famous views about the nature of propositional-attitude content. The first view I will call  

Internalism: Propositional-attitude content is narrow  

-- where content C is defined as “narrow” if and only if all possible microphysical duplicates of someone who has a belief (desire, fear, etc.) with C also have such a belief (desire, fear, etc.). According to the internalist about mental content, Saul’s belief and Shmaul’s belief share the content that Aristotle was wise.
The second view is

**Externalism:** By and large, propositional-attitude content is wide

-- where content C is defined as "wide" if and only if it is not narrow. According to the externalist, Saul's belief and Shmaul's belief may well have different contents: the content of Saul's belief, for instance, is *that Aristotle was wise*; while the content of Shmaul's belief is *that Shmaristotle was wise*.

The third view says that our initial question is ill-defined -- since by and large and on the whole, *propositional-attitudes have more than one kind of content*. In particular, this third view defines "folk-psychological content" as the denotation of "that"-clauses in ordinary, *de dicto* ascriptions of propositional attitudes, and then asserts what I call

**The Middle Way:** By and large, propositional-attitudes have two kinds of content: folk-psychological content, which is wide; and non-folk-psychological content, which is narrow.

According to defenders of the Middle Way, Saul's and Shmaul's beliefs differ with respect to folk-psychological content -- Saul's belief having the folk-psychological content *that Aristotle was wise*, and Shmaul's belief having the folk-psychological
content that Shmaristotle was wise – while Saul’s and Shmaul’s beliefs nevertheless share some non-folk-psychological kind of content.

Brian Loar, for instance, says that “narrow content is not in general captured by ordinary ‘that’-clauses”; and Jerry Fodor says that this is not just a contingent fact about ordinary “that”-clauses, but instead that “narrow content is radically inexpressible”1. Defenders of different versions of the Middle Way have different theories of the exact nature of this non-folk-psychological content; but for my purposes here, those differences will not matter.2

Internalism was arguably the commonsense view of the mind until the middle of the 20th century: Descartes, for instance, was a paradigm internalist. Revolutionary arguments by Hilary Putnam, Tyler Burge, and Saul Kripke in the 1970s have since made externalism the majority view. Versions of the Middle Way, finally, were introduced by Jerry Fodor and Brian Loar in the mid 1980s and have been endorsed in different forms by Colin McGinn, David Lewis, Frank Jackson, and David Chalmers, in an attempt to do justice both to the intuitions supporting internalism and to the intuitions supporting externalism.

If the Middle Way is accepted, then past disputes between internalists and externalists can simply be written-off as misunderstandings: internalists and externalists

---


2 Colin McGinn seems to take narrow content to be “descriptive” in nature; Jerry Fodor takes narrow content to be akin to David Kaplan’s notion of “character”; and Brian Loar takes narrow content to be individuated according to its functional role.
were simply talking about different kinds of content. The attractiveness of this peaceful Middle Way has led to its jargon being co-opted by the general philosophical community: even philosophers who would not admit to agreeing with anything Jerry Fodor ever said often speak unabashedly about a belief’s “narrow” content in addition to its “wide” content. Due to its having just this sort of influence, Fodor’s argument for the Middle Way has been claimed by Ned Block to have “set the main agenda for current philosophy of mind”. 3

I believe that each of the above three views is more misguided and less-well-motivated than its temporal predecessor. Descartes was right to hold that propositional-attitude content is narrow; Putnam, Burge, and Kripke were wrong, but at least reasonable, to hold that propositional-attitude content is wide; while Fodor, Loar, and their followers were simply irrational to believe in a peaceful Middle Way.

In this paper I will defend the strongest of these last claims: in particular, that the historically influential argument for the Middle Way, pushed by Jerry Fodor (1987) in Psychosemantics, is simply incoherent. As I will show, that influential argument rests on three premises which, upon close inspection, cannot be motivated simultaneously – so the rational conclusion, even if one feels attracted to each of the three premises, is that one had better go back and think again. This will leave internalism and externalism as the only reasonable options. I will end by sketching why I prefer internalism over

externalism; but the central moral will be that common appearances notwithstanding, we have to make some such choice.

I. The Middle Way argument

The following passage contains one of Fodor’s clearest summaries of his argument for the Middle Way:

At a minimum [the Burge and Putnam stories] show that propositional attitudes, as common sense understands them, don’t supervene on brain states.

But it’s arguable that any scientifically useful notion of psychological state ought to respect supervenience; mind/brain supervenience (and/or mind/brain identity) is, after all, the best idea that anyone has had so far about how mental causation is possible.

The moral would appear to be that you can’t make respectable science out of the attitudes as commonsensically individuated. … that we need, when doing psychology, other identity conditions for mental states than those that common sense prefers. This doesn’t bother me much, because … redrawing these boundaries doesn’t jeopardize the major claim on which the vindication of the attitudes as explanatory constructs depends -- viz., that scientific psychological explanation, like commonsense belief/desire explanation, is committed to states to which semantic and causal properties are simultaneously ascribable.4

I have broken this passage into three separate paragraphs, which I take to constitute the three key premises of Fodor’s argument. To get clearer about what these premises are, let me begin by laying out some definitions.

---

Let *folk-psychology* be the tacit theory that underlies our everyday explanations of intentional behavior in terms of folk-psychological contents of the behaving subjects’ propositional attitudes – where folk-psychological contents, again, are the denotations of “that”-clauses in ordinary *de dicto* attitude-ascriptions. Folk-psychological explanations, in other words, are explanations in terms of “that”-clauses.

Let *scientific* psychology, furthermore, be an *ideal* theory that explains intentional behavior in terms of those mental properties of the behaving subject that are *immediately causally relevant* to the behavior. Perhaps folk-psychology is properly scientific, or perhaps not; that will depend upon whether havings of propositional attitudes with folk-psychological content are ever immediately causally relevant to behavior.

The three separate paragraphs in Fodor’s above argument for the Middle Way, now, can be captured in the following three premises, respectively.

**Premise 1. Folk-psychological Externalism (FE):** By and large, folk-psychological content is wide.

*That Aristotle was wise,* for instance, is the content of Saul’s but not Shmaul’s belief. Note that (FE) is just externalism carefully phrased so as to avoid the tacit assumption that folk-psychological content is the only kind of propositional-attitude content.
Premise 2. Methodological Solipsism (MS): Only what is shared among all possible microphysical duplicates of a behaving subject can be mentioned in scientific-psychological explanations of that subject's intentional behavior.

A scientific-psychological explanation of Saul's signing-up for a philosophy course, for instance, cannot mention his having read books written by Aristotle; because not all possible microphysical duplicates of Saul have that same relational property. Hence it follows from (FE) that folk-psychological contents cannot normally be mentioned in scientific-psychological explanations: Saul's believing that Aristotle was wise, for instance, cannot be mentioned in a scientific-psychological explanation of his signing-up for a philosophy course any more than his having read books written by Aristotle can be.

Premise 3. Content-Conservatism (CC): Intentional behavior can be given scientific-psychological explanations that mention contents of the behaving subjects' propositional attitudes.

Saul's signing-up for a philosophy course, for instance, can be given a scientific-psychological explanation that mentions contents of at least some of his propositional attitudes. (CC) combines with the previous two premises to give us
**Conclusion.**  

**The Middle Way:** By and large, propositional-attitudes have two kinds of content: folk-psychological content, which is wide; and non-folk-psychological content, which is narrow.

(FE), (MS), and (CC) constitute my best interpretation of Fodor's argument for the Middle Way. My interpretation might be misguided. If so, then the upcoming criticism of Fodor's argument might be irrelevant. But my interpretation of Fodor's argument *works*; furthermore it certainly sounds like what he is getting at; and finally it is, at least in large part, fairly convincing. Hence I will leave it as a challenge to someone who wants to defend Fodor's argument for the Middle Way to come up with a better interpretation of that argument, an interpretation that avoids the criticism I am about to raise.

**II. Defending Content-Conservatism**

My criticism of the above argument for the Middle Way rests on how, exactly, the final premise (CC) is defended.

Fodor's main line of defense of (CC) begins from the premise that folk-psychology is remarkably successful as a tool for predicting intentional behavior. For instance: from the claim that "by all reasonable empirical criteria, the predictive adequacy of [folk-psychology] is not susceptible to serious doubt", Fodor concludes that
“we have substantial reason to believe that it is possible to have a scientific psychology that vindicates commonsense belief/desire explanation”\(^5\) – where he takes “vindication” of folk-psychology to be the retention of use of the notion of content in scientific psychology, even if the \textit{particular kind} of content mentioned in folk-psychology will not itself be retained.

How exactly is this “Pragmatic Defense” of (CC), as I will call it, to be filled-in? My most charitable hypothesis is that the first premise (that folk-psychology is remarkably successful as a tool for predicting intentional behavior) is meant to combine with something like the following principle to give us good reason to believe in (CC):

\[(P) \text{ To the extent that a psychological theory succeeds as a tool for predicting intentional behavior, to that extent there is good reason to believe that the properties it mentions are } \text{immediately causally relevant} \text{ to the behavior and so can be mentioned in scientific-psychological explanations.}\]

(P) certainly captures a plausible way of reasoning. To take an analogy: before Watson and Crick discovered the double-helix, we already had excellent reason to believe genetic theory as a causal explanation of the passing on of physical traits from parent to child;

precisely because that theory worked so well as a tool in predicting which traits would be passed on in which kinds of circumstance.

When (P) is combined with the initial premise that folk-psychology is remarkably successful as a tool for predicting intentional behavior, furthermore, we find good reason to believe in (CC). For if havings of the kinds of contentful propositional attitudes mentioned in folk-psychology were not in fact immediately causally relevant to intentional behavior, then it would be a mystery why folk-psychology is as predictively powerful as it in fact is. So we have reason to believe that havings of the kinds of contentful propositional attitudes mentioned in folk-psychology are immediately causally relevant to intentional behavior; so we have reason to believe that propositional-attitude content can be mentioned in scientific-psychological explanations.

This Pragmatic Defense of (CC), I think, works. In the upcoming section, however, I will show that it works only by undermining the argument for the Middle Way.

III. Incoherence regained

The Pragmatic Defense of (CC) does, as I said, justify that premise of the Middle Way argument. The problem comes when you note that the only way the Pragmatic Defense works is by way of the implicit conclusion I will call
Folk-psychological Content-Conservatism (FCC): Intentional behavior can be given scientific-psychological explanations that mention *folk-psychological* contents of the behaving subjects' propositional attitudes.

Note that (FCC) is just (CC) with the added requirement that the scientifically respectable kind of content is *folk-psychological*.

From the first two premises of the Pragmatic Defense of (CC), *that* is what immediately follows: since folk-psychology is so successful as a predictive tool; and since we have reason to believe that whatever is mentioned in a predictively successful psychological theory can be mentioned in a scientific-psychology; we therefore have reason to believe that *folk-psychological* contents can be mentioned in a scientific-psychology. To get (CC) we then must existentially generalize, from the claim that *folk-psychological* content is scientifically respectable to the conclusion that *some* kind of content *or other* is scientifically respectable.

As I will summarize this point: (CC) *is reasonable only if (FCC) is*. But now the argument for the Middle Way is in trouble; for (FE), (MS), and (FCC) form a mutually inconsistent triad. Hence any two of Fodor's three premises can form a coherently-motivated pair; but then the third, remaining premise cannot be coherently motivated. This can be seen by considering the three possible options, in turn.
First, a rational person who believes both (FE) and (MS) should reason as follows. “By (FE), I know that folk-psychological contents are wide. And by (MS), I know that wide contents cannot figure in scientific-psychological explanations; so I know that folk-psychological contents cannot figure in scientific-psychological explanations. So, I cannot believe (FCC), the claim that folk-psychological content can figure in scientific-psychological explanations; so I have no reason to believe (CC), since (CC) is reasonable only if (FCC) is. And so, since I have no reason to believe in the third premise of the argument for the Middle Way, I have no reason to believe in the Middle Way.”

This is how eliminativists like Stephen Stich (1983) and maybe Paul and Patricia Churchland reason: they are externalists about mental content, but believe that good science should stick to what is shared among microphysical duplicates; so they conclude that scientific psychology will eliminate all mention of mental content.

Second, a rational person who believes both (MS) and (CC) should reason as follows. “I believe (CC) because I believe (FCC), the claim that scientific-psychological explanations can mention folk-psychological content. By (MS), furthermore, I know that scientific-psychological explanations cannot mention wide content; so I conclude that folk-psychological content must not be wide. So I cannot believe (FE), the claim that folk-psychological content is wide. And so, since I cannot believe the first premise of the Middle Way argument, I have no reason to believe in the Middle Way.”
This is how full-blown internalists like Tim Crane (1991) and maybe John Searle reason: they see evidence for the scientific respectability of folk-psychology, but also think good science should stick to what is shared among microphysical duplicates; so they conclude that folk-psychological content must be shared among microphysical duplicates.

Third and finally, a rational person who believes both (CC) and (FE) should reason as follows: “I believe (CC) because I believe (FCC), the claim that scientific-psychological explanations can mention folk-psychological content. By (FE), then, which says that folk-psychological content is wide, it follows that scientific-psychological explanations can mention wide content. So I cannot believe (MS), which would have the consequence that scientific-psychological explanations cannot mention wide content. So, since I can’t believe the second premise of the Middle Way argument, once again I have no reason to believe in the Middle Way.”

This last line is how externalists like Tyler Burge and Fred Dretske are happy to reason: they accept the scientific respectability of folk-psychology, but also accept externalism about mental content; so they’re happy to conclude that good science need not stick to what is shared among microphysical duplicates.

These are three perfectly understandable ways to reason; but none of them leads to the Middle Way between internalism and externalism full-stop. As a certain philosopher might put the point: *cut the pie any way you like, Fodor’s argument for the Middle Way ain’t gonna work.*
IV. An alternative version of the Pragmatic Defense

Maybe, though, I’ve been uncharitable in fleshing out the Pragmatic Defense of (CC). In particular, you might think that to get from the initial premise that folk-psychology is a successful predictive tool to the conclusion that we have good reason to believe in (CC), we don’t need anything as strong as our previous premise (P) – which premise assumed that all properties mentioned by a successful tool for predicting behavior are themselves immediately causally relevant to the behavior and hence available for mention in a scientific-psychology.

In particular, one might instead assume something weaker, along the lines of

(P') To the extent that a psychological theory succeeds as a tool for predicting intentional behavior, to that extent there is good reason to believe that the properties it mentions are tightly correlated with properties that are immediately causally relevant to the behavior and that hence can be mentioned in scientific-psychological explanations.

We do not need to conclude from (P') that folk-psychological content itself can be mentioned in scientific-psychology; so if the Pragmatic Defense of (CC) rests on (P') instead of (P) then the argument for the Middle Way will not be undermined.
(P'), furthermore, is certainly reasonable. To take an analogy: suppose that solely on the basis of my knowledge that the Cowardly Lion has a heart while the Tin Man lacks one (plus knowledge of the environmental conditions in which both creatures are situated; e.g. that the humidity where they live is very low, that there is an abundance of vitamin D, and so on), I predict that the Lion will develop kidney-stones before the Tin Man does. In fact I will be right the vast majority of the time: in an infinite number of relevantly similar situations, the Lion will develop kidney stones before the Tin Man does an infinite number of times. However, my predictive success is not traceable to the fact that having a heart is immediately causally relevant to the development of kidney-stones; presumably, it is no such thing. Rather, my predictive success is traceable solely to the fact that having a heart goes hand-in-hand with some property that is causally relevant — namely, having a kidney. As all good philosophers know, the set of creatures with hearts is identical to the set of creatures with kidneys; so good predictions about which sorts of creatures are likely to develop kidney-stones can be made on the sole basis of knowledge about which creatures have hearts, even though having a heart need not be immediately causally relevant to the development of kidney-stones.

Unfortunately, the quite reasonable premise (P') does not help the Pragmatic Defense of (CC). For from the initial premise of the Pragmatic Defense (that folk-psychology is a successful predictive tool), plus (P') (the premise that what is mentioned in a psychological theory that is a successful predictive tool is likely tightly correlated with causally relevant properties), it only follows that folk-psychological contents are
likely tightly correlated with properties that can be mentioned in a scientific-psychology. Nothing at all follows about what kind of properties those causally relevant properties are: they might be neurophysiological properties, as the Churchlands think; or they might be syntactical properties of a neural code, as Stich thinks; but whatever the case, we have no reason at all to think that mentioning those properties requires mentioning mental content. And it is that last condition that is required for the justification of (CC).

V. One last try

One last try. Instead of (P'), you might be think, we should suppose instead that

(P'') To the extent that a psychological theory succeeds as a tool for predicting intentional behavior, to that extent there is good reason to believe that the properties it mentions are importantly similar to properties that are immediately causally relevant to the behavior and that hence can be mentioned in scientific-psychological explanations.

So even if we must believe for independent reasons that folk-psychological content cannot itself be mentioned in a scientific-psychology, we still have reason to believe that content, of some kind or other, can be.
Premise (P''), unlike our original premise (P), does not lead to (FCC) and so does not undermine the argument for the Middle Way. Unlike (P'), furthermore, (P'') might well give us reason to believe in (CC). Unfortunately, (P'') is entirely implausible.

If we are not allowed to explain the predictive success of folk-psychology by saying that the very properties it mentions are immediately causally relevant to behavior and so can be mentioned in a scientific-psychology, then we have no reason at all to suppose that what it mentions is in any important way similar to what can be mentioned in a scientific-psychology. To see this point, return with me to Oz.

Based solely on my knowledge that the Cowardly Lion has a heart but the Tin Man doesn't, I correctly predict that the Lion will get kidney-stones before the Tin-Man. Now does that give us reason to believe that some property relevantly similar to having a heart is immediately causally relevant to the development of kidney-stones? In particular, if we were to learn that hearts are causally irrelevant to the development of kidney-stones (for instance, by conducting experiments in which we have a control group of Lions and a group whose hearts we remove, though of course making sure the blood is circulated via some other means; and we note that the control group and the heartless group develop kidney-stones in identical numbers), should we then conclude that the Lion develops kidney-stones because he has a heart of some previously undiscovered kind?

Of course not – but that is just the sort of inference this last version of the Pragmatic Defense of (CC) asks us to be prepared to make. We are asked to accept that
even if we find reason to believe that the kind of content mentioned in folk-psychology cannot be mentioned in a scientific-psychology, still we have reason to believe that \textit{content of some previously undiscovered kind can} be so mentioned. But if there is no reason to believe that the Lion has a heart of some previously undiscovered kind, then premise \((P'')\) must be wrongheaded.

Perhaps, of course, there is some relevant difference between this imaginary Oz example and real-life cases of psychological explanation of intentional behavior, a difference which would make some restricted version of \((P'')\) acceptable. But the burden is certainly on a defender of the Middle Way to say what such a difference might be. On the face of it, I see no reason to think that a version of \((P'')\) that spoke only of psychological explanations of intentional behavior will fare any better than \((P'')\) itself.

\section*{VI. Conclusion}

To recap: a burgeoning view in the philosophy of mind contends that by and large and on the whole, propositional attitudes have two kinds of content – ordinary content of the kind ascribed in folk-psychology, which is “wide”; and some non-folk-psychological kind of content which is “narrow”.

The historically most influential line of argument for this sort of “Middle Way” between internalism and externalism rests on three premises: Folk-psychological Externalism (\textit{FE}); Methodological Solipsism (\textit{MS}); and Content-Conservatism (\textit{CC}).
That third premise, furthermore, is reasonable only if *Folk-psychological Content-Conservatism (FCC)* is; and (FCC) is incompatible with (FE) and (MS).

The upshot is that it would not be rational -- even for someone who is initially attracted to each of the premises in the historically influential argument for the Middle Way -- to believe in the Middle Way. Instead, the rational conclusion would be to go back and think again.

My own opinion, after going back and thinking, is that the obvious and enduring success of folk-psychology, along with the clear success of "solipsism" as a general scientific methodology, should make philosophers reconsider their current infatuation with folk-psychological externalism. I.e. we should go along with Crane, Segal, Searle, and others in accepting that *folk-psychological content is the only kind of propositional-attitude content there is, and it is narrow.*

But whatever the case, the main moral of this paper has been that we have to make some such choice. And recognizing that point should make many philosophers who write about mental content -- in particular, the increasingly many who unthinkingly co-opt the Middle Way jargon, simply because it seems to let them avoid taking sides -- go back and rethink their philosophical commitments.
Chapter 4: Two Grades of Internalism (Pass and Fail)

O. Introduction

Imagine a guy named “Sam”, who is in the midst of living a full and exciting mental life: he’s wondering if he’ll ever develop arthritis, craving a tall glass of ice-water, pondering the virtues of Aristotle, realizing that his pants are on fire; and so on. As it happens, Sam is a competent neuroscientist -- so competent, in fact, that in his spare time he has constructed, and keeps alive in a jar on his mantel, a particle-for-particle duplicate of his own brain. Call this envatted duplicate of Sam’s brain “Sam’s vat-brain”.

It seems reasonable to suppose that if strange enough things had happened after Sam’s birth, Sam’s brain could have been, for most or even all of its existence, kept alive in a jar in a mad scientist’s laboratory. In fact, if things had gone very differently from the beginning of time onward, it seems possible that a vat containing Sam’s brain could have been the only thing that ever existed. Imagine such a possible world, in which nothing has ever existed except a vat containing Sam’s brain; call Sam’s brain, in that possible world “vat-Sam”.

On even a very modest assumption about the extent to which the mental depends on the physical, presumably some mental events are occurring in both Sam’s vat-brain and vat-Sam. At the very least, for instance, since Sam believes that Aristotle was wise and craves a glass of ice-water, presumably some believing-event and desiring-event are
occurring in both Sam's vat-brain and vat-Sam. Hence we can ask: what relations of similarity and difference hold among the mental events in Sam's brain, Sam's vat-brain, and vat-Sam?

On only a slightly less modest assumption about the dependence of the mental on the physical, we must accept that many properties are shared among the three. If Sam feels a headache or a general sense of ennui, for instance, then presumably a similar feeling is occurring in both Sam's vat-brain and vat-Sam; if a visual experience as of some red wavy lines is occurring in Sam's vat-brain, then presumably a similar experience is occurring in both Sam and vat-Sam; and so on. But when it comes to properties consisting of the having of particular mental contents, the answer is not so clear.

On the popular view I will call "externalism about mental content", first of all, many mental events in Sam's brain will have contents not shared by corresponding events in either Sam's vat-brain or vat-Sam. An externalist might say, for instance, that no belief in either Sam's vat-brain or vat-Sam can share the content of Sam's belief that Aristotle was wise -- because in order for a belief to be about Aristotle it must be caused by Aristotle in appropriate ways, and no beliefs in either Sam's vat-brain or vat-Sam were caused by Aristotle in appropriate ways. A view I will call "grade-1 internalism", on the contrary, holds that all contents of mental events in Sam's brain must be shared by corresponding events in both Sam's vat-brain and vat-Sam. A grade-1 internalist will
say, for instance, that there must be beliefs in both Sam's vat-brain and vat-Sam that share the content *that Aristotle was wise*.

Finally, a view I will call “grade-2 internalism” falls somewhere in-between the previous two views. In particular, grade-2 internalism holds that all contents of mental events in Sam’s brain will be shared by corresponding events in vat-Sam but at least some such properties will not be shared by any events in Sam’s vat-brain. A grade-2 internalist might say, for instance, that although all contents of mental events in Sam’s brain must be shared by corresponding events in vat-Sam, no belief in Sam’s vat-brain shares the content of Sam’s belief *that his pants are on fire*. The reason: both Sam’s belief and the corresponding belief in vat-Sam, unlike any belief in Sam’s vat-brain, is about *Sam*.

In the considerable literature surrounding the debate between externalism and internalism, I have seen the possibility of distinguishing between the above grades of internalism recognized only once.¹ In this paper I will clarify that distinction, show how important it is to recognize, and argue that the more popular of the two grades of internalism is untenable. Although many internalists might have hoped to agree with externalists on this point, I will argue that they will have to accept that Sam’s belief that his pants are on fire shares its propositional content with both the corresponding belief in vat-Sam and the corresponding belief in Sam’s vat-brain.

I. **Externalism and two grades of internalism**

Sam, remember, is wondering if he'll ever develop arthritis, craving a tall glass of ice-water, pondering the virtues of Aristotle, and realizing that his pants are on fire. The question we have set ourselves is: "To what extent are the contents of mental events in Sam's actual brain shared by corresponding events in either Sam's vat-brain or vat-Sam?" The correct answer to this question depends on the correct view about the nature of mental content. And although only two relevant views are widely recognized, the distinction among relevant views is in fact threefold. I will next lay out this threefold distinction and illustrate the lesser-recognized view with a paradigm example of it from the literature.

I.i **The basic distinction**

First, some terminology. Let me say that a property of a thing at a time is *external-world dependent* (EW-dependent) if the thing's having the property at the time depends, as a matter of metaphysical necessity, on what the world is like either outside of the thing's location or at at least some other times. *Occurring near a burning barn,* for instance, is always EW-dependent since its instantiation metaphysically demands the existence of a nearby burning barn; similarly *postdating Aristotle's death* is always EW-dependent since its instantiation demands the past death of Aristotle. *Having molecule M as a physical part,* on the other hand, and *having such-and-such rest-mass* are always
EW-independent since something that has those properties could do so no matter what the external world around it were like.

In these terms, we can state the leading view of mental content as

Externalism: Many content-properties of mental events are EW-dependent.

Some possible externalist claims, for instance, are that for my desire to have the content that I have a drink of water requires my having adequately interacted with genuine samples of H₂O (e.g. by drinking some); and that for a belief to have the content that Aristotle was wise requires my having been suitably causally impacted by Aristotle himself (e.g. by reading his books).

If externalism is true then many mental events in Sam’s brain will presumably have contents that are not shared by corresponding events in either Sam’s vat-brain or vat-Sam. Since neither Sam’s vat-brain nor vat-Sam has interacted in the appropriate way with samples of H₂O, for instance (e.g. by drinking some), or has been suitably causally impacted by Aristotle (e.g. by reading his books), neither can share Sam’s thirst for water or his belief that Aristotle was wise.

The opposition to externalism is often stated as

Internalism: All content-properties of any mental event are intrinsic properties of that event.
An *intrinsic property* can be thought of as a property that is necessarily shared among *microphysical duplicates* of anything that has the property, where microphysical duplicates are things that are indiscernible from the perspective of a completed particle physics: they are things that are composed of all of the same material parts arranged in exactly the same way. Thus this "grade-1" internalism says that content-properties are necessarily shared among microphysical duplicates.

In general, any property that supervenes on the intrinsic properties of its owners is EW-independent, since microphysical duplicates of can in general occupy arbitrarily different external-world contexts. Hence grade-1 internalism entails the falsity of externalism: it judges that all of the contents of events in Sam's brain must be shared *both* by events in Sam's vat-brain *and* by events in vat-Sam. Since each of the three brains is a microphysical duplicate of the others, any content had by an event in one of the brains must be shared by some event in each of the others: both Sam's vat-brain and vat-Sam are thirsting for a glass of ice-water even though they cannot drink; and they are pondering the virtues of Aristotle even though they cannot read.

Grade-1 internalism is stronger than a mere denial of externalism, however, since not all EW-independent properties are intrinsic. Of the two previous examples of EW-independent properties, for instance — namely, *having molecule M as a physical part* and *having such-and-such rest-mass* — only the second is intrinsic. For microphysical duplicates of something with molecule M as a part need not have molecule M as a part;
whereas microphysical duplicates must, presumably, be indiscernible with respect to rest-
mass. The following view, then, is incompatible with externalism just like grade-1
internalism, but is weaker than grade-1 internalism:

Internalism₂: All content-properties of mental events are EW-independent.

Since the only difference between Sam’s actual brain and vat-Sam consists in
what the world around that brain is like, this “grade-2” internalism judges that all
contents of events in Sam’s actual brain must be shared by events in vat-Sam. Thus
grade-2 internalism is incompatible with externalism. But grade-2 internalism is weaker
than grade-1 because it remains silent about whether all contents of events in Sam’s
actual brain are shared by corresponding events in Sam’s vat-brain. And this is because,
even though Sam’s vat-brain is a microphysical duplicate of Sam’s actual brain, many
properties of events in Sam’s brain are EW-independent but not intrinsic.

For instance: supposing neuron N to be a constituent of some event in Sam’s
actual brain, *having neuron N as a constituent* is an EW-independent property of that
event that is presumably not shared by any event in Sam’s vat-brain. It is EW-
independent, first of all, because its being had does not depend on what the world is like
outside of the place occupied by the event that has it or at other times; but it is not shared
by any event in Sam’s vat-brain, presumably, since Sam’s vat-brain has no neural parts in
common with Sam’s actual brain. So grade-2 internalism leaves open the possibility,
whereas grade-1 internalism does not, that some contents of events in Sam’s actual brain are not shared by corresponding events in Sam’s vat-brain.

According to both of the above grades of internalism, all mental-content properties can be instantiated by events in a brain in a vat; for everything that determines mental content is “in the head”. But only according to grade-1 must your current mental contents be shared by events in any brain that is a microphysical duplicate of your own. It will help to bring out this distinction by looking at a theory of mental content that is paradigmatically internalist yet is incompatible with grade-1 internalism.

I.ii. Searle’s theory

John Searle is a committed internalist, yet he goes to great lengths to avoid grade-1 internalism. As I will put it, Searle is a “mere” grade-2 internalist, endorsing something like the conjunction of grade-2 internalism with the negation of grade-1.²

All of the fun happens around propositional attitudes that are what I will call “indexical”: an indexical attitude is one whose content is naturally expressed by a sentence containing indexicals. Sam’s belief that his pants are on fire, for instance, is an indexical belief since Sam would naturally express its content by saying “My pants are on fire”; similarly with my hope that it is now twelve-o’clock (the natural expression of which would use the indexical “now”) and your fear that real-estate prices around here are too high (the natural expression of which would use the indexical “here”).

According to Searle, indexical attitudes have propositional contents that are "causally self-referential". An example he uses is the perceptual belief *that that man (over there) is wearing a red cap*. This is an indexical belief in the relevant sense because the natural expression of its content contains "that". According to Searle, furthermore, the content of this belief is something like *that there is a man causing this very belief and the man is wearing a red cap.*[^3] This content is "self-referential" in the sense that it makes essential reference to the belief that has it: no attitude that is not about that very belief (and also about the man wearing the cap) can share that content. For you or I to share Searle’s belief, we would have to believe that there is a man wearing a red cap causing *Searle’s belief*.

Putting the point, as will prove helpful, in terms of content-properties, and also returning to the example with which we began, consider the property of Sam’s belief that consists in that belief’s having the content that it does – viz., the property consisting of a relation to the proposition Sam would express by saying “my pants are on fire”. Call that property “SAM’S PANTS”. Then according to Searle, SAM’S PANTS amounts to something like the following property:

\[ [\text{SAM’S PANTS}: \text{having the content } \text{that the pants of the subject of [Sam’s current believing-event] are on fire}] \]

Chapter 4: “Two grades of Internalism”

– where I have bracketed “Sam’s current believing-event” in order to signify that that phrase is a rigid designator, denoting in all possible worlds the token believing-event actually occurring in Sam.

What this means is that the proposition Sam believes necessarily refers to the token-event consisting in Sam’s having that belief. No one else who believes that their pants are on fire shares Sam’s belief, because in order to share Sam’s belief one would have to believe that Sam’s pants are on fire. In fact even when, all those times in the past, Sam himself believed that his pants were on fire he did not share the belief he has now; for those past believing-events of Sam’s were token-distinct from his current believing.

The intuition that Searle’s theory captures is that Sam’s belief is true if and only if his pants are on fire, while anyone else’s belief that their pants are on fire will be true if and only if their pants are on fire. The intuition is undeniable: when Hume believes that he is Hume, he is right, while when Heimson believes that he is Hume, he is wrong; so Hume’s belief must be true if he is Hume, while Heimson’s belief must be true only if he is Hume. Yet Searle’s theory captures this intuition in an internalist-friendly way, since there is nothing stopping “self-referential” propositional contents from being had by events in brains-in-vats. If Sam’s brain, in its actual state, had been in a vat instead of in his skull, there would still have occurred in it a belief with the content had by his actual belief; so that event would have had [SAM]’S PANTS just as Sam’s belief actually does; so some belief in vat-Sam must share the content of Sam’s belief. Hence Searle’s theory amounts to at least grade-2 internalism.
Searle’s theory is incompatible with grade-1 internalism, however, and hence amounts to mere grade-2 internalism, because properties consisting in the having of self-referential contents need not be shared among numerically distinct but intrinsically indiscernible events. Since Sam’s vat-brain is an intrinsic duplicate of Sam’s own brain, some belief presumably occurs in it that is about pants and fire: that much will be had in common by the content of Sam’s belief and the content of the belief in Sam’s vat-brain. But if Searle is right then the belief in Sam’s vat-brain will be about itself (viz., the belief in Sam’s vat-brain), just as Sam’s belief is about itself (viz., Sam’s belief). So Sam’s belief will have [SAM]’S PANTS while the belief in Sam’s vat-brain will not. Which gives us mere grade-2 internalism. 4

The distinction between two grades of internalism, then, is far from academic: one influential internalist, at least, seems to feel that it is quite important to steer clear of the stronger grade. And in fact this view arguably extends as far back as Descartes. The father of skepticism was a paradigm internalist, since he thought that all of our beliefs about the external world could be exactly as they are even if the external world did not exist. But he was also arguably a mere grade-2 internalist, since he arguably thought that

---

4 In fact Searle (1983, ch. 8, sec. 1) extends his grade-2 internalism beyond attitudes that are explicitly indexical in the way that Sam’s belief that his pants are on fire is. Searle seems to agree with Putnam’s original idea that attitudes about water are implicitly indexical: for original-Putnam, for instance, the belief that water is wet is in reality an indexical belief whose content is something like that [the stuff that fills the oceans and lakes around here] is wet; while for Searle, the content of that belief is something like that the stuff that fills the oceans and lakes and is among the causes of [that very believing-event] is wet. Accordingly, Searle would agree with Putnam that Sam’s vat-brain lacks Sam’s desire for water; but he would disagree with Putnam about whether that desire is shared by vat-Sam. Putnam again would say ‘No’, while Searle would say ‘Yes’.
different persons believe different things when they each believe, of themselves, that they exist.

Venerable opinion notwithstanding, I will next argue that it is in fact the weaker of the two grades of internalism that is untenable. I will argue in particular against Searle’s theory of mental content, which tactic will admittedly leave open the possibility that a version of mere grade-2 internalism could be constructed that avoids the pitfalls of Searle’s own version. But at the very least my arguments will pose the following challenge: unless a version of mere grade-2 internalism can be constructed that avoids the pitfalls of Searle’s theory, internalists should agree that microphysical duplicates must be mental duplicates as well.

II. First argument against mere grade-2 internalism: communication

What I will call the “communication argument” contends that if mere grade-2 internalism is true, then we cannot communicate the contents of many of our propositional attitudes to others in the way that we ordinarily believe we can. As promised above, I will focus on Searle’s version of mere grade-2 internalism in particular.

The heart of the communication argument rests on noticing that Searle’s theory entails what I will call a “Fregean view” of indexical propositional attitudes. Indexical attitudes, again, are beliefs, desires, and so forth whose content is most naturally expressed by the use of indexicals. According to Searle, indexical attitudes have
propositional contents which essentially involve the particular token events consisting of
the having of those attitudes. Hence anyone's having an attitude with the same
propositional content as Sam's belief that his pants are on fire, for instance, depends, as a
matter of metaphysical necessity, on the occurrence of Sam's particular token believing-
event: were Sam not to have that particular belief, no one could have a propositional
attitude that shared its actual content.

When combined with the falsity of externalism, this last consequence entails –
albeit to Searle's dismay⁵ -- what I will call a "Fregean view" about my belief that my
pants are on fire. First of all: if no one could have a propositional attitude that shared the
actual content of my belief unless I also were to have that belief, then if anyone did have
such an attitude, their attitude's having that content would not be EW-independent. In
which case externalism would have to be true. Thus if externalism is false, no one else's
propositional attitudes can share the content of my belief that my pants are on fire. This
idea, that no one else can have a propositional attitude that shares the content of my
belief, is properly called "Fregean" because according to Frege indexical attitudes are
possible only because "everyone is presented to himself in a particular and primitive way,
in which he is presented to no-one else".⁶ When I believe that my pants are on fire I

---

⁵ "I have called my account of indexicals 'Fregean' in spirit, but it is quite different from Frege's few actual remarks about indexicals.... The idea of incommunicable senses is profoundly anti-Fregean, since the notion of sense was introduced, in part, to provide a publicly graspable content ..." (Searle, 1983, p. 229). But Searle says some confusing things in addition: "Of course [my and your indexical] beliefs will be different in the trivial sense that any self-referential content makes reference to a particular token and not to qualitatively similar tokens, but that is a result we want anyway ..." (pp. 212-3).

stand in relation to a proposition that no one other than me can believe, disbelieve, or bear any other such attitude toward.

The communication argument against mere grade-2 internalism, then, can be brought out by considering a similar argument against the Fregean view just described. In particular, it is a well-rehearsed complaint against the Fregean view just described that if no one else can believe, disbelieve, or bear any other such attitude toward the proposition I believe when I believe that my pants are on fire, then I cannot communicate the content of that belief to others in the way we ordinarily believe I can. I will state the argument as follows.

If the Fregean view is right then no matter what I show or tell others, I cannot lead them to believe (or to disbelieve, or to bear any other attitude toward) what I believe. If being able to believe (or to disbelieve, or to bear at least some attitude toward) what a person believes is taken to be a necessary condition on knowing what that person believes, furthermore, then since communication is just the process of leading others to know what one believes, the Fregean view makes communication impossible. But communication is certainly possible: when I tell you that my pants are on fire, you learn precisely what it is I believe; namely, that my pants are on fire. Hence the Fregean view is false.

A Fregean will presumably want to challenge the assumption that learning what a person believes requires being able to bear at least some attitude toward the proposition
believed. I am not going to defend that assumption directly; but neither do I feel a very strong need to, for two reasons.

First of all, the assumption is certainly a *prima facie* plausible as part of an account of the nature of communication: ordinarily when you are unable to mull over the proposition a person expresses, or consider believing it, or bear any other attitude at all toward it, the explanation is simply that you do not know which proposition it is. For instance, a Heideggerian informs you excitedly that *the Nothing nothings*, and asks whether you agree: the reason that you would be unable to answer, or at least to consider whether or not you *do* agree, is plausibly that you simply do not know which proposition the Heideggerian is expressing. Accordingly, if the Fregean thinks that learning which proposition a person believes does not require being able to bear at least some attitude toward it, the burden is on her to show so.

And second of all, as I will next argue, the Fregean has little available as an alternative account of the nature of communication. If we accept that you cannot bear belief or any other attitude toward the proposition that my pants are on fire, then there remain only a few minimal senses in which you might be said to know which proposition I believe when I tell you that my pants are on fire. And those senses, I will next argue, do not go nearly deep enough to capture what we ordinarily mean when we say that you know what I believe.

II.i. **First response to the communication argument: metalanguage**
Harold Noonan, in the following defense of the Fregean view, draws out one sense in which you can know what it is I believe when I express the content of my belief in the ordinary way:

Fregean 'I'-thoughts are indeed in one sense private and incommunicable, but not in any objectionable sense. They are private and incommunicable in that only their subject can think them; but other people can know exactly what thoughts he is thinking when he does so ... When I say, for example ... 'I have a headache', I thereby give expression to a thought you cannot think; but you can say exactly what thought it is which I thus express and you cannot think: namely, the thought (for there is only one) expressible by me using the sentence 'I have a headache'. (Noonan 1984, p. 216)

Returning to the previous example, Noonan's point is that when I tell you that my pants are on fire, you can know what it is I believe in the following sense: you can know that I believe the proposition I express with the sentence “my pants are on fire”.

This sense, however, is too weak to do justice to the commonsense claim that you can know what it is I believe. Consider an analogy: deep in the jungle you come upon a native who utters “bananas are yummy”. You infer that the native believes that bananas are yummy, so you reach into your bag to offer him a banana; but a passing linguist stops you, informing you that although the native expresses some proposition with the sentence “bananas are yummy”, he does not thereby express the same proposition you would express. You ask, naturally, what proposition the native does express with that sentence; and the linguist tells you that unfortunately, the native expresses with that sentence a
Chapter 4: "Two grades of Internalism"

proposition that you not only cannot express, but also cannot believe, disbelieve, or bear any other such attitude toward. What can you conclude about the belief that the native is expressing with the sentence "bananas are yummy"? The obvious answer is "not much".

Certainly, you know a lot about what the content of the native’s belief is not: in particular, for any proposition that you can entertain, in the sense of believing, considering, mulling over, doubting, and so on, you know that the native’s belief does not have that proposition as its content. You know, then, something about what the native believes. But at least in this instance, knowing what something is not does not come close to knowing in any relevant sense what that thing is. If, when I believe that my pants are on fire, you can only know what I believe in the sense of knowing many things that it is not, then you are far less capable of knowing what I believe than we ordinarily assume.

II.ii. Second response to the communication argument: de re belief

There is, however, a second and more promising sense in which a Fregean might be able to know what it is I believe: namely, by knowing that I believe, with respect to the x that is identical with me, that x’s pants are on fire. In other words: even though it would be incorrect to ascribe de dicto to you knowledge of my belief in the proposition I believe, still it might be correct to ascribe de re to you such knowledge. There might be, for instance, some proposition of the form that the pants of ___ are on fire such that the
constituent that fills in the blank represents me, and such that it would be correct to ascribe \textit{de dicto} to you knowledge that I believe some proposition of a similar form.

I accept that often, when ascribing propositional attitudes to others, we make just this type of \textit{de re} maneuver. In particular, when the particular proposition believed is less important than the subject-matter of the belief, i.e. the entities and properties that the belief is about, then making the above type of \textit{de re} maneuver is often wiser, given its less-committal nature, than ascribing a belief \textit{de dicto}. For instance: when explaining to a NASA official why I need to borrow a spaceship, I might point to Venus and mention that my friend believes that that planet is beautiful and wants to visit it. This is presumably all I need to do, in order to secure the spaceship for the desired purpose; even if it is only by virtue of believing \textit{that the morning-star is beautiful} that my friend believes, with respect to Venus, that it is beautiful. Even if, for instance, all my friend knows about Venus is that it is the morning-star, in which case I could correctly report his belief \textit{de dicto} only by saying that he believes that the \textit{morning star} is beautiful, I will lose nothing by going \textit{de re} in my explanation to the NASA official.

Although the above \textit{de re} maneuver has its place, however, I do not think it is enough to do justice to our intuition that the Fregean can know, in a suitably deep sense, what it is I believe when I believe that my pants are on fire. To see this, notice that the sense in which you know what it is I believe is a strong enough sense to allow you to you know part of my reason for acting, when I act as a result of that belief. For instance: not wanting my pants to be ruined, I jump into a nearby lake. By knowing that I believe that
my pants are on fire, then, you know at least part of my reason for jumping into the lake: that knowledge, along with knowledge of certain other of my beliefs and desires, allows you to reasonably infer that I jump into the lake.

But the *de re* response to the communication argument against the Fregean view does not allow this. For if your knowledge of what it is I believe consists only in the knowledge that I believe, *with respect to the x that is identical with me*, that x’s pants are on fire, then you do not thereby know part of my reason for jumping into the lake in any relevant sense. This becomes clear by noticing, as John Perry has famously pointed out, that my believing *that my pants are on fire* can give me reason to behave in ways I would have no reason to behave simply by virtue of believing *with respect to the x that is identical with me* that x’s pants are on fire. If I were to see myself in a mirror, for instance, but fail to realize that it is I who I am seeing, then I might still believe, with respect to the x that is identical with me, that x’s pants are on fire; yet I would have no reason at all to jump into the lake. (Instead I might, for instance, throw a bucket of water toward where I believe the man with the burning pants to be.)

Hence even when accompanied by knowledge of my other relevant beliefs and desires, mere knowledge that I believe, with respect to the x that is identical with me, that x’s pants are on fire does not allow you to reasonably infer how I will behave. Which means that the *de re* response cannot protect the Fregean view, and hence cannot protect mere grade-2 internalism, from the communication argument.

---

III. Second argument against mere grade-2 internalism: causation

A much-discussed argument for internalism contends that if externalism is true, then many mental properties must be unable to play the important causal role in producing intentional behavior that we think they often do. As I will explain in a moment, this "causal" argument has often been fleshed out so as to remain neutral between the two grades of internalism I have distinguished. But when the underlying idea behind the causal argument is fleshed out correctly, it tells against mere grade-2 internalism even more strongly than it tells against externalism simpliciter. In particular, when it is understood exactly why the internalist is right to claim that many mental properties as they must be if externalism is true cannot be play the desired causal role, it becomes apparent that the same holds for many mental properties as they must be if mere grade-2 internalism is true.

A sample of the causal argument against externalism runs as follows. Sam believes that Aristotle was wise and wants to learn more about a wise man; so he signs up to take a course in ancient Greek philosophy. The instantiation of the following property by one of Sam's believing-events, accordingly, plays a particularly important causal role with respect to his signing-up for the philosophy course:

\[ \text{ARISTOTLE: having the content that Aristotle was wise.} \]

---

Now a paradigmatic externalist claim is that part of what it is for a belief to be about Aristotle is for it to have the following sort of property:

EXISTOTLE:  *occurring at the end of a causal chain in which Aristotle himself suitably participated.*

Being caused at least partly by hearing about Aristotle and reading books by him, for instance, might be claimed by an externalist to be metaphysically necessary conditions on a belief's being about Aristotle. But, the internalist chimes in, the instantiation of EXISTOTLE by one of Sam's believing-events being does not play anything like the important causal role played by the instantiation of ARISTOTLE with respect to Sam's signing-up for the philosophy course. It follows that externalism is false.

Much past rhetoric behind the internalist's idea that properties like EXISTOTLE cannot play the desired causal role has focused on the fact that those properties consist at least partly in relations between to entities in the mental subjects' present or past environment – i.e., that they are EW-dependent. Colin McGinn, for instance, offers a paradigm instance of this rhetoric:

According to externalism, contentful states are identified by reference to entities that lie outside the subject's body .... And now the causal objection to citing such contents in psychological explanations is just this: ... these correspondence relations cannot themselves be implicated in the causal
transaction being reported. This is because … the features of the cause that lead to the effect must be right there where the causal interaction takes place. Causation is the same with brains and minds as it is with billiard balls. Their effects depend upon local properties of these entities. The causal powers of a state or property … cannot depend essentially upon relations to what lies quite elsewhere. (McGinn, 1989, p. 133)

Hence much past rhetoric has allowed internalists to remain neutral between the two grades of internalism I previously distinguished: it is as true on mere grade-2 internalism as it is on grade-1 that all mental-content properties are EW-independent. This fact might explain, it seems to me, why the distinction between two grades of internalism has so long been overlooked. However, when the underlying idea behind the causal argument is fleshed out correctly, I believe that it tells against mere grade-2 internalism even more strongly than it tells against externalism simpliciter.

At least one aspect of the important causal role that the instantiation of ARISTOTLE plays consists in its being ceteris paribus necessary for his behaving as he does; just as at least one aspect of the important causal role played by a firecracker’s exploding, with respect to the shattering of a nearby window, consists in its being c.p. necessary. There are two ways to make this intuition vivid. First: just as the window would not shatter, holding other relevant facts constant, if the firecracker did not explode, so Sam would not sign up for the philosophy course if he did not believe that Aristotle was wise.⁹ And second: just as any other firecracker that differs from the first only by failing to explode will, again holding other relevant facts constant, fail to shatter any

---

⁹ This broadly counterfactual test for the relevant sort of causal role has been endorsed by Ernest Lepore, Barry Loewer, and Jean Kazez among many others.
windows, so any other person who differs from Sam only by lacking the belief that Aristotle was wise will not sign up for a philosophy course.\footnote{This type of test for the relevant causal role is used by Jerry Fodor in his elaborations of Putnam’s Twin-Earth thought experiment.}

The internalist’s causal argument can hence be defended by pointing out that the instantiation of EXISTOTLE by one of Sam’s beliefs is not c.p. necessary, in the appropriate sense, for his behaving as he does. First: if it had been Aristotle’s nephew who (unbeknownst to Sam) wrote the corpus actually written by Aristotle, that would not have affected Sam’s reasons for wanting to sign up for the philosophy course and so would not have changed the fact that he eventually signs up. And second: even a microphysical duplicate of Sam, for instance, will sign up for a philosophy course even if he was just yesterday conjured up in a biogenetics laboratory and so never actually heard of Aristotle or read any of his writings.

When fleshed out in this way, the causal argument can be seen to tell against mere grade-2 internalism at least as strongly as it tells against externalism \textit{simpliciter}. To see this, recall the property of Sam’s belief that his pants are on fire that consists in that belief’s having the content that it does -- i.e. the property I called “SAM’S PANTS”. Also recall that according to Searle (our paradigm grade-2 internalist), SAM’S PANTS is identical with
[SAM]'S PANTS: having the content that the pants of the subject of [Sam's current believing-event] are on fire.

The instantiation of SAM'S PANTS by one of Sam's believing-events is presumably c.p. necessary for his jumping into a nearby lake. If Sam did not believe that his pants are on fire, for instance, then holding other relevant facts constant he would not jump into the lake. Thus if there is no corresponding intuition that the instantiation of [SAM]'S PANTS is c.p. necessary, we will have to conclude that Searle is wrong to identify SAM'S PANTS with [SAM]'S PANTS.

In fact the intuition that the instantiation of [SAM]'S PANTS is not c.p. necessary Sam's jumping into the lake can be made vivid in the same two ways as before. First: if [SAM]'S PANTS had not been instantiated by any of Sam's beliefs, then holding other relevant facts constant, he still could easily have jumped into the lake -- in fact for exactly the same reasons that he actually jumped. To see this, just imagine a counterfactual scenario in which Sam's actual believing-event does not occur but one very much like it does. In that scenario none of Sam's beliefs would have [SAM]'S PANTS, since Sam would not believe anything about his actual believing-event (instead, he would have a belief about the believing-event that occurs in him in that counterfactual scenario); but Sam would have all of the same reasons to jump into the lake that he actually has.
Second: any other relevantly similar person will jump into a lake just the same, even if (and the internalist must say though) none of their beliefs has [SAM]'S PANTS. This point can be seen most vividly by imagining that I am a microphysical duplicate of Sam and that my environment that is a microphysical duplicate of Sam’s environment. None of my beliefs, presumably, will have [SAM]'S PANTS, since I will presumably not believe anything about Sam’s believing-event: instead, I will have a belief about my own believing-event. (In fact an internalist must say that none of my beliefs can have [SAM]'S PANTS, since my having that belief would have to be EW-dependent.) Nevertheless, I will have all of the same reasons to jump into a lake that Sam has: I, too, believe that my pants are on fire; I, too, do not want my pants to be ruined; I, too, believe there is a lake in front of me; and so on.

III.1. First response to the causal argument: causal convergence

Historically, externalists have offered two main responses to the internalist’s causal argument; both of which can be raised in response to the version of that argument I just sketched against mere grade-2 internalism. The first response points out that the instantiation of SAM’S PANTS, by one of Sam’s mental events, might very well play the desired causal role with respect to Sam’s jumping into the lake on the one hand even if Sam would have jumped into the lake for some other reason had he not believed that
Aristotle was wise, and on the other hand even though many other persons in fact jump into lakes for just such other reasons.\textsuperscript{11}

This response draws on the truism that one and the same type of effect can often be caused in more than one way. One firecracker, for instance, might explode as a result of its wick being lit while another explodes as a result of being heated in an oven. Let us call cases in which one and the same type of effect is caused in more than one way cases of "causal convergence". The mere grade-2 internalist need only accept such causal convergence in the case of Sam and me: Sam's jumping into the lake is at least partly caused by the instantiation of [SAM]'S PANTS by one of his beliefs, while my jumping into the lake is at least partly caused by the instantiation of a different property (namely, having the content that [my] pants are on fire). The general idea here is certainly not difficult to accept, since I could have jumped into the lake as a result of deciding that it was a nice day for a relaxing swim, or as a result of wanting to train for an upcoming race -- and many others jump into lakes for just such reasons.

However, I see a crucial difference between these uncontroversial cases of causal convergence and the case of Sam and my jumping into the lake. The difference is that the causes in the uncontroversial cases differ in their causal powers; whereas Sam and I have identical causal powers. This is demonstrated by the fact that for all of the uncontroversial cases, it is easy to imagine some possible context such that, when each of

\textsuperscript{11} E.g. Klein (1996).
the actually-convergent causal processes is placed in that context, they do not converge; while for Sam and I, the causal processes leading to our behavior will cause identical effects whenever they are placed in identical contexts.

For instance: when one firecracker's wick is lit and another firecracker is placed in an oven, we can make the first remain intact and the second explode by cutting off both of their wicks; or we can make the first explode and the second remain intact by shutting off the electricity. Or again: inform me and everyone else that despite first appearances it is a terrible day for a swim, and I will still jump into the lake while those who had decided it was nice day for a swim will not; inform me and everyone else that the race has been cancelled, and I will still jump into the lake while those who wanted to train for the race will not. Or dump a bucket of water over each of us, and I will no longer need to jump into the lake while the others will only want to do so sooner. Do any of these things to both Sam and I, however, and we will be affected in exactly the same way: we will both still jump into the lake if informed that it is a terrible day for a swim or that the race is cancelled; neither of us will jump if a bucket of water is dumped on us first.

Since the causes in the uncontroversial cases of causal convergence differ in their causal powers, i.e. in what sorts of effects they are capable of producing under different sorts of circumstances, it is perfectly reasonable to see them as differing with respect to properties that can play an important causal role with respect to those causes' effects. Since, however, Sam and I do not differ in our causal powers, I see no reason at all to
suppose that events in our brains differ with respect to properties that can play such a role. In fact if one thinks, as seems reasonable to me, that differences with respect to such properties should always make for differences in causal powers, then one must accept that all such properties of events in my brain are shared by corresponding events in Sam's brain. In which case Sam must believe the same thing I do, which belies mere grade-2 internalism.

III.ii. Second response to the causal argument: wide behavior

A second common externalist response to the causal argument, as I have made it out, attacks in particular the second way I chose to make vivid the idea that the instantiation of [SAM]'S PANTS does not play any relevantly important causal role with respect to Sam's behaving as he does. That way of making the idea vivid pointed out that I, a microphysical duplicate of Sam in a microphysically identical context, will behave exactly as Sam does even though no event in the my brain has [SAM]'S PANTS.

The Fregean can deny, however, that Sam and I exhibit behavior that is of the relevantly same type. My behavior, after all, is an instance of AN's jumping into a lake, while Sam's is an instance of Sam's so jumping. Since our behaviors fall under these different types, it is reasonable to expect that at least some properties are such that their instantiation by Sam's belief plays an important causal role with respect to my behavior though no such instantiation by my belief plays any such role with respect to my behavior; and vice-versa.
The instantiation of [SAM]'S PANTS by Sam’s belief, in particular, might not play the desired important causal role with respect to Sam’s simply jumping into the lake, i.e. with respect to his jumping’s occurring; but it does, this response runs, play an extremely important causal role with respect to Sam’s behavior’s being Sam’s. After all: no one relevantly similar to Sam can exhibit behavior which shares that property with Sam’s unless a belief that causes their so behaving has [SAM]'S PANTS. This follows from two facts: first, that no one relevantly similar to Sam will jump into a lake without believing that their pants are on fire; and second, that no such behavior can belong to Sam unless it is caused by mental events belonging to Sam.12

Jerry Fodor (“A Modal Argument for Narrow Content”) has developed a complex and influential objection to this “wide behavior” response to the causal argument. The gist of Fodor’s objection, with respect to the current example, is that even though the instantiation of belonging to Sam by Sam’s belief that his pants are on fire obviously plays a very important role with respect to the resulting behavior’s being his, we should not take that role to be causal like the role played by the firecracker’s exploding with respect to the window’s shattering. I think that although Fodor’s objection demands a bit more clarification, it looks promising. Nevertheless, Stephen Yablo has suggested a simpler objection to the “wide behavior” response that should certainly be paid notice.13

---


In particular, even if the mere grade-2 internalist were somehow better-equipped than the grade-1 internalist to explain my behavior’s being AN’s and Sam’s behavior’s being Sam’s, that would not dissolve the heart of the causal argument. For that argument rested on the intuition that the instantiation of [SAM]’S PANTS by Sam’s belief that his pants are on fire does not play any important causal role with respect to Sam’s jumping into the lake. It is entirely irrelevant to that intuition whether or not the instantiation of [SAM]’S PANTS by Sam’s belief plays an important role— even, pace Fodor, an important causal role— with respect to Sam’s behavior’s being Sam’s.

In other words: perhaps the instantiation of [SAM]’S PANTS by Sam’s belief plays an important causal role with respect to some properties of the resulting behavior; but we have been shown no reason to believe that it plays any such role with respect to that behavior’s simply occurring. And that is what the causal argument demands from the mere grade-2 internalist. Not only does the content of Sam’s belief that his pants are on fire need to be had, in the circumstances, in order to cause a jumping that is Sam’s; it also has to be had, in those circumstances, in order for a jumping to occur period. And mere grade-2 internalism cannot account for that intuition.

IV. Defending grade-1 internalism

I have just argued that for several compelling reasons, mere grade-2 internalism is untenable. Thus unless grade-1 internalism is defensible, externalism is the only sensible
option. In this section I want to defend grade-1 internalism from what I take to be its most serious objections.

What distinguishes grade-1 internalism from mere grade-2 internalism is its judgment about Sam's belief that his pants are on fire. According to grade-1 internalism, the property which consists in that belief's having the content it does — i.e., SAM'S PANTS — must be shared by all events that are intrinsically indiscernible from that belief. Remembering that I am a microphysical duplicate of Sam, grade-1 internalism hence entails that a belief in my brain has SAM'S PANTS as well.

Why do internalists want to shy away from this consequence? The reason, I believe, is apparent if we assume, as I have largely been doing, that the contents of propositional attitudes are propositions — i.e. truth-apt things whose truth-values determine the truth-values of beliefs in them, the satisfaction of desires toward them, and so forth. If SAM'S PANTS, for instance, consists of a relation to the proposition Sam believes, when he believes that his pants are on fire, then my belief must bear that same relation to that same proposition. So according to grade-1 internalism, the content of Sam's belief that his pants are on fire is the same proposition as the content of my belief that my pants are on fire. This claim about propositions, I believe, is what underlies the feeling shared by externalists and internalists alike that grade-1 internalism is unacceptable.

Since Sam's pants might very well be on fire while mine are not, or vice-versa, the above consequence forces us, on pain of denying the law of non-contradiction, that
the proposition believed by both Sam and I is what I will call an "indexical proposition": it has a truth-value only relative to particular persons. It is false relative to me, which fact makes my belief false; and it is true relative to Sam, which fact makes his belief true. Indexical propositions go against the deeply-entrenched idea that all propositions bear their truth-values "absolutely", i.e. independent of who bears which attitudes toward them at which places and times. Hence if there are good reasons for insisting that propositions must bear their truth-values "absolutely", grade-1 internalism is in trouble.

One can avoid this seemingly problematic consequence of grade-1 internalism, viz. the postulation of indexical propositions, if one denies that the contents of propositional attitudes are propositions at all. Though this might sound absurd at first, it is one way of understanding David Lewis's (1979) view. Lewis proposes that the contents of propositional attitudes, the things toward which beliefs, desires and so forth are directed, are best taken to be properties. The content of Sam's belief that his pants are on fire, for instance, is the property having burning pants; and instead of bearing the belief-relation toward that property, Sam is best seen as bearing toward it the relation of self-ascription. In other words: Sam believes that his pants are on fire by virtue of self-ascribing the property having burning pants.

Lewis's view is compatible with grade-1 internalism since there is nothing stopping the content of Sam's belief from being shared by beliefs in both vat-Sam and Sam's vat-brain. In particular, both vat-Sam and Sam's vat-brain might self-ascribe the property having burning pants. But the potentially troubling conclusion of indexical
propositions does not follow. For the only analogous consequence of Lewis's view is that not all properties are instantiated "absolutely", but rather that at least some properties can be instantiated only relative particular persons. And that is entirely untroubling: it does not even make sense to think of having burning pants, for instance, as something that is instantiated independently of whether any particular person has burning pants.

I find Lewis's view of propositional attitudes implausible, however, for several reasons. Top on the list is that on Lewis's view, all of my beliefs are in some sense about me -- even ones that seem only to be about things in the world around me. When I believe that water contains hydrogen, for instance, Lewis would say that I am self-ascribing the property being such that water contains hydrogen. The implausibility of this claim becomes apparent when you see that it is exactly what Lewis would have to say about my belief that I am such that water contains hydrogen: when I have that belief, Lewis would say, I am again self-ascribing the property being such that water contains hydrogen. Hence, according to Lewis, I cannot believe that water contains hydrogen without believing that I am such that water contains hydrogen. But certainly the two beliefs seem distinct: it seems, at least on the face of it, possible for me to have either one of them without the other. Often, it seems, I am able to think about water without thinking about myself. On Lewis's view this is not possible.
Perhaps Lewis’s view is salvageable. But I want to show that the fate of grade-1 internalism does not rest on that possibility. In particular, I want to conclude by arguing that the main historical objections to indexical propositions are uncompelling.

**IV.i. Objectivity**

One of the central ideas behind the traditional resistance to indexical propositions – e.g. traceable to Frege (1918/67) -- is the admittedly vague idea that in general, the world is as it is “objectively”, e.g. independently of whom we consider it in relation to. Whether or not my pants are on fire, for instance, is independent of whom we consider them in relation to: if they are on fire relative to me, then they are on fire relative to you; and if they are not on fire relative to me, then they are not on fire relative to you.

The objection is that since indexical propositions are not true “absolutely”, but rather only relative to particular persons, the existence of indexical propositions is incompatible with the above sort of worldly objectivity. If what I believe when I believe that my pants are on fire, for instance, can be true relative to me but false relative to you – i.e. be such that my belief in it is true while your belief in it is false -- then it is but a short step to the conclusion that *my pants can be on fire* relative to me but *not* on fire relative to you. And that consequence is incompatible with the compelling idea that if my pants are on fire, they are on fire independently of whom they are considered in relation to.
This “admittedly vague” objection, however, contains a hidden fallacy that becomes apparent when spelled out more clearly. In particular, the objection seems to run as follows:

1. If my pants are on fire, then they are so “objectively”, i.e. independently of whom they are considered in relation to.

2. Hence if my pants are on fire, then my belief that my pants are on fire is true independently of whom it is considered in relation to.

3. Hence if my pants are on fire, then what I believe when I believe that my pants are on fire is true independently of whom it is considered in relation to; i.e. it is not an indexical proposition.

The fallacy lies in the step from 2 to 3: although my belief is true independently of whom it is considered in relation to, it does not follow that what I believe is true in the same way. For a defender of indexical propositions will reject the implicit linking premise, viz. that my belief is true if and only if what I believe is true. Instead, the defender of indexical propositions will say that my belief is true if and only if what I believe is true relative to me. If my belief is true if and only if what I believe is true relative to me, then 3 does not follow from 2.
IV.ii. Perry and believing-true

A more recently influential objection to indexical propositions comes from John Perry's (1979) discussion of “The Essential Indexical”. Perry's objection can be stated by supposing that the proposition I believe when I believe that my pants on fire – call it “P” -- bears a truth-value only relative to particular persons. It is true relative to me, for instance, and false relative to you, provided my pants are on fire and yours are not.

Since my belief is at least possibly true, it cannot be a belief that \( P \) is true simpliciter -- since \( P \) cannot be true simpliciter. Instead, Perry says:

Once we have adopted these new-fangled [indexical] propositions, which are only true at times for persons, we have to admit also that we believe them as true for persons at times, and not absolutely. (Perry, 1979, p. 36.)

If Perry is right, then my belief that my pants are on fire must be a belief that \( P \) is true relative to me. In other words it must be a belief in the following proposition:

\[(P_1) \text{ the proposition that } P \text{ is true relative to me.}\]

But even if I believe \( P_1 \), Perry argues, it is not that belief that plays the important causal role with respect to my jumping into the lake that my belief in \( P \) plays. For any other relevantly similar person will jump into a lake just as I did, even if they do not believe \( P_1 \). All they need to believe, instead, is that their pants are on fire – e.g., that \( P \) is true relative to them – and they will have all of the same reasons to jump into the lake as I
do. Since a belief in \( P_1 \) does not play the important causal role with respect to my behavior that my belief in \( P \) does, then, \( P \) must not be identical with \( P_1 \). And hence, by *reductio*, \( P \) is not an indexical proposition.

A clever objection. But fallacious for at least two reasons. First, as Richard Feldman (1980) points out,\(^{14}\) we need not follow Perry in his first step – namely, in moving from the claim that my belief cannot be *that \( P \) is true simpliciter* to the conclusion that my belief must be *that \( P \) is true relative to me*, i.e. that it must be a belief in \( P_1 \). For why need we assume that my belief in \( P \) is any sort of belief at all about \( P \)’s *being true*?

This step of Perry’s rests on the implicit assumption that in general, a belief in any proposition is identical with a belief *that that proposition is true*. But that assumption will be justified, as far as I can tell, only if we assume deflationism about truth. And that is a hefty assumption for the opponent of indexical propositions to make. The defender of grade-i internalism, on the contrary, can simply hold that believing a proposition is in general different from believing that the proposition is true: after all, only one of those beliefs is about a proposition and the property *truth*. My believing \( P \), then, can play an important causal role with respect to my jumping into the lake even though my believing \( P_1 \) does not.

In fact, however, Perry’s objection is fallacious even if the grade-i internalist is for some reason forced to suppose that believing \( P \) is identical to believing \( P_1 \). For in that

\(^{14}\) Sosa (1995), p. 245, makes a similar point.
case, we still must ask whether $P_1$ is itself an indexical proposition or not. And then Perry encounters the following dilemma. On the one hand, if the grade-1 internalist takes $P_1$ not to be indexical, then Perry's objection works; but any internally consistent version of grade-1 internalism will have to take $P_1$ to be indexical in the first place. On the other hand, if the grade-1 internalist avoids internal inconsistency by taking $P_1$ to be indexical, then Perry's objection fails. For in that case, my belief in $P_1$ can perfectly well play the important causal role with respect to my jumping into the lake that my belief in $P$ plays.

On the one hand, Perry's objection does work against a grade-1 internalist who takes $P_1$ not to be indexical. For if $P_1$ is not indexical, then it must be about my (AN's) pants no matter who bears an attitude toward it; and as we saw above, someone suitably similar to me can certainly have such a belief about my (AN's) pants without having any reason at all to jump into the lake. Instead, anyone suitably similar to me who has such a belief about my (AN's) pants would be more likely to shove me (AN) into the lake than to jump in herself.

However, any internally consistent grade-1 internalist who identifies $P$ with $P_1$ in the first place will have to take $P_1$ itself to be an indexical proposition. For it is straightforwardly inconsistent to believe (i) that $P$ is an indexical proposition, (ii) that $P$ is identical with $P_1$, and (iii) that $P_1$ is not an indexical proposition. Hence Perry's objection works only against a straw-man.

On the other hand, however, Perry's objection fails against an internally consistent grade-1 internalist. For as just explained, an internally consistent grade-1
internalist will take \( P_1 \) itself to be an indexical proposition; and if \( P_1 \) is an indexical proposition then my belief in it can perfectly well play the important causal role with respect to my jumping into the lake that my belief in \( P \) is supposed to play.

For if \( P_1 \) is an indexical proposition, then just like \( P \), relative to any person \( x \) who bears an attitude toward it, it will be about \( x \). When I believe \( P_1 \) I will thereby believe that \( P \) is true relative to \textit{me} (AN), while when \textit{Sam} believes \( P_1 \) he will thereby believe that \( P \) is true relative to \textit{him} (SAM). And in that case, just like my belief in \( P \), my belief in \( P_1 \) is perfectly capable of playing the desired important causal role with respect to my jumping into the lake. For anyone suitably similar to me who shares my belief in \( P_1 \) will thereby believe that \textit{their} pants are on fire, and hence will have every reason to jump into the lake just as I do.

\textbf{V. Conclusion}

Externalists believe that many mental-content properties depend as a matter of metaphysical necessity on the world external to the events that have them. Internalists often characterize their own view, in contrast, as the view that all mental-content properties are \textit{intrinsic}. That view I have called “grade-1 internalism”. But a mere denial of externalism, which several influential internalists seem to prefer, holds only that all mental-content properties are \textit{independent} of the external world. I have called that view “grade-2 internalism”.

I argued in two ways that the conjunction of grade-2 internalism with a denial of grade-1 internalism, which view I called "mere" grade-2 internalism, is untenable. On the one hand, that view becomes a version of the Fregean view that many propositional attitudes are *unsharable*; and that Fregean view is notoriously difficult to reconcile with our commonsense commitment to the possibility of communicating with others. On the other hand, one of the central arguments against externalism tells equally against mere grade-2 internalism. What I called the "causal" argument, in particular, claimed that if mere grade-2 internalism is true then many mental properties cannot play the important causal role with respect to the production of intentional behavior that we ordinarily think they do.

The upshot, then, is that any reasonable internalist should part company with Searle and others who have opted for a version of mere grade-2 internalism, and instead insist that all mental-content properties are intrinsic. (This upshot holds in spades for those internalists who are driven to internalism at least partly thanks to the causal argument against externalism – because of what I just explained about that argument telling against mere grade-2 internalism as well.)

This might be thought problematic since grade-1 internalism arguably forces us to accept that at least some propositions bear their truth-values not absolutely, but rather only relative to particular persons. However, for reasons I sketched above, that consequence is far less troubling than traditional objections make it out to be.
**Postscript: Twin-Earth and Indexical Propositions**

**O. Introduction**

The preceding chapters make a sustained case for internalism about mental content, the view that *all mental-content properties are intrinsic*. That view flies in the face of famous arguments by Hilary Putnam, Tyler Burge, and Saul Kripke that are nowadays commonly taken to prove externalism about mental content, the view that at least some perfectly ordinary mental-content properties are extrinsic. Accordingly, unless more is said, I have only posed a puzzle: there are arguments in favor of both internalism and externalism.

In this postscript I will say what I think is wrong with Putnam's, Burge's, and Kripke's arguments — or rather, in particular, what is wrong with a version of their arguments that would, if sound, prove the truth of externalism. In short, what I will call “canonical externalist arguments” rest on the assumption that no propositions are what I previously called “indexical”; but I will argue that a problem famously raised by John Perry shows the need for indexical propositions.

---

1 Putnam (1975), Burge (1979), Kripke (1980). Putnam's arguments instigated the externalist/internalist debate. Burge's arguments were designed to prove "anti-individualism" rather than mere externalism; but anti-individualism is just a subcategory of externalism as I have defined it. Much of Kripke's aim was to attack the Russellian "description-theory" of proper names, but I will consider a minor variation which leads to externalism.

2 Putnam's and Kripke's original arguments concerned *linguistic* content; but Burge helped extend the the underlying idea to mental content.
I. Canonical externalist arguments

Canonical externalist arguments begin as follows. First, we imagine some person $S$ in a particular physical, social, and historical context $C$. For Putnam, Burge, and Kripke, $C$ is very much like North America in the 1970s: it contains lakes and oceans filled with $\text{H}_2\text{O}$, competent English-speakers who use "arthritis" to denote an ailment of the joints, and is part of a possible world in which Aristotle wrote the *Nicomachean Ethics*.

Next, we imagine a person $S^*$ who is microphysically indiscernible from $S$: the two are composed of the exact same number and types of subatomic particles arranged in exactly the same way. $S^*$, furthermore, inhabits a physical, social, and historical context $C^*$ that is subtly different from $C$. For Putnam, $S^*$ is a numerically distinct but microphysically indiscernible "twin" of $S$, and $C^*$ is a distant planet that differs from Earth only in containing the strange chemical $\text{XYZ}$ wherever Earth contains $\text{H}_2\text{O}$. For Burge, $S^*$ is $S$ herself and $C^*$ is a counterfactual context in which competent English-speakers use "arthritis" to denote, in addition to arthritis, a certain muscle-tissue ailment. For Kripke,$^3$ $S^*$ is again $S$ herself, while $C^*$ is a counterfactual context in which the person from whom the person from whom ... (etc.) the person from whom $S^*$ learned the name "Aristotle" used it to dub not Aristotle but rather another guy, whom we can call "Shmaristotle".

---

$^3$ Again: the slightly modified version of Kripke who argues for externalism as I have defined it.
Finally, imagine that S believes a proposition P about a constituent of C — call it “c” — that is at the root of the difference between C and C*. For Putnam c is water, which S believes to be wet; for Burge c is arthritis, which S believes to afflict her thigh; and for Kripke c is Aristotle, whom S believes to have written the *Nicomachean Ethics*. Presumably S* believes *something*, given how similar she is to S and how similar her environment is to S’s. In particular, S* presumably believes something she would express with an utterance that’s phonetically indiscernible from the utterance with which S would express her belief: Putnam’s S* believes, as she would put it, “that water is wet”; Burge’s S* believes “that I have arthritis in my thigh”; and Kripke’s S* believes “that Aristotle wrote the *Nicomachean Ethics*”. Summarize all of this by saying that that belief of S*’s “corresponds” to S’s.

Canonical externalist arguments then run as follows:

1. S’s belief in P is about c.
2. S*’s corresponding belief is not about c.
3. Content Determines Reference (CDR): The content of a propositional attitude determines what the attitude is about.
4. Hence, P is not the propositional content of S*’s corresponding belief.

Since S and S* are microphysical duplicates, furthermore, all of S’s mental events are microphysical duplicates of S*’s corresponding mental events; so the above argument
entails that the property having content P is not necessarily shared among microphysical duplicates. Which gives us the falsity of internalism. Since the content-properties in question externalism are perfectly ordinary, furthermore, we then get externalism as I have defined it.

Premise 1 is simply assumed. Premise 2, furthermore, seems plausible when we look at the truth-conditions of S*'s belief: for Putnam, S*'s belief is true if and only if XYZ (not water) is wet; for Burge, S*'s belief is true if and only if she has either arthritis or a certain muscle-tissue ailment (not arthritis simpliciter) in her thigh; and for Kripke, S*'s belief is true if and only if Shmaristotle (not Aristotle) wrote the Nicomachean Ethics. The truth-conditions of a belief, furthermore, are plausibly determined by what the belief is about: in general, if I believe something about X, e.g. that it is F, then my belief is true if and only if X is F. Hence S*'s belief cannot be about all of the same things that S's corresponding belief is about: since the truth-conditions differ, so must the subject-matter.

Tim Crane once defended internalism by rejecting premise 2. That seems to me a last resort: who or what else could Kripke's S*'s belief could be about, for instance, other than Shmaristotle? The only alternative is that both S's and S*'s beliefs are about a single “superficial kind” of which both Aristotle and Shmaristotle are instances. But what a belief is about arguably determines its truth-conditions: I believe something about you only if how you are matters to the truth of my belief. And S*'s belief, it

---

seems, is true only if Shmaristotle, not Aristotle, wrote the Nicomachean Ethics. So $S^*$’s belief is about Shmaristotle, not Aristotle.

Premise 3, or CDR (to borrow a name from Tim Crane, 2001\(^5\)) seems plausible by example: what more do you need to do, in order to have a belief about water (arthritis, Aristotle), than to believe *that water (arthritis, Aristotle) is such-and-such*? I will argue, however, that the plausibility of CDR is illusory, and that what at least some attitudes are about is not determined solely by their propositional contents. Of course the content of any particular attitude presumably contributes to determining what the attitude is about; but, I will argue, it is metaphysically possible for one and the same proposition to be the content of two attitudes that are about different things.

II. Rejecting CDR

Crane in later writings (2001), and also Segal (2001), both resist the canonical externalist arguments by rejecting CDR: they suggest that the subject-matter of at least some propositional attitudes is determined by more than just their propositional contents. I want to defend that response, in particular in the case of what I will call “indexical” propositional attitudes. In the section following this one I will argue against CDR directly; but first I want to explain how CDR can coherently be rejected.

---

\(^5\) In its linguistic version dubbed the “meaning determines reference” principle, by Crane (1991); and neutrally dubbed the “intension determines extension” principle by John Searle (1980).
Indexical attitudes, remember, are attitudes whose contents are naturally expressed with sentences containing indexicals – such as my belief that my pants are on fire, your hope that it is now noon, and her expectation that real-estate prices around here are exorbitant. So defined, the existence of indexical attitudes is uncontentious. What I want to argue is that the propositional contents of indexical attitudes are themselves indexical – in the sense of having constituents that refer, and hence having a determinate subject-matter, only relative to a particular context (e.g. person, place, or time). The postulation of indexical propositions is certainly contentious: it is a substantive theory about what distinguishes indexical from non-indexical propositional attitudes.

An indexical proposition can be thought of as the propositional analogue of an indexical sentence: just as the sentence “I am happy here, right now” has a determinate subject-matter only relative to a person who utters it and a place and time at which the utterance occurs, so, I claim, the contents of indexical attitudes have a determinate subject-matter only relative to a person who bears some attitude toward it and a place and time at which that attitude is borne. On this view, the context of the owner of an indexical attitude (which consists of that person and the time and place at which the person bears attitude) combines with that attitude’s propositional content to determine what the attitude is about, by providing referents for certain constituents of the indexical propositions involved.

Take the proposition I believe when I believe that my pants are on fire. That proposition, much like the sentence “My pants are on fire”, contains a constituent –
namely, the concept expressed by the indexical "my" -- that does not refer "absolutely", i.e. independently of a particular context. Accordingly, not everything that my belief is about (namely, whom) is determined solely by the identity of that proposition. Instead, whom any particular belief in that proposition is about is determined by the identity of the particular believer (and the time and place of belief). When I believe that proposition, the resulting belief will be about me (at the time and place of belief); and in general, when any particular person bears an attitude toward that proposition the resulting attitude will be about that person (at the time and place of the bearing). In this way the proposition is again similar to the sentence "My pants are on fire": when I utter that sentence, the resulting utterance will be about me; and in general, any person's utterance will be about that person. Admit indexical propositions and you can safely reject CDR, the principle that the contents of all propositional attitudes are sufficient to determine what those attitudes are about.

Of course, defenders of CDR do not think that rejecting it can be "safe". In particular, in the preceding chapter I mentioned the influential neo-Fregean idea that because the truth-values of indexical propositions depend as a matter of metaphysical necessity on who bears which attitudes toward them at which places and times, accepting indexical propositions and hence rejecting CDR is incompatible with a commonsense commitment to the objectivity of the world around us. If my pants are on fire, then commonsense has it that they are on fire no matter whom they are considered in relation to: they are on fire relative to me if and only if they are on fire relative to anyone. If
what I believe when I believe that my pants are on fire is an indexical proposition, on the other hand, then it is not true or false independently of whom it is considered in relation to.

However, I argued in the preceding chapter that the neo-Fregean idea is a red herring: there is nothing wrong with accepting certain propositions to have truth-values only relative to particular persons, as long as the things those propositions are about either have or lack the properties that the propositions say they have independently of whom they are considered in relation to. Certainly, if my pants are on fire then they are so relative to anyone. Our commonsense commitment to the objectivity of the world around us requires only this, that most of what we think about be as it is independently of whom it is considered in relation to. Our commitment to objectivity does not in any similar sense require, it seems to me, that the things we think have their actual truth-values independently of whom they are considered in relation to.

Certainly, again, commonsense requires that our acts of thought have their actual truth-values independently of whom they are considered in relation to. But it does not follow that commonsense requires the same of what we think. This should be obvious from the fact that the very distinction between thoughts and what is thought – i.e. between propositional attitudes and their contents – is a philosopher's distinction, one that is consistently blurred outside of philosophical discussions.

The state of play in which we are left, then, is that canonical externalist arguments rest on a premise (CDR) that can coherently be rejected and for which the traditional
arguments are unconvincing. I pass rather quickly over those points here, since I elaborated them in more detail in the preceding chapter. Here, instead, I want to complete the internalist’s case by objecting to CDR directly.

III. The essential indexical

CDR is incompatible with the existence of indexical propositions, propositions that can be true or false only relative to a particular context (e.g. a person, place, and time). I will next argue that John Perry’s (1979) “Problem of the Essential Indexical” provides the basis for a compelling argument in favor of indexical propositions — and hence against CDR, and hence against canonical externalist arguments.

Perry’s problem begins with the plausible claim that which propositional attitudes a person has at a time in general and in a relatively natural sense explains that person’s intentional behavior, at or around that time. For instance: I come to believe that my pants are on fire, so I jump into a nearby lake. It is entirely natural to say that my belief that my pants are on fire explains my jumping into the lake, in at least the following sense: that belief is all I need to have, ceteris paribus, in order to be caused to behave as I do. Modify the situation as you will, and as long as I stand in the belief-relation to the proposition that my pants are on fire (and still desire not to get burned, am physically unconstrained, near a lake, and so on), I will jump into the lake. The same is not true about, say, my belief that 2+2=4: even if I desire not to get burned, am physically unconstrained, near a lake, and so on, I need to have more than just that belief in order to
be caused to jump into the lake. (In particular, I need to believe as well that my pants are on fire.)

Perry’s problem arises when we inquire in a bit more detail about what I believe, when I believe that my pants are on fire. Of course, what I believe is the proposition that *my pants are on fire*. But what are the identity-conditions of that proposition? On what I will call a “coarse-grained” theory of propositions, first of all, the proposition that my pants are on fire is identical with any proposition that is true in all and only possible worlds in which my (AN’s) pants are on fire. The proposition *that AN’s pants are on fire*, for instance, is identical with the proposition that *my* pants are on fire -- since I am AN. But this coarse-grained theory of propositions, Perry argues, is incompatible with the above “plausible claim” – in particular, in this case, with the idea that my belief that my pants are on fire explains my jumping into the lake.

For even if I in fact believe that AN’s pants are on fire, it is not *this* belief that explain my jumping into the lake. For this belief is not all I need to have, *ceteris paribus*, in order to be caused to behave as I do. Suppose, for instance, that I am touched by amnesia and so have no idea that I am AN. A defender of the coarse-grained theory of propositions would not describe the relevant possibility in those words, since even with amnesia I presumably know *that AN is AN* and hence, on a coarse-grained theory of propositions, know *that I* am AN. But still, they will want to describe the relevant possibility *somehow*. And no matter how one chooses to describe it, the relevant possibility shows that I might well stand in the belief-relation to the coarse-grained
proposition that AN’s pants are on fire without having any reason at all to jump into the lake – even if I still desire not to get burned, am physically unconstrained, near the lake, and so on. Instead, in the relevant case, I would be much more likely to simply pace and worry helplessly about the fate of poor AN.

On what I will call a “fine-grained” theory of propositions, what I believe when I believe that my pants are on fire is distinct from the proposition that AN’s pants are on fire. On a fine-grained theory, instead, what I believe when I believe that my pants are on fire contains something like the concept I would express by “me” (and the property owning burning pants); while the proposition that AN’s pants are on fire contains the concept I would express by “AN” (and the property owning burning pants). When I believe that my pants are on fire I conceive of myself in one way, namely as me; while when I believe that AN’s pants are on fire I conceive of myself in a different way, namely as AN. On a fine-grained theory of propositions then, it is perfectly possible for my belief that my pants are on fire to explain my jumping into the lake while my belief that AN’s pants are on fire does not. However, as Perry notes, a fine-grained theory leads straight to indexical propositions.

The reason is that the way I conceive of myself, when I believe that my pants are on fire, must be a way in which no one else can conceive of me. Take any way in which someone other than I might conceive of me: as AN; as DN’s brother; or what have you. None of these can be the way I conceive of myself, when I believe that my pants are on fire, for the same reasons that above spoke against a coarse-grained theory of
propositions. If the way I conceive of myself, when I believe that my pants are on fire, is as AN or as DN's brother, then the content of my belief must be *that AN's pants are on fire* or *that DN's brother's pants are on fire*; but again, even if I in fact believe those things, neither of *those* beliefs explains my jumping into the lake. For it is perfectly possible that I have those beliefs, along with all of my other actual propositional attitudes, and yet have no reason at all to jump into the lake. In particular I would have to have, in addition to those beliefs, the belief *that I am AN or that I am DN's brother*.

Even ways of conceiving of me that are best expressed as "you" or "that guy", furthermore, cannot be the way I conceive of myself, when I believe that my pants are on fire. For if anyone other than I conceives of me in a way they would best express as "you" or "that guy", then I can share that concept only by conceiving of myself in a way *I too* should express as "you" or "that guy": any other way of conceiving of myself would be different from the way the other person is conceiving of me. Hence if the way I conceive of myself when I believe that my pants are on fire is best expressed as "you" or "that guy", then the content of that belief must be *that your pants are on fire* or *that that guy's pants are on fire*. But again, even if I do in fact hold those beliefs, neither of *them* explains my jumping into the lake, for the same reasons as above.

Finally, if the way I conceive of myself when I believe that my pants are on fire is a way in which no one else can conceive of me, two possibilities remain. Either no one other than I can entertain the concept by which I conceive of myself, or else anyone other than I who entertains that same concept must thereby conceive of someone other than me.
The first option implies what in the preceding chapter I called a "Fregean view" of indexical attitudes, the view that attitudes like my belief that my pants are on fire cannot be had by more than one person. For if no one other than I can conceive of me as I do when I believe that my pants are on fire, and if propositions are composed of ways of conceiving of the things and properties they are about, then no one other than I can believe what I do when I believe that my pants are on fire. In the preceding chapter, furthermore, I reviewed what I take to be a few serious objections to that Fregean view. Hence we are left with the conclusion that anyone other than I who entertains the concept by which I conceive of myself, when I believe that my pants are on fire, must thereby conceive of someone other than me.

In fact it is obvious who we should say such a person conceives of: herself. On the view Perry's problem has taken us to, when any person x entertains the concept by which I conceive of myself when I believe that my pants are on fire, she thereby conceives of x. Accordingly, the proposition I believe, when I believe that my pants are on fire, is the same proposition that you believe when you believe your pants to be on fire. Since my pants might well be on fire while yours are not, furthermore, my belief might well be true while yours is false. Which means that the proposition we believe is true or false only relative to a particular context (person, place, or time); i.e. it is indexical.
IV. Perry’s character

That is one interpretation of how Perry’s “Problem of the Essential Indexical” leads straightforwardly to indexical propositions. Perry himself hopes to resist indexical propositions by simply denying the “plausible claim” with which his problem begins, viz. that which propositional attitudes a person has at a time explains her intentional behavior at or around that time. Perry might accept that this is sometimes the case: a desire for rain and a belief in the efficacy of a certain rain-dance, for instance, might well explain their owner’s dancing. But at least in the case of indexical propositional attitudes, Perry insists, the situation is different.

It is not my belief that my pants are on fire, for instance, that explains my jumping into the lake. By denying that plausible claim, Perry removes the immediate obstacle to adopting a coarse-grained view of propositions. For by denying that my belief that my pants are on fire explains my jumping into the lake, Perry can accept that that belief is identical with the belief that AN’s pants are on fire. Thus he can give identity conditions for the proposition I believe without accepting indexical propositions. The question Perry then has to answer, however, is “What exactly does explain my jumping into the lake, if not my belief that my pants are on fire?” If Perry cannot point out something in the general vicinity of my belief which can do the needed explanatory work, then his view will be no more plausible than the claim he asks us to reject. That will not necessarily mean that Perry’s view is false. But we will have no more reason to accept Perry’s view than to reject it.
Perry's answer to the question "What does explain my jumping into the lake?" relies on an extension of what David Kaplan calls "character". The Kaplanian character of a sentence, first of all, is a function from contexts (persons, places, or times) to coarse-grained propositions consisting of the things and properties utterances of that sentence are about. The character of "my pants are on fire", for instance, is the function that maps every speaker x onto the coarse-grained proposition that is true in all and only those possible worlds in which x's pants are on fire. The character of "AN's pants are on fire", on the other hand, maps all speakers onto the coarse-grained proposition that is true in all and only those possible worlds in which AN's pants are on fire.

Analogously, what Perry calls propositional-attitude "states" are functions from contexts to attitudes borne toward coarse-grained propositions. When we would ordinarily say that my behavior is explained by my belief that my pants are on fire, according to Perry what we should instead say is that it is explained by my being in a particular belief-state: namely, the belief-state consisting of the function that maps any believer x onto a belief in the coarse-grained proposition that is true in all and only those possible worlds in which x's pants are on fire. When we would ordinarily say that my behavior is not explained by my belief that AN's pants are on fire, on the other hand, what we should instead say is that it is not explained by my being in a different belief-state: namely, the belief-state consisting of the constant function that maps all believers onto a belief in the coarse-grained proposition that is true in all and only those possible worlds in which AN's pants are on fire.
Return to the case I described above in which, although I believe that my pants are on fire, I am "touched by amnesia and so have no idea that I am AN." What explains my jumping into the lake in that case, Perry agrees, is not my belief. For it is perfectly possible, he agrees, for me to believe what I do – namely, the coarse-grained proposition that is true in all and only those possible worlds in which AN’s pants are on fire -- without having any reason at all to jump into the lake. In particular, if I were to believe what I do by virtue of instantiating anything other than the function that maps any believer x onto a belief in the coarse-grained proposition that is true in all and only those possible worlds in which x’s pants are on fire, I would not jump into the lake. Instead, in order to be caused to jump into the lake I have to instantiate that function: i.e., I have to be in the belief-state which would determine, in any person x, a belief that x’s pants are on fire.

VI. A flaw in character

Both Kaplan’s notion character and Perry’s related notion of propositional-attitude state are representative of what is a very popular way of conceiving of the mental causes of intentional behavior. That Perry’s notion of propositional-attitude state is not up to the job of explaining intentional behavior, however, can be seen by developing a
point made by Gabriel Segal (2001):\(^6\) propositional-attitudes states, just like coarse-grained propositions, are too coarse-grained to do the needed explanatory work.

To develop one of Segal’s own examples, imagine that the following is true. I believe that today is Friday, but I am ignorant of the order of the days of the week: I do not know, for instance, that Friday follows Thursday. Knowing that yesterday was Grandma’s birthday, however, and knowing that Grandma does not accept birthday-cards that are more than a day late, I rush to the post-office.

Ordinarily, we would say that my belief that yesterday was Grandma’s birthday explains my rushing to the post-office. Perry, however, denies this: I could perfectly well bear the belief-relation toward that same coarse-grained proposition, Perry says, but have no reason at all to go to the post-office. For instance: although yesterday was August 20\(^{th}\), simply believing that Grandma’s birthday is August 20\(^{th}\) will not suffice to send me to the post-office; for I might also believe that today is August 25\(^{th}\) and hence only sigh with despair, believing that I have missed Grandma’s birthday yet again. Instead, Perry tells us that what explains my rushing to the post-office is my belief-state, i.e. my instantiating the function that maps any day x onto a belief in the coarse-grained proposition that Grandma’s birthday is the day before x. According to Perry, that belief-state is all I need to be in, ceteris paribus, in order to be caused to behave as I do: modify the situation as you will, and as long as I instantiate that same function, (and still

\(^6\) Segal (2001), ch. 4, sec. 2.
believe that Grandma doesn’t accept birthday-cards that are more than a day late, am
physically unconstrained, and so on), I will rush to the post-office.

Perry is wrong, however: what explains my behavior cannot be that belief-state,
for the same reason it cannot be my belief in the coarse-grained proposition that
Grandma’s birthday is August 20th. To see this, imagine a counterfactual scenario in
which we would ordinarily say that instead of believing that Grandma’s birthday was
yesterday, I believe only

\[(P_1) \quad \text{that Grandma’s birthday was either yesterday or, if today is Friday, then}
\]
\[\quad \text{the most recent Thursday.}\]

(Perhaps that bit of knowledge is all that remains in my mind from a previous glance at
the calendar.)

Since yesterday is the most recent Thursday as long as today is Friday, the
function that determines my belief in the proposition just mentioned is the function that
maps any day \(x\) onto a belief in the proposition that Grandma’s birthday is the day before
\(x\). On any day on which I instantiate the function by virtue of which I have the above
belief, I will thereby believe that Grandma’s birthday is the previous day. Hence that
same function is just the function that determines my actual belief Grandma’s birthday
was yesterday. And hence in the imagined counterfactual scenario I would instantiate all
of the same functions from believers to beliefs in coarse-grained propositions that I actually instantiate.

Since I believe that today is Friday, however, and do not know that the day before Friday is Thursday, in the imagined counterfactual scenario I will have no reason to rush to the post-office. Instead, I might again only sigh with despair, believing that I have missed Grandma’s birthday: her birthday was the most recent Thursday; but for all I know, that could have been several days ago. Accordingly, my belief-state is not all I need to have, ceteris paribus, in order to be caused to behave as I do: for it is perfectly possible for me to be in that belief-state without having any reason at all to rush to the post-office (even if I still believe that Grandma doesn’t accept birthday-cards that are more than a day late, am physically unconstrained, and so on).

Once Segal’s basic strategy is understood, similar counterexamples can be found to Perry’s view of my jumping into the lake. According to Perry, what explains my jumping into the lake is my instantiating the function that maps any believer x onto a belief in the coarse-grained proposition that x’s pants are on fire. But I would instantiate that same function were I to instead believe any of what we would ordinarily describe in the following ways:

\[(P_2) \text{ that my pants are on fire unless there is a largest prime number}\]
In other words: since there is no largest prime, since I must be the person 4 feet to the left of the place 4 feet to the right of me, and since AN is necessarily born of sperm S and egg E, the function that would determine my belief in any of P₂, P₃, or P₄ is the function that maps any believer x onto a belief in the proposition that x’s pants are on fire: anyone who were to instantiate the function by virtue of which I believed any of the above beliefs would thereby believe a proposition that ascribes to their pants the property being on fire. Hence that same function is just the function that determines my actual belief that my pants are on fire. And hence, were I to believe any of P₂, P₃, or P₄, I would instantiate all of the same functions from believers to beliefs in coarse-grained propositions that I instantiate when I believe that my pants are on fire.

But we would not ordinarily say that my belief in any of P₂, P₃, or P₄ explained my jumping into the lake, since I could easily believe any of P₂, P₃, or P₄ without having any reason at all to jump into the lake. Believing P₂, first, will give me no reason to jump into the lake if I believe that there is a largest prime number. Believing P₃, next, will give

(P₃)  that the pants of the person 4 feet to the left of the place 4 feet to the right of me are on fire

(P₄)  that either my pants are on fire or, if I am AN, then the pants of the person born of sperm S and egg E are on fire [where AN is necessarily born of sperm S and egg E].
me no reason to jump into the lake as long as I do not infer \textit{that I am} 4 feet to the left of the person 4 feet to the right of me. Believing P, finally, will give me no reason to jump into the lake unless I know that \(AN\) is necessarily born of sperm \(S\) and egg \(E\).

Accordingly, my belief-state is not all I need to have, \textit{ceteris paribus}, in order to be caused to behave as I do: for it is perfectly possible for me to be in that belief-state without having any reason at all to jump into the lake (even if I still desire not to get burned, am physically unconstrained, near the lake, and so on). Again, therefore, Perry's propositional-attitude states are too coarse-grained to explain intentional behavior.

\textbf{VII. Conclusion}

If indexical propositional attitudes have propositional contents that are themselves indexical, then CDR is false. In which case the canonical externalist arguments, as I have made them out, contain a false premise.

Externalists could, of course, insist that even if CDR is false, still a \textit{restricted} version of CDR is true which is sufficient to let at least some canonical externalist arguments go through. For instance: Putnam could insist that even if the subject-matter of \textit{indexical} propositional attitudes is not determined solely by their propositional contents, still what the belief \textit{that water is wet} is about (namely, water) is determined solely by its propositional content. And Burge could insist the same for anyone's belief \textit{that she has arthritis in her thigh}; and Kripke could insist the same for the belief \textit{that Aristotle wrote the Nicomachean Ethics}. This line of response would block the objection
to canonical externalist arguments I have just made out. But I can give two reasons why I find the response unattractive.

First, the common belief that \textit{at least some} propositional contents determine what attitudes toward them are about is usually motivated by considerations of a general character, considerations which equally well motivate the common belief that \textit{all} propositional contents are of that sort. I mentioned some of these considerations in the preceding chapter. What I previously called a "commonsense commitment to objectivity", for instance, speaks not only in favor of the idea that what I believe when I believe that water is wet is true or false independent of whom it is considered in relation to, but also and equally well in favor of the idea that what I believe when I believe that my pants are on fire is true or false independent of whom it is considered in relation to. To the extent that we should believe any version of CDR at all, then, we should believe CDR itself. Or conversely: to the extent that the objection I developed above should lead us to reject CDR, to that same extent it should lead us to reject what I have just called a "restricted" version of CDR.

Second, even a "restricted" version of CDR -- a version which admits indexical propositions only as contents of indexical attitudes -- will not save the canonical externalist arguments if the particular propositional attitudes mentioned in those arguments turn out to be, despite first appearances, indexical. In particular, one might think that my belief that water is wet is \textit{implicitly} indexical in the following sense: although its content would not ordinarily be expressed by a sentence containing
indexicals, such a sentence should be used when spelling out its content in more detail. When I believe that water is wet I might be said to believe, for instance, that the stuff filling the oceans and lakes around here is wet. One might similarly think that when I believe that I have arthritis in my thigh, what I believe is something like that what medical experts in my socio-linguistic community call “arthritis” afflicts my thigh; and one might think that when I believe that Aristotle wrote the Nicomachean Ethics, what I believe is something like that the ancient Greek guy whom I heard about in philosophy class wrote the Nicomachean Ethics.

The idea that the sorts of propositional attitudes mentioned in are implicitly indexical is not uncommon. It was Putnam’s original view in the paper in which he put forth his famous Twin-Earth argument, for one -- at least with respect to beliefs about natural kinds like water. And Michael Dummett has since defended a similar claim with respect to beliefs about things and people we refer to with proper names like “Aristotle”. If claims like these are true, furthermore, then a restricted version of CDR that admits indexical propositions as contents of indexical attitudes will have to admit them as contents of the attitudes mentioned in at least some of the canonical externalist arguments. Which will spell failure for even a version of those canonical arguments that relies on only a restricted version of CDR.

My conclusion, then, is that at the very least, the outlook for canonical externalist arguments is dimmer than it is normally believed to be. On the one hand, the most straightforward expression of those arguments relies on a premise (CDR) which can
coherently be rejected, for which traditional arguments are uncompelling, and that Perry’s “Problem of the Essential Indexical”, despite common opinion, speaks directly against. And on the other hand, versions of canonical externalist arguments which instead relied on something weaker than CDR would be largely unmotivated and are incompatible with the common idea that many ordinary propositional attitudes are implicitly indexical. Combined with the anti-externalist arguments developed in the preceding chapters, all of this seems to me to add up to a strong case for internalism about mental content.
References


--------------- (1992): “Mental Causation and Mental Reality”, *Proceedings of the
References


References

(1975b): "Philosophy and our mental life", reprinted in N. Block (ed.)
University Press), pp. 134-44.


Schiffer, Stephen (1990): "Fodor's character", in E. Villanueva (ed.), *Information,


