PHILOSOPHY OF MIND

THIRD EDITION

JAEGWON KIM

A Member of the Perseus Books Group

Westview Press was founded in 1975 in Boulder, Colorado, by notable publisher and intellectual Fred Praeger. Westview Press continues to publish scholarly titles and high-quality undergraduate- and graduate-level textbooks in core social science disciplines. With books developed, written, and edited with the needs of serious nonfiction readers, professors, and students in mind, Westview Press honors its long history of publishing books that matter.

Published by Westview Press,
A Member of the Perseus Books Group

Find us on the World Wide Web at www.westviewpress.com.

Westview Press books are available at special discounts for bulk purchases in the United States by corporations, institutions, and other organizations. For more information, please contact the Special Markets Department at the Perseus Books Group, 2300 Chestnut Street, Suite 200, Philadelphia, PA 19103, or call (800) 810-4145, ext. 5000, or e-mail special.markets@perseusbooks.com.

Designed by Trish Wilkinson
Set in 10.5 point Minion Pro

Library of Congress Cataloging-in-Publication Data

Kim, Jaegwon.
Philosophy of mind / Jaegwon Kim.—3rd ed.
p. cm.
ISBN 978-0-8133-4458-4 (alk. paper)
1. Philosophy of mind. I. Title.
BD418.3.K54 2011
128'.2—dc22
E-book ISBN 978-0-8133-4520-8 2010040944

10 9 8 7 6 5 4 3 2 1

CHAPTER 8

Mental Content

You hope that it will be warmer tomorrow, and I believe that it will be. But Mary doubts it and hopes that she is right. Here we have various "intentional" (or "content-bearing" or "content-carrying") states: your *hoping* that it will be warmer tomorrow, my *believing*, and Mary's *doubting*, that it will be so. All of these states, though they are states of different persons and involve different *attitudes* (believing, hoping, and doubting), have the same *content*: the proposition that it will be warmer tomorrow, expressed by the embedded sentence "it will be warmer tomorrow." This content *represents* a certain state of affairs, its being warmer tomorrow. Different subjects can adopt the same intentional attitude toward it, and the same subject can have different attitudes toward it (for example, you believe it and are pleased about it; later you come to disbelieve it).

But how do these intentional states, or propositional attitudes, come to have the content they have and represent the state of affairs they represent? More specifically, what makes it the case that your hope and my belief have the same content? There is a simple, and not wholly uninformative, answer: Because they each have the content expressed by the same content sentence "it will be warmer tomorrow." But then a more substantive question awaits us: What is it about your hope and my belief that makes it the case that the same sentence can capture their content? We do not expect it to be a brute fact about these mental states that they have the content they have or that they share the same content; there must be an explanation. These are the basic questions about mental content.

The questions can be raised another way. It is not just persons who have mental states with content. All sorts of animals perceive their surroundings through their perceptual systems, process information gained thereby, and use it in coping with things and events around them. We humans do this in our own distinctive ways, though perhaps not in ways that are fundamentally different

from those of other higher species of animals. It seems, then, that certain physical-biological states of organisms, presumably states of their brains or nervous systems, can carry information about their surroundings, representing them as being this way or that way (for example, here is a red apple, or a large, brown, bear-shaped hulk is approaching from the left), and that processing and using these representations in appropriate ways is highly important to their surviving and flourishing in their environments. These physical-biological states have representational content—they are *about* things, inside or outside an organism, and *represent them as being a certain way*. In a word, these states have *meanings*: A neural state that represents a bear as approaching *means* that a bear is approaching. But how do neural-physical states come to have meanings—and come to have the particular meanings that they have? Just what is it about a configuration of nerve fibers or a pattern of their activation that makes it carry the content "there is a red apple on the table" rather than, say, "there are cows in Canada," or perhaps nothing at all?

This question about the nature of mental content has a companion question, a question about how contents are *attributed* to the mental states of persons and other intentional systems. We routinely ascribe states with content to persons, animals, and even some nonbiological systems. If we had no such practice—if we were to stop attributing to people around us beliefs, desires, emotions, and the like—our communal life would surely suffer a massive collapse. There would be little understanding or anticipating of what other people will do, and this would seriously undermine interpersonal interactions. Moreover, it is by attributing these states to ourselves that we come to understand ourselves as cognizers and agents. A capacity for self-attribution of beliefs, desires, intentions, and the rest is arguably a precondition of personhood. Moreover, we often attribute such states to nonhuman animals and sometimes even to purely mechanical or electronic systems. (Even such humble devices as supermarket doors are said to "see that a customer is approaching.") What makes it possible for us to attribute content-carrying states to persons and other organisms? What procedures and principles do we follow when we do this? According to some philosophers, the two questions, one about the nature of mental content and the other about its attribution, are intimately connected.

Interpretation Theory

Suppose you are a field anthropologist-linguist visiting a tribe of people never before visited by an outsider. Your project is to find out what these people believe, remember, desire, fear, hope, and so on, and to be able to understand their

speech. That is, your project is to map their "notional world" and develop a grammar and dictionary for their language. So your job involves two tasks: first, interpreting their minds, to find out what they believe, desire, and so on; and, second, interpreting their speech, to determine what their utterances mean. This is the project of "radical interpretation": You are to construct an interpretation of the natives' speech and their minds from scratch, based on your observation of their behavior and their environment, without the aid of a native translator-informant or a dictionary. (This is what makes it "radical" interpretation.)[1]

Brief reflection shows that the twin tasks are interconnected and interdependent. In particular, belief, among all mental states, can be seen to hold the key to radical interpretation: It is the crucial link between a speaker's utterances and their meanings. If a native speaker sincerely asserts sentence S (or more broadly, "holds S true," as Donald Davidson says) and S means that there goes a rabbit, then the speaker believes that there goes a rabbit, and in asserting S she expresses her belief that there goes a rabbit. Conversely, if the speaker believes that there goes a rabbit and uses sentence S to express this belief, then S means that there goes a rabbit. If you knew how to interpret the natives' speech, it would be a simple matter to find out what they believe: All you would need to do is observe their speech behavior—their assertions, denials, and so on. Similarly, if you had knowledge of what belief a native is expressing by uttering S on a given occasion, you know what S, as a sentence of her language, means. When you begin, you have knowledge of neither her beliefs nor her meanings, and your project is to secure them both through your observation of how she behaves in her environment. There are, then, three variables involved: behavior, belief, and meaning. Through observation, you have access to one of them, behavior. Your task is to solve for the two unknowns, belief and meaning. How is this possible? Where do you start?

Karl is one of the subjects you are trying to interpret. Suppose you observe that Karl affirmatively utters, or holds true,[2] the sentence "Es regnet" when, and only when, it is raining in his vicinity. (This is highly idealized, but the main point should apply, with suitable provisos, to real-life situations.) You observe a similar behavior pattern in many others in Karl's speech community, and you are led to posit the following proposition:

> (R) Speakers of language L (Karl's language) utter "Es regnet" at time *t* if and only if it is raining at *t* in their vicinity.

So we are taking (R) to be something we can empirically establish by observing the behavior, in particular, speech behavior, of our subjects in the context

of what is happening in their immediate environment. Assuming, then, that we have (R) in hand, it would be natural to entertain the following two hypotheses:

> (S) In language L, "Es regnet" means that it is raining (in the speaker's vicinity).
>
> (M) When speakers of L utter "Es regnet," this indicates that they believe that it is raining (in their vicinity) and they use "Es regnet" to express this belief.

In this way you get your first toehold in the language and minds of the natives, and something like this seems like the only way.

These hypotheses, (S) and (M), are natural and plausible. But what makes them so? What sanctions the move from (R) to (S) and (M)? When you observe Karl uttering the words "Es regnet," you see yourself that it is raining out there. You have determined observationally that Karl is expressing a belief about the current condition of the weather. This assumption is reinforced when you observe him, and others in his speech community, do this time after time. But what belief is Karl expressing when he makes this utterance? What is the content of the belief that Karl expresses when he says "Es regnet"? Answering this question is the crux of the interpretive project. The obvious answer seems to be that Karl's belief has the content "it is raining." But why? Why not the belief with the content "it is a sunny day" or "it is snowing"? What are the tacit principles that help to rule out these possibilities?

You attribute the content "it is raining" to Karl's belief *because you assume that his belief is true*. You know that his belief is about the weather outside, and you see that it is raining. What you need, and all you need, to get to the conclusion that his belief has the content "it is raining" is the further premise that his belief is true. In general, then, what you need is the famous "charity principle":

> *Principle of Charity*. Speakers' beliefs are by and large true. (Moreover, they are largely correct in making inferences and rational in forming expectations and making decisions.)[3]

With this principle in hand, we can make sense of the transition from (R) to (S) and (M) in the following way:

> In uttering "Es regnet," Karl is expressing a belief about the current weather condition in his vicinity, and we assume, by the charity principle, that this belief is true. The current weather condition is that it is raining. So Karl's belief has the content that it is raining, and he is using the sentence "Es regnet" to express this belief (M), whence it further follows that "Es regnet" means that it is raining (S).

We do not attribute the content "it is clear and sunny" or "it is snowing" because that would make Karl's and his friends' beliefs about whether it is raining around them almost invariably, and unaccountably, false. There is no logical contradiction in the idea that a group of speakers are almost always wrong about rains in their vicinity, but it is not something that can be taken seriously. We would have to posit serious, and unexplainable, cognitive deficits in Karl and his friends, and this is not a reasonable possibility. For one thing, they seem able to cope with their surroundings, including good and bad weather, as well as we do.

Clearly, the same points apply to interpreting utterances about colors, shapes, and other observable properties of objects and events around Karl. When Karl and his friends invariably respond with "Rot" when we show them cherries, ripe tomatoes, and McIntosh apples and withhold it when they are shown lemons, eggplants, and snowballs, it would make no sense to speculate that "rot" might mean *green*, that Karl and his friends systematically misperceive colors, and that in consequence they have massively erroneous beliefs about the colors of objects around them. The only plausible thing to say is that "rot" means *red* in Karl's language and that Karl is expressing the (true) belief that the apple held in front of him is red. All this is not to say that our speakers never have false beliefs about colors or about anything else; they may have them in huge numbers. But unless we assume that their beliefs, especially those about the manifestly observable properties of things and events around them, are largely correct, we have no hope of gaining entry into their notional world.

So what happens is that we interpret the speakers in such a way as to credit them with beliefs that are by and large true and coherent. But since *we* are doing the interpreting, this in effect means *true and coherent by our light*. Under our interpretation, therefore, our subjects come out with *beliefs that are largely in agreement with our own*. The attribution of a system of beliefs and other intentional states is essential to the understanding of other people, of what they say and do. From all this an interesting conclusion follows: We can

interpret and understand only those people whose belief systems are largely like our own.

The charity principle therefore rules out, a priori, interpretations that attribute to our subjects beliefs that are mostly false or incoherent; any interpretive scheme according to which our subjects' beliefs are massively false or manifestly inconsistent (for example, they come out believing that there are round squares) cannot, for that very reason, be a correct interpretation. Further, we can think of a generalized charity principle that enjoins us to interpret all of our subjects' intentional states, including desires, aversions, hopes, fears, and the rest, in a way that renders them maximally coherent and intelligible among themselves and in relation to the subjects' actions and behaviors.

But we should note the following important point: There is no reason to think that in any interpretive project there is a single unique interpretation that best meets this requirement. This is evident when we reflect on the fact that the charity principle requires only that the entire *system* of beliefs attributed to a subject be by and large true but it does not tell us which of her beliefs must come out true. In practice as well as in theory, there are likely to be ties, or unstable near-ties, among possible interpretations: That is, we are likely to end up with more than one maximally true, coherent, and rational scheme of interpretation that can explain all the observational data. (This phenomenon is called "indeterminacy of interpretation.") We can appreciate such a possibility when we note that our criteria of coherence and rationality are bound to be somewhat vague and imprecise (in fact, this is probably necessary to ensure their flexible application to a wide and unpredictable range of situations) and that their applications to specific situations are likely to be fraught with ambiguities. At any rate, it is easy to see how interpretational indeterminacy can arise by considering a simple example.

We observe Karl gorging on raw spinach leaves. Why is he doing that? We can see that there are indefinitely many belief-desire pairs that we could attribute to Karl that would explain why he is eating raw spinach. The following are only some of the possibilities:

Karl believes that eating raw spinach will improve his stamina, and he wants to improve his stamina.

Karl believes that eating raw spinach will help him get rid of his bad breath, and he has been very self-conscious about his breath.

Karl believes that eating raw spinach will please his mother, and he will do anything to make her happy.

> Karl believes that eating raw spinach will annoy his mother, and he will go to any length to annoy her.

You get the idea: This can go on without end. We can expect many of these potential explanations to be excluded by further observation of Karl's behavior and by consideration of coherence with other beliefs and desires that we want to attribute to him. But it is difficult to imagine that this will eliminate all but one of the indefinitely many possible belief-desire pairs that can explain Karl's spinach eating. Moreover, it is likely that any one of these pairs could be protected no matter what if we were willing to make drastic enough adjustments elsewhere in Karl's total system of beliefs, desires, and other mental states.

Suppose, then, that there are two interpretive schemes of Karl's mental states that, as far as we can determine, satisfy the charity principle to the same degree and work equally well in explaining his behavior. Suppose further that one of these systems attributes to Karl the belief that eating raw spinach is good for one's stamina, and the second instead attributes to him the belief that eating spinach will please his mother. As far as interpretation theory goes, the schemes are in a tie, and neither could be pronounced to be superior to the other. But what is the fact of the matter concerning Karl's belief system? Does he or doesn't he believe that eating raw spinach improves stamina?

There are two possible approaches we could take in response to these questions. The first is to take interpretation as the rock-bottom foundation of content-carrying mental states by embracing a principle like this:

> For S to have the belief that *p* is for that belief to be part of the best (most coherent, maximally true, and so on) interpretive scheme of S's total system of propositional attitudes (including beliefs, desires, and the rest). There is no further fact of the matter about whether S believes that *p*.

It will be natural to generalize this principle so that it applies to all propositional attitudes, not just beliefs. On this principle, then, interpretation is *constitutive* of intentionality; it is what ultimately determines whether any supposed belief exists.[4] Interpretation is not merely a procedure for finding out what Karl believes. This constitutive view of interpretation, when combined with the indeterminacy of interpretation, can be seen to have some apparently puzzling consequences. Suppose that several interpretive schemes are tied for first and the belief that *p* is an element of some but not all of these schemes. In such

a case we would have to conclude that there is no fact of the matter about whether Karl has this belief. Whether Karl believes that *p* therefore is a question without a determinate answer. To be sure, the question about this particular bit of belief may be settled by further observation of Karl; however, indeterminacies are almost certain to remain even when all the observations are in. (Surely, at some point after Karl's death, there is nothing further to observe that will be relevant!) Some will see in this kind of position a form of *content irrealism*. If beliefs are among the objectively existing entities of the world, either Karl believes that raw spinach is good for his stamina or he does not. There must be a fact about the existence of this belief, independent of any interpretive scheme that someone might construct for Karl. So if the existence of beliefs is genuinely indeterminate, we would have to conclude, it seems, that beliefs are not part of objective reality. Evidently, the same conclusion would apply to all intentional states.[5]

An alternative line of consideration can lead to *content relativism* rather than content irrealism: Instead of accepting the indeterminacy of belief, we might hold that whether a given belief exists is *relative to a scheme of interpretation*. It is not a question that can be answered absolutely, independently of a choice of an interpretive scheme. Whether Karl has that particular belief depends on the interpretive theory relative to which we view Karl's belief system. But a relativism of this kind is not free from difficulties either. What is it for a belief to "exist relative to a scheme" to begin with? Is it anything more than "the scheme attributes the belief to Karl"? If so, shouldn't we ask the further question whether what the scheme says is *correct*? But this takes us right back to the nonrelativized notion of belief existence. Moreover, is all existence relative to some scheme or other, or is it just the existence of belief and other propositional attitudes that is relative in this way? Either way, many more questions and puzzles await us.

There is a further point to think about: Interpretation involves an interpreter, and the interpreter herself is an intentional system, a person with beliefs, desires, and so forth. How do we account for *her* beliefs and desires—*how do her intentional states get their contents*? And when she tries to maximize agreement between her beliefs and her subject's beliefs, how does she know what she believes? That is, *how is self-interpretation possible*? Don't we need an account of how we can know the contents of our own beliefs and desires? Do we just look inward, and are they just there for us to "see"? Or do we need to be interpreted by a third person if we are to have beliefs and meaningful speech? It is clear that the interpretation approach to mental content must, on pain of circularity, confront the issue of self-interpretation.

All this may lead you to reject both the constitutive and the relativist views of interpretation and pull you toward a realist position about intentional states, which insists that there is a fact of the matter about the existence of Karl's belief about spinach that is independent of any interpretive schemes. If Karl is a real and genuine believer, there must be a determinate answer to the question whether he has this belief. Whether someone happens to be interpreting Karl, or what any interpretive scheme says about Karl's belief system, should be entirely irrelevant to that question. This is content realism, a position that views interpretation only as a way of finding out something about Karl's belief system, not as constitutive of it. Interpretation therefore is given only an epistemological function, that of ascertaining what intentional states a given subject has; it does not have the ontological role of grounding their existence.

You may find content realism appealing. If so, there is more work to do; you must provide an alternative realist account of what constitutes the content of intentional states. It is only if you take the constitutive view of interpretation that interpretation theory gives you a solution to the problem of mental content—that is, an answer to the question "How does a belief get to have the content it has?"

The Causal-Correlational Approach: Informational Semantics

A fly flits across a frog's visual field, and the frog's tongue darts out, snaring the fly. The content of the frog's visual perception is a moving fly (which is a complicated way of saying that the frog sees a moving fly). Suppose now that in a world pretty much like our own (this could be some remote region of this world), frogs that are like our frogs exist but there are no flies. Instead there are "schmies," very small lizards roughly the size, shape, and color of earthly flies, and they fly around just the way our flies do and are found in the kind of habitat that our flies inhabit. In that world frogs feed on schmies, not flies. Now, in this other world, a schmy flits across a frog's visual field, and the frog flicks out its tongue and catches it. What is the content of this frog's visual perception? What does the frog's visual percept represent? The answer: a moving schmy.

From the frogs' "internal," or "subjective," perspectives, there is no difference, we may suppose, between our frog's perceptual state and the other-worldly frog's perceptual state: Both register a black speck flitting across the visual field. However, we attribute different contents to them, and the difference lies outside the frogs' perceptual systems; it is a difference in the kind of object that stands in a certain relationship to the perceptual states of the frogs. It is not only that in

these particular instances a fly caused the perceptual state of our frog and a schmy caused a corresponding state in the other-worldly frog; there is also a more general fact, namely, that the habitat of earthly frogs includes flies, not schmies, and it is flies, not schmies, with which they are in daily perceptual and other causal contact. The converse is the case with other-worldly frogs and schmies. Our frogs' perceptual episodes involving a flitting black speck *indicate*, or *mean*, the presence of a fly; qualitatively indistinguishable perceptual episodes in other-worldly frogs *indicate* the presence of a schmy.

Consider a mercury thermometer: The height of the column of mercury indicates the ambient air temperature. When the thermometer registers 32°C, we say, "The thermometer says that the temperature is 32°C"; we also say that the current state of the thermometer carries the information that the air temperature is 32°C. Why? Because there is a lawful correlation—in fact, a causal connection—between the state of the thermometer (that is, the height of its mercury column) and air temperature. It is for that reason that the device is a thermometer, something that carries *information* about ambient temperature.

Suppose that under normal conditions a certain state of an organism covaries regularly and reliably with the presence of a horse. That is, this state occurs in you when, and only when, a horse is present in your vicinity (and you are awake and alert, sufficient illumination is present, you are appropriately oriented in relation to the horse, and so on). The occurrence of this state, then, can serve as an *indicator*[6] of the presence of a horse; it carries the information "horse" (or "a horse is out there"). And it seems appropriate to say that this state *indicates* or *represents* the presence of a horse and has it as its content. The suggestion is that something like this account works for intentional content in general, and this is the basic idea of the causal-correlational approach. (The term "causal" is used because on some accounts based on this approach, the presence of horses is supposed to cause the internal "horse-indicator" state.)

The strategy seems to work well with contents of perceptual states, as we saw in the fly-schmy case. I perceive red, and my perceptual state has "red" as its content because I am having the kind of perceptual experience typically correlated with—in fact, caused by—the presence of a red object. Whether I perceive red or green has little to do with the intrinsic experienced qualities of which I am conscious; rather, it depends essentially on the properties of the objects with which I am in causal-correlational relations. Those internal states that are typically caused by red objects, or that lawfully correlate with the presence of red objects nearby, have the content "red" for that very reason, not because of any of their intrinsic properties. Two thermometers of very different construction—

say, a mercury thermometer and a gas thermometer—both represent the temperature to be 30°C in spite of the fact that the internal states of the two thermometers that covary with temperature—the height of a column of mercury in the first and the pressure of a gas in the second—are different. In a similar way, two creatures, belonging to physiologically quite diverse species, can both have the belief that there are red fruits on the tree. The causal-correlational approach to content, also called informational semantics, has been influential; it explains mental content in a naturalistic way and seems considerably simpler than the interpretational approach considered earlier.

How well does this approach work with intentional states in general? We may consider a simple version of this approach, perhaps something like this:[7]

> (C) Subject S has the belief with content *p* (that is, S believes that *p*) just in case, under optimal conditions, S has this belief (as an occurrent belief)[8] if and only if *p* obtains.

To make (C) at all viable, we should restrict it to cases of "observational beliefs"—beliefs about matters that are perceptually observable to S. For (C) is obviously implausible when applied to beliefs like the belief that God exists or that light travels at a finite velocity and beliefs about abstract matters (say, the belief that there is no largest prime number). It is much more plausible for observational beliefs like the belief that there are red flowers on my desk or that there are horses in the field. The proviso "under optimal conditions" is included since for the state of affairs *p* (for example, the presence of horses) to correlate with, or cause, subject S's belief that *p*, favorable perceptual conditions must obtain, such as that S's perceptual systems are functioning properly, the illumination is adequate, S's attention is not seriously distracted, and so on.

Although there seem to be some serious difficulties that (C) has to overcome, remember that (C) is only a rough-and-ready first pass, and none of the objections enumerated here need be taken as a disabling blow to the general approach.

1. The belief that there are horses in the field correlates reliably, let us suppose, with the presence of horses in the field. But it also correlates reliably with the presence of horse genes in the field (since the latter correlate reliably with the presence of horses). According to (C), someone observing horses in the field should have the belief that there are horse genes in the field. But this surely is wrong.

Moreover, the belief that there are horses in the field also correlates with the presence of undetached horse parts. But again, the observer does not have the belief that there undetached horse parts in the field. The general problem, then, is that an account like (C) cannot differentiate between belief with *p* as its content and belief with *q* as its content if *p* and *q* reliably correlate with each other. For any two correlated states of affairs *p* and *q*, (C) entails that one believes that *p* if and only if one believes that *q*, which evidently is incorrect. Restricting (C) to observational beliefs can relieve some of this problem, however.

2. Belief is *holistic* in the sense that what you believe is shaped, often crucially, by what else you believe. When you observe horselike shapes in the field, you are not likely to believe that there are horses in the field if you have read in the papers that many cardboard horses have been put up for a children's fair, or if you believe you are hallucinating, and so on. Correlational accounts make beliefs basically atomistic, at least for observational beliefs, but even our observational beliefs are constrained by other beliefs we hold, and the correlational approach as it stands is not sensitive to this aspect of belief content.
3. The belief that there are horses in the field is caused not only by horses in the field but also by cows and moose at dusk, cardboard horses at a distance, robot horses, and so on. In fact, this belief correlates more reliably with the disjunction "horses or cows and moose at dusk or cardboard horses or . . ." If so, why should we not say that when you are looking at the horses in the field, your belief has the *disjunctive* content "there are horses *or* cows *or* moose at dusk *or* cardboard horses *or* robot horses in the field"? This so-called disjunction problem has turned out to be a recalcitrant difficulty for the causal-correlational approach; it has been actively discussed, but there seems no solution that commands a consensus.[9]
4. We seem to have direct and immediate knowledge of what we believe, desire, and so on. I know, directly and without having to depend on evidence, that I believe it will rain tomorrow. That is, I seem to have direct knowledge of the content of my beliefs. There may be exceptions, but that does not overturn the general point. According to the correlational approach, my belief that there are horses in the field has the content it has because it correlates, or covaries, with the presence of horses in my vicinity. But this correlation is not some-

> thing that I know directly, without evidence or observation. So the correlational approach appears inconsistent with the special privileged status of our knowledge of the contents of our own mental states. (We discuss this issue further later, in connection with content externalism.)

These are some of the initial issues and difficulties for the correlational approach; whether, or to what extent, these difficulties can be overcome without compromising the naturalistic-reductive spirit of the theory remains an open question. Quite possibly, most of the difficulties are not really serious and can be resolved by further elaborations and supplementations. It may well be that this approach is the most promising one—in fact, the only viable one that promises to give a non-question-begging, naturalistic account of mental content.

Misrepresentation and the Teleological Approach

One important fact about representation is the possibility of *misrepresentation*. Misrepresentation does occur; you, or a mental-neural state of yours, may represent that there are horses in the field when there are none in sight. Or your perception may represent a red tomato in front of you when there is none (think about Macbeth and his bloody dagger). In such cases, misrepresentation occurs: The representational state misrepresents, and the representation is false. Representations have contents, and contents are "evaluable" in respect of truth, accuracy, fidelity, and related criteria of representational "success." It seems clear, then, that any account of representation must allow for the possibility of misrepresentation as well of course as correct, or successful, representation, just as any account of belief must allow for the possibility of false belief. One way of seeing how this could be a problem with the correlational approach is to go back to the disjunction problem discussed earlier. Suppose you form a representation with the content "there are horses over there" when there are no horses but only cows seen in the dusk. In such a case it would be natural to regard your purported representation as a misrepresentation—namely, as an instance of your representing something that does not exist, or representing something to be such and such when it is not such and such. But if we follow (C) literally, this seems impossible. If your representation was occasioned by cows seen in the dusk as well as horses, we would have to say that the representation has the content "horses or cows seen in the dusk" and that

that would make the representation correct and veridical. It would seem that (C) does not allow false beliefs or misrepresentations. But there surely are cases of misrepresentation; our cognitive systems are liable to produce false representations, even though they may be generally reliable.

This is where the teleological approach comes in to help out.[10] The basic concept employed in the teleological approach is that of a "function." For representation R to indicate (and thus represent) C, it is neither sufficient—nor necessary—that "whenever R occurs C occurs" holds. Rather, what must hold is that R has the *function* of indicating C—to put it more intuitively, R *is supposed* to indicate C and it is R's *job* to indicate C. Your representation has the content "there are horses over there" and not "there are horses or cows in the dusk over there" because it has the function of indicating the presence of horses, not horses or cows in the dusk. But things can go wrong, and systems do not always perform as they are supposed to. You form a representation of horses in the absence of horses; such a representation is *supposed* to be formed only when horses are present. That is exactly what makes it a case of misrepresentation. So it seems that the correlational-causal approach suitably supplemented with reference to function could solve the problem of misrepresentation.

But how does a state of a person or organism acquire a function of this kind? It is easy enough to understand function talk in connection with artifacts because we can invoke the purposes and intentions of their human designers and users. A thermometer reads 30°C, when the temperature is 20°C. What makes this a case of misrepresentation is that the thermometer's function is to indicate current air temperature, which is 20°C. That is the way the thermometer was designed to work and the way it is expected to work. It is the purposes and expectations external to the thermometer that give sense to the talk of functions. But this is something that we are not able to say, at least literally, about representations of natural systems, like humans and other higher animals. What gives a mental state (or a neural state) in us the function of representing some particular object or state of affairs? What gives a natural representation the job of representing "horses" rather than "horses or cows in the dusk"?

Philosophers who favor the teleological approach attempt to explain function in terms of evolution and natural selection. To say that representation R has the function of indicating C is to say that R has been selected, in the course of the evolution of the species to which the organism belongs, for the job of indicating C. This is like the fact that the heart has the function of pumping blood, or that the pineal gland has the function of secreting melatonin, because these organs have evolved for their performance of these tasks. Proper performance of these tasks presumably conferred adaptive advantages to our

ancestors. Similarly, we may presume that if R's function is to indicate C, performance of this job has given our ancestors biological advantages and, as some philosophers put it, R has been "recruited" by the evolutionary process to perform this function.

Exactly how the notion of function is to be explained is a further question that appears relatively independent of the core idea of the teleological approach. There are various and diverse biological-evolutionary accounts of function in the literature (see "For Further Reading" at the end of this chapter). Even if the theory of evolution were false and all biological organisms, including us, were created by God (so that we are God's "artifacts"), something like the teleological approach could still be right. It is God who gave our representations the indicating functions they have. But almost all contemporary philosophers of mind and of biology are naturalists, and it is important to them that function talk does not need to involve references to supernatural or transcendental plans, purposes, or designs. That is why they appeal to biology, learning and adaptation, and evolution for an account of function.

Narrow Content and Wide Content: Content Externalism

One thing that the correlational account of mental content highlights is this: Content has a lot to do with what is going on in the world, outside the physical boundaries of the creature. As far as what goes on inside is concerned, the frog in our world and the other-worldly frog are indistinguishable—they are in the same neural-sensory state, both registering a moving black dot. But in describing the representational content of their states, or what they "see," we advert to the conditions in the environments of the frogs: One frog sees a fly and the other sees a schmy. Or consider a simpler case: Peter is looking at a tomato, and Mary is also looking at one (a different tomato, but we suppose that it looks pretty much the same as Peter's tomato). Mary thinks to herself, "This tomato has gone bad," and Peter too thinks, "This tomato has gone bad." From the internal point of view, Mary's perceptual experience is indistinguishable from Peter's (we may suppose their neural states too are relevantly similar), and they would express their thoughts using the same words. But it is clear that the contents of their beliefs are different. For they involve different objects: Mary's belief is about the tomato she is looking at, and Peter's belief is about a different object altogether. Moreover, Mary's belief may be true and Peter's false, or vice versa. On one standard understanding of the notion of "content," beliefs with the same content must be true together or false together (that is,

contents serve as "truth conditions"). Obviously, the fact that Peter's and Mary's beliefs have different content is due to facts external to them; the difference in content cannot be explained in terms of what is going on inside the perceivers. It seems, then, that at least in this and other similar cases belief contents are differentiated, or "individuated," by reference to conditions external to the believer.

Beliefs whose content is individuated in this way are said to have "wide" or "broad" content. In contrast, beliefs whose content is individuated solely on the basis of what goes on inside the persons holding them are said to have "narrow" content. Alternatively, we may say that the content of an intentional state is narrow just in case it supervenes on the internal-intrinsic properties of the subject who is in that state, and that it is wide otherwise. This means that two individuals who are exactly alike in all intrinsic-internal respects must have the same narrow content beliefs but may well diverge in their wide content beliefs. Thus, our two frogs are exactly alike in internal-intrinsic respects but unlike in what their perceptual states represent. So the contents of these states do not supervene internally and are therefore wide.

Several well-known thought-experiments have been instrumental in persuading most philosophers that many, if not all, of our ordinary beliefs (and other intentional states) have wide content, that the beliefs and desires we hold are not simply a matter of what is going on inside our minds or heads. This is the doctrine of *content externalism*. Among these thought-experiments, the following two, the first due to Hilary Putnam and the second to Tyler Burge,[11] have been particularly influential.

Putnam's Thought-Experiment: Earth and Twin Earth

Imagine a planet, "Twin Earth," somewhere in the remote region of space, which is just like the Earth we inhabit, except in one respect: On Twin Earth, a certain chemical substance with the molecular structure XYZ, which has all the observable characteristics of water (it is transparent, dissolves salt and sugar, quenches thirst, puts out fire, freezes at 0°C, and so on), replaces water everywhere. So lakes and oceans on Twin Earth are filled with XYZ, not H_2O (that is, water), and Twin Earth people drink XYZ when they are thirsty, bathe and swim in XYZ, do their laundry in XYZ, and so on. Some Twin Earth people, including most of those who call themselves "Americans," speak English, which is indistinguishable from our English, and their use of the expression "water" is indistinguishable from its use on Earth.

But there is a difference: The Twin Earth "water" and our "water" refer to different things. When a Twin Earth inhabitant says, "Water is transparent," what she means is that XYZ is transparent. The same words when uttered by you, however, mean that water is transparent. The word "water" from a Twin Earth mouth means XYZ, not water, and the same word on your mouth means water, not XYZ. If you are the first visitor to Twin Earth and find out the truth about their "water," you may report back to your friends on Earth as follows: "At first I thought that the stuff that fills the oceans and lakes around here, and the stuff people drink and bathe in, was water, and it really looks and tastes just like water. But I just found out that it isn't water at all, although people around here call it 'water.' It's really XYZ, not water." You will not translate the Twin Earth word "water" into the English word "water"; you will translate it into "XYZ," or invent a new vernacular word, say "twater." We have to conclude then that the Twin Earth word "water" and our word "water" have different meanings, although what goes on inside the minds, or heads, of Twin Earth people may be exactly the same as what goes in ours, and their speech behavior involving their word "water" is indistinguishable from ours with our word "water." This semantic difference between our "water" and Twin Earth "water" is reflected in the way we describe and individuate mental states of people on Earth and people on Twin Earth. When a Twin Earth person says to the waiter, "Please bring me a glass of water!" she is expressing her desire for twater, and we will report, in *oratio obliqua*, that she wants some twater, not that she wants some water. When you say the same thing, you are expressing a desire for water, and we will say that you want water. You believe that water is wet, and your Twin Earth doppelganger believes that twater is wet. And so on. To summarize, people on Earth have water-thoughts and water-desires, whereas Twin Earth people have twater-thoughts and twater-desires; this difference is due to differences in the environmental factors external to the subjects, not to any differences in what goes on "inside" their heads.

Suppose we send an astronaut, Jones, to Twin Earth. She does not realize at first that the liquid she sees in the lakes and coming out of the tap is not water. She is offered a glass of this transparent liquid by her Twin Earth host and thinks to herself, "That's a nice, cool glass of water—just what I needed." Consider Jones's belief that the glass contains cold water. This belief is false, since the glass contains not water but XYZ, that is, twater. Although she is now on Twin Earth, in an environment full of twater and devoid of water, she is still subject to the standards current on Earth: Her words mean, and her thoughts are individuated, in accordance with the criteria that prevail on Earth. What

this shows is that a person's *past associations* with her environment play a role in determining her present meanings and thought contents. If Jones stays on Twin Earth long enough—say, a dozen years—we will likely interpret her word "water" to mean twater, not water, and attribute to her twater-thoughts rather than water-thoughts—that is, eventually she will come under the linguistic conventions of Twin Earth.

If these considerations are by and large correct, they show that two supervenience theses fail: First, the meanings of our expressions do not in general supervene on our *internal*, or *intrinsic*, physical-psychological states. I and my molecule-for-molecule-identical Twin Earth doppelganger are indistinguishable as far as our internal lives, both physical and mental, are concerned, and yet our words have different meanings—my "water" means water and his "water" means XYZ, that is, twater. Second, and this is what is of immediate interest to us, the contents of beliefs and other intentional states also fail to supervene on internal physical-psychological states. You have water-thoughts and your doppelganger has twater-thoughts, in spite of the fact that you two are in the same internal states, physical and psychological. Beliefs, or thoughts, are individuated by content—that is, that we regard beliefs with the same content as the same belief, and beliefs with different content count as different. So your water-thoughts and your twin's twater-thoughts are different thoughts. What beliefs you hold depends on your relationship, both past and present, to the things and events in your surroundings, as well as on what goes on inside you. The same goes for other content-carrying intentional states. If this is right, intentional states have wide content.

Burge's Thought-Experiment: Arthritis and "Tharthritis"

Consider a person, call him Peter, in two situations. (1) *The actual situation*: Peter thinks "arthritis" means inflammation of the bones. (It actually means inflammation of the bone joints.) Feeling pain and swelling in his thigh, Peter complains to his doctor, "I have arthritis in my thigh." His doctor tells him that people can have arthritis only in their joints. Two points should be noted: First, Peter believed, before he talked to his doctor, that he had arthritis in his thigh; and second, this belief was false.

(2) *A counterfactual situation*: Nothing has changed with Peter. Experiencing swelling and pain in his thigh, he complains to his doctor, "I have arthritis in my thigh." What is different about the counterfactual situation concerns the use of the word "arthritis" in Peter's speech community: In the situation we are imagining, the word is used to refer to inflammation of bones, not just bone

joints. That is, in the counterfactual situation Peter has a correct understanding of the word "arthritis," unlike in the actual situation. In the counterfactual situation, then, Peter is expressing a true belief when he says "I have arthritis in my thigh." But how should we report Peter's belief concerning the condition of his thigh in the counterfactual situation—that is, report in *our* language (in the actual world)? We cannot say that Peter believes that he has arthritis in his thigh, because in our language "arthritis" means inflammation of joints and he clearly does not have that, making his counterfactual belief false. We might coin a new expression (to be part of our language), "tharthritis," to mean inflammation of bones as well as of joints, and say that Peter, in the counterfactual situation, believes that he has tharthritis in his thigh. Again, note two points: First, in the counterfactual situation, Peter believes not that he has arthritis in his thigh but that he has tharthritis in his thigh; and second, this belief is true.

What this thought-experiment shows is that the content of belief depends, in part but crucially, on the speech practices of the linguistic community in which we situate the subject. Peter in the actual situation and Peter in the counterfactual situation are exactly alike when taken as an individual person (that is, when we consider his internal-intrinsic properties alone), including his speech habits (he speaks the same idiolect in both situations) and inner mental life. Yet he has different beliefs in the two situations: Peter in the actual world has the belief that he has arthritis in his thigh, which is false, but in the counterfactual situation he has the belief that he has tharthritis in his thigh, which is true. The only difference in the two situations is that of the linguistic practices of Peter's community (concerning the use of the word "arthritis"), not anything intrinsic to Peter himself. If this is right, beliefs and other intentional states do not supervene on the internal physical-psychological states of persons; if supervenience is wanted, we must include in the supervenience base the linguistic practices of the community to which people belong.

Burge argues, persuasively for most philosophers, that the example can be generalized to show that almost all contents are wide—that is, externally individuated. Take the word "brisket" (another of his examples): Some of us mistakenly think that brisket comes only from beef, and it is easy to see how a case analogous to the arthritis example can be set up. (The reader is invited to try.) As Burge points out, the same situation seems to arise for any word whose meaning is incompletely, or defectively, understood—in fact, any word whose meaning *could* be incompletely understood, which includes pretty much every word. When we profess our beliefs using such words, our beliefs are identified and individuated by the socially determined meanings of these

words (recall Peter and his "arthritis" in the actual situation), and a Burge-style counterfactual situation can be set up for each such word. Moreover, we seem to identify our own beliefs in terms of the words we would use to express them, even if we are aware that our understanding of these words is incomplete or defective. (How many of us know the correct meaning of, say, "mortgage," "justice of the peace," or "galaxy"?) This shows, it has been argued, that almost all of our ordinary belief attributions involve wide content.

If this is right, the question naturally arises: Are there beliefs whose content is not determined by external factors? That is, are there beliefs with "narrow content"? There appear to be beliefs, and other intentional states, that do not imply the existence of anything, or do not refer to anything, outside the subject who has them. For example, Peter's belief that he is in pain or that he exists, or that there are no unicorns, does not require anything other than Peter to exist, and it would seem that the content of these beliefs is independent of conditions external to Peter. If so, the narrowness of these beliefs is not threatened by considerations of the sort that emerged from the Twin Earth thought-experiment. But what of Burge's arthritis thought-experiment? Consider Peter's belief that he is in pain. Could we run on the word "pain" Burge's argument on "arthritis"? Surely it is possible for someone to misunderstand the word "pain" or any other sensation term. Suppose Peter thinks that "pain" applies to both pains and severe itches and that on experiencing a bad itch on his shoulder, he complains to his wife about an annoying "pain" in the shoulder. If the Burge-style considerations apply here, we have to say that Peter is expressing his belief that he is having a pain in his shoulder and that this is a false belief.

The question is whether that is indeed what we would, or should, say. It would seem not unreasonable that knowing what we know about Peter's misunderstanding of the word "pain" and the sensation he is actually experiencing, the correct thing to say is that he believes, and in fact knows, that he is experiencing an itch on his shoulder. It is only that in saying, "I am having a pain in my shoulder," he is misdescribing his sensation and hence misreporting his belief.

Now, consider the following counterfactual situation: In the linguistic community to which Peter belongs, "pain" is used to refer to pains and severe itches. How would we report, in our own words, the content of Peter's belief in the counterfactual situation when he utters "I have a pain in my shoulder"? Remember that both in the actual and counterfactual situations, Peter is having a bad itch, and no pain. There are these possibilities: (i) We say "He believes that he has a pain in his shoulder"; (ii) we say "He believes that he has a

bad itch in his shoulder"; and (iii) we do not have a word in English that can be used for expressing the content of his belief (but we could introduce a neologism, "painitch," and say "Peter believes that he is having a painitch in his shoulder"). Obviously, (i) has to be ruled out; if (iii) is what we should say, the arthritis argument applies to the present case as well, since this would show that a change in the social environment of the subject can change the belief content attributed to him. But it is not obvious that this, rather than (ii), is the correct option. It seems to be an open question, then, whether the arthritis argument applies to cases involving beliefs about one's own sensations, and there seems to be a reason for the inclination to say of Peter in the actual world that he believes he is having severe itches rather than that he believes he is having pains. The reason is that if we were to opt for the latter, it would make his belief false, and this is a belief about his own current sensations. But we assume that under normal circumstances people do not make mistakes in identifying their current sensory experiences. This assumption need not be taken as a contentious philosophical doctrine; arguably, recognition of first-person authority on such matters also reflects our common social-linguistic practices, and this may very well override the kinds of considerations advanced by Burge in the case of arthritis and the rest.

These considerations should give us second thoughts concerning Burge's thought-experiment involving arthritis and tharthritis. As you will recall, this involved a person, Peter, who misunderstands the meaning of "arthritis" and, on experiencing pain in his thigh, says to his doctor, "I have arthritis in my thigh." With Burge, we said that Peter believes that he has arthritis in his thigh, and that this belief is false. Is this what we should really say? Isn't there an option, perhaps a more reasonable one, of saying that Peter, in spite of the words he used, doesn't believe that he has arthritis in his thigh; rather, the content of the belief he expresses when he says to the doctor "I have arthritis in my thigh" is to the effect that he has pain in his thigh, or that he has an inflammation of his thigh bone. He does have a false, or defective, belief—about the meaning of the word "arthritis"—and this leads him to misreport the content of his belief. Of course, it is no surprise that the meanings of words depend on the linguistic practice of the speech community. The reader is invited to ponder this way of responding to Burge's thought-experiment.

Another point to consider is beliefs of animals without speech. Do cats and dogs have beliefs and other intentional states whose contents can be reported in the form: "Fido believes that *p*," where *p* stands in for a declarative sentence? We do say things like "Fido believes that Charlie is calling him to come upstairs," "He believes that the mail carrier is at the door," and so on. But it is clear

that the arthritis-style arguments cannot be applied to such beliefs since Fido does not belong to any speech community and the only language that is involved is our own, namely, the language of the person who makes such belief attributions. In what sense, then, could animal beliefs be externally individuated? It seems that Putnam's Twin Earth–style considerations can be applied to animal beliefs (also recall our fly-schmy example), but Burge-style argument cannot. However, the case of animal beliefs can cut both ways as far as Burge's argument is concerned, for we might argue, as some philosophers have,[12] that nonlinguistic animals are not capable of having intentional states (in particular, beliefs) and, therefore, the inapplicability of Burge's considerations is only to be expected. Some will find this line of thinking highly implausible, namely that only animals that use language for social communication are capable of having beliefs and other intentional states.

The Metaphysics of Wide Content States

Considerations involved in the two thought-experiments show that many, if not all, of our ordinary beliefs and other intentional states have wide content. Their contents are "external": They are determined, in part but importantly, by factors outside the subject—factors in her physical and social environment and in her history of interaction with it. Before these externalist considerations were brought to our attention, philosophers used to think that beliefs, desires, and the like were "in the mind," or at least "in the head." Putnam, the inventor of the Twin Earth parable, declared, "Cut the pie any way you like, 'meanings' just ain't in the head."[13] Should we believe that beliefs and desires are not in the head, or in the mind, either? If so, where are they? *Outside* the head? If so, just where? Does that even make sense? Let us consider some possibilities.

1. We might say that the belief that water and oil do not mix is constituted in part by water and oil—that the belief itself, in some sense, involves the actual stuff, water and oil, in addition to the person (or her "head") having the belief. A similar response in the case of arthritis would be that Peter's belief that he has arthritis is in part constituted by his linguistic community. The general idea is that all the factors that play a role in determining the content of a belief *ontologically constitute* that belief; the belief is a state that comprises these items within itself. Thus, we have a simple explanation for just how your belief that water is wet differs from your Twin Earth doppelganger's belief that twater is wet: Yours includes water as a con-

stituent, and hers includes twater as a constituent. On this approach, then, beliefs extrude from the subject's head into the world, and there are no bounds to how far they can reach. The whole universe would, on this approach, be a constituent of your beliefs about the universe! Moreover, all beliefs about the universe would appear to have exactly the same constituent, namely, the universe. This sounds absurd, and it is absurd. We can also see that this general approach would make the causal role of beliefs difficult to understand—beliefs as either causes or effects.

2. We might consider the belief that water and oil do not mix as a *relation* holding between the subject, on the one hand, and water and oil, on the other. Or alternatively, we take the belief as a *relational property* of the subject involving water and oil. (That Socrates is married to Xanthippe is a relational fact; Socrates also has the relational property of being married to Xanthippe, and conversely, Xanthippe has the relational property of being married to Socrates.) This approach makes causation of beliefs more tractable: We can ask, and will sometimes be able to answer, how a subject came to bear this belief relation to water and oil, just as we can ask how Xanthippe came to have the relational property of being married to Socrates. But what of other determinants of content? As we saw, belief content is determined in part by the history of one's interaction with one's environment. And what of the social-linguistic determinants, as in Burge's examples? It seems at least awkward to consider beliefs as relations with respect to these factors.
3. The third possibility is to consider beliefs to be wholly internal to the subjects who have them but consider their contents, when they are wide, as giving *relational specifications*, or *descriptions*, of the contents. On this view, beliefs may be neural states or other types of physical states of organisms to which they are attributed, and as such they are "in" the believer's head, or mind. Contents, then, are construed as ways of specifying, or describing, the representational properties of these states; wide contents are thus specifications in terms that involve factors and conditions external to the subject, both physical and social, both current and historical. We can refer to, or pick out, Socrates by relational descriptions, that is, in terms of his relational properties—for example, "the husband of Xanthippe," "the Greek philosopher who drank hemlock in a prison in Athens," "Plato's mentor," and so on. But this does not mean that

> Xanthippe, hemlock, or Plato is a constituent part of Socrates, nor does it mean that Socrates is some kind of a "relational entity." Similarly, when we specify Jones's belief as the belief that water and oil do not mix, we are specifying this belief relationally, by reference to water and oil, but this does not mean that water and oil are constituents of the belief or that the belief itself is a relation to water and oil.

Let us look at this last approach in a bit more detail. Consider physical magnitudes such as mass and length, which are standardly considered to be paradigm examples of intrinsic properties of material objects. How do we *specify*, *represent*, or *measure* the mass or length of an object? The answer: relationally. To say that this metal rod has a mass of three kilograms is to say that it bears a certain relationship to the International Prototype Kilogram. (It would balance, on an equal-arm balance, three objects that each balance the Standard Kilogram.) Likewise, to say that the rod has a length of two meters is to say that it is twice the length of the Standard Meter (or twice the distance traveled by light in a vacuum in a certain specified fraction of a second). These properties, mass and length, are intrinsic, but their specifications or representations are extrinsic and relational, involving relationships to other things and properties in the world. Moreover, the availability of such extrinsic representations may be essential to the utility of these properties in the formulation of scientific laws and explanations. They make it possible to relate a given intrinsic property to other significant properties in theoretically interesting and fruitful ways. Similar considerations might explain the usefulness of wide contents, or relational descriptions of beliefs, in vernacular explanations of human behavior.

In physical measurements, we use numbers to specify properties of objects, and these numbers involve relationships to other objects (see the above discussion of what "three kilograms" refers to). In attributing to persons beliefs, we use propositions, or content sentences, to specify their contents, and these propositions often involve references to objects and events outside the believers. When we say that Jones believes that water is wet, we are using the content sentence "water is wet" to specify this belief, and the appropriateness of this sentence as a specification of the belief depends on Jones's relationship, past and present, to her environment. What Burge's examples show is that the choice of a content sentence may depend also on the social-linguistic facts about the person holding the belief. In a sense, we are "measuring" people's mental states using sentences, just as we measure physical magnitudes using numbers.[14] Just as the assignment of numbers in measurement involves relationships to things

other than the things whose magnitudes are being measured, the use of content sentences in the specification of belief contents makes use of, and depends on, factors outside the subject. In both cases the informativeness and utility of the specifications—the assigned numbers or sentences—depend crucially on the involvement of external factors and conditions.[15]

This approach seems to have much to recommend itself over the other two. It locates beliefs and other intentional states squarely within the subjects; ontologically, they are states of the persons holding them, not something that somehow extrudes from them into the outside, like some green goo we see in science fiction films! This is a more elegant metaphysical picture than its alternatives. What is "wide" about these states is their specifications or descriptions, not the states themselves. And there are good reasons for using wide content specifications. For one, we want them to indicate the representational contents of beliefs (and other intentional states)—what states of affairs they represent—and it is no surprise that this involves reference to external conditions. After all, the whole point of beliefs is to represent states of affairs in the world, outside the believer. For another, the sorts of social-linguistic constraints involved in Burge's examples may be crucial to the uniformity, stability, and intersubjectivity of content attributions. The upshot is that it is important not to conflate the ontological status of intentional states with the modes of their specification.

Is Narrow Content Possible?

You believe that water extinguishes fires, and your twin on Twin Earth believes that twater extinguishes fires. The two beliefs have different contents: What you believe is not the same as what your twin believes. But leaving the matter here is unsatisfying; it misses something important—something *psychologically* important—that you and your twin share in holding these beliefs. "Narrow content" is supposed to capture this something you and your twin share.

First, we seem to have a strong sense that both you and your twin conceptualize the same state of affairs in holding the beliefs about water and twater, respectively; the way things seem to you when you think that freshwater fills the Great Lakes must be the same, we feel, as the way things seem to your twin when she thinks that fresh twater fills the Twin Earth Great Lakes. From an internal psychological perspective, your thought and her thought seem to have the same significance. In thinking of water, you perhaps have the idea of a substance that is transparent, flows a certain way, tastes a certain way, and so on; in thinking of twater, your twin has the same associations. Or take the frog case: Isn't it plausible to suppose that the frog in our world that detects a fly and the

other-worldly frog that detects a schmy are in the same perceptual state—a state whose "immediate" content consists in a black dot flitting across the visual field? There is a strong intuitive pull toward the view that there is something important that is common to your psychological life and your twin's, and to our frog's perceptual state and the other-worldly frog's, that could reasonably be called "content."

Second, consider your behavior and your twin's behavior: They show a lot in common. For example, when you find your couch on fire, you pour water on it; when your twin finds her couch on fire, she pours twater on it. If you were visiting Twin Earth and found a couch on fire there, you would pour twater on it too (and conversely, if your twin is visiting Earth). In ordinary situations your behavior involving water is the same as her behavior involving twater; moreover, your behavior would remain the same if twater were substituted for water everywhere, and this goes for your twin as well mutatis mutandis. It seems then that the water-twater difference is *psychologically irrelevant*—irrelevant for behavior causation and explanation. The difference between water-thoughts and twater-thoughts cancels itself out, so to speak. What is important for psychological explanation seems to be what you and your twin share, namely, thoughts with narrow content. So the question arises: Does psychological theory need wide content? Can it get by with narrow content alone?

We have seen some examples of beliefs that plausibly do not depend on the existence of anything outside the subject holding them: your beliefs that you exist, that you are in pain, that unicorns do not exist, and the like. Although we have left open the question of whether the arthritis argument applies to them, they are at least "internal" or "intrinsic" to the subject in the sense that for these beliefs to exist, nothing outside the subject needs to exist. It appears, then, that these beliefs do not involve anything external to the believer and therefore that these beliefs supervene solely on the factors internal to him (again barring the possibility that the Burge-style considerations generalize to all expressions without exception).

However, a closer look reveals that some of these beliefs are not supervenient only on internal states of the believer. For we need to consider the involvement of the subject herself in the belief. Consider Mary's belief that she is in pain. The content of this belief is that she—that is, Mary—is in pain. This is the state of affairs represented by the belief, and this belief is true just in case that state of affairs obtains—that is, just in case Mary is in pain. Now we put Mary's twin on Twin Earth in the same internal physical state that Mary is

in when she has this belief. If mind-body supervenience, as *intuitively understood*, holds, it would seem that Mary's twin too will have the belief that she is in pain. However, her belief has the content that *she* (Twin Earth Mary) is in pain, not that Mary is in pain. The belief is true if and only if Mary's twin is in pain. Beliefs with the same content are true together, or false together. It follows, then, that belief contents in cases of this kind do not supervene on the internal-intrinsic physical properties of persons. This means that the following two ideas that are normally taken to lie at the core of the notion of "narrow content" fail to coincide: (1) Narrow content is internal and intrinsic to the believer and does not involve anything outside her current state; and (2) narrow content, unlike wide content, supervenes on the current internal physical state of the believer.[16]

One possible way to look at the situation is this: What examples of this kind show is not that these beliefs do not supervene on the internal physical states of the believer, but rather that we should revise the notion of "same belief"—that is, we need to revise the criteria of belief individuation. In our discussion thus far, individual beliefs (or "belief tokens") have been considered to be "the same belief" (or the same "belief type") just in case they have the same content; on this view, two beliefs have the same content only if their truth condition is the same (that is, necessarily they are true together or false together). As we saw, Mary's belief that she, Mary, is in pain and her twin's belief that she, the twin Mary, is in pain do not have the same truth condition and hence must count as belonging to different belief types. That is why supervenience fails for these beliefs. However, there is an obvious and natural sense in which Mary and her twin have "the same belief"—even beliefs with "the same content"—when each believes that she is in pain. More work, however, needs to be done to capture this notion of content or sameness of belief,[17] and that is part of the project of explicating the notion of narrow content.

As noted, it is widely accepted that most of our ordinary belief attributions, as well as attributions of other intentional states, involve wide content. Some hold not only that all contents are wide but that the very notion of narrow content makes no sense. One point often made against narrow content is its alleged ineffability: How do we capture the shared content of Jones's belief that water is wet and her twin's belief that twater is wet? And if there is something shared, why is it a kind of "content"?

One way the friends of narrow content have tried to deal with such questions is to treat narrow content as an abstract technical notion, roughly in the following sense. The thing that Mary and her twin share plays the following

role: If anyone has it and has acquired her language on Earth (or in an environment containing water), her word "water" refers to water and she has water-thoughts; if anyone has it and has acquired her language on Twin Earth (or in an environment containing twater), her word "water" refers to twater and she has twater-thoughts; for anyone who has it and has acquired her language in an environment in which a substance with molecular structure PQR replaces water everywhere, her word "water" refers to PQR; and so on. The same idea applies to the frog case: What the two frogs, one in this world and the other in a world with schmies but no flies, have in common is this: If a frog has it and inhabits an environment with flies, it has the capacity to have flies as part of its perceptual content, and similarly for frogs in a schmy-inclusive environment. Technically, narrow content is a function from environmental contexts (including contexts of language acquisition) to wide contents (or truth conditions).[18] One question that has to be answered is why narrow content in that sense is a kind of content. For isn't it true, by definition, that content is "semantically evaluable"—that is, that it is something that can be true or false, accurate to various degrees, and so on? Narrow content, conceived as a function from environment to wide content, does not seem to meet this conception of content; it does not seem like the sort of thing that can be said to be true or false. Here various strategies for meeting this point seem possible; however, whether any of them will work is an open question.

Two Problems for Content Externalism

We briefly survey here two outstanding issues confronting the thesis that most, perhaps all, of our intentional mental states have wide content. (The first was briefly alluded to earlier.)

The Causal-Explanatory Efficacy of Wide Content

Even if we acknowledge that commonsense psychology individuates intentional states widely and formulates causal explanations of behavior in terms of wide content states, we might well ask whether this is an ineliminable feature of such explanations. Several considerations can be advanced to cast doubt on the causal-explanatory efficacy of wide content states. First, we have already noted the similarity between the behaviors of people on Earth and those of their Twin Earth counterparts in relation to water and twater, respectively. We saw that in formulating causal explanations of behaviors, the difference between water-thoughts and twater-thoughts somehow cancels itself out

by failing to manifest itself in a difference in the generation of behavior. Second, to put the point another way, if you are a psychologist who has already developed a working psychological theory of people on Earth, formulated in terms of content-bearing intentional states, you obviously would not start all over again from scratch when you want to develop a psychological theory for Twin Earth people. In fact, you are likely to say that people on Earth and those on Twin Earth have "the same psychology"—that is, the same psychological theory holds for both groups. In view of this, isn't it more appropriate to take the difference between water-thoughts and twater-thoughts, or water-desires and twater-desires, merely as a difference in the values of a contextual parameter to be fixed to suit the situations to which the theory is applied rather than as an integral element of the theory itself? If this is correct, doesn't wide content drop out as part of the theoretical apparatus of psychological theory?

Moreover, there is a metaphysical point to consider: The proximate cause of my physical behavior (say, my bodily motions), we feel, must be "local"—it must be a series of neural events originating in my central nervous system that causes the contraction of appropriate muscles, which in turn moves my limbs. This means that what these neural events represent in the outside world is irrelevant to behavior causation: If the same neural events occur in a different environment so that they have different representational (wide) content, they would still cause the same physical behavior. That is, we have reason to think that proximate causes of behavior are *locally* supervenient on the internal physical states of an organism, but that wide content states are not so supervenient. Hence, the wideness of wide content states is not relevant to causal explanations of physical behavior. (You may recall discussion of the irrelevance of representational contents of computational states to the course of computational process, in chapter 5.)

One way in which the friends of wide content have tried to counter these considerations goes as follows. What we typically attempt to explain in commonsense psychology is not physical behavior but action—not why your right hand moved thus and so, but why you turned on the stove, why you boiled the water, why you made the tea. To explain why your hand moved in a certain way, it may suffice to advert to causes "in the head," but to explain why you turned on the stove or why you boiled the water, we must invoke wide content states: because you wanted to heat the kettle of water, because you wanted to make a cup of tea for your friend, and so on. Behaviors explained in typical commonsense explanations are given under "wide descriptions," and we need wide content states to explain them. So the point of the reply is that we need wide content to explain "wide behavior." Whether this response is

sufficient is something to think about. In particular, we might raise questions as to whether the wideness of thoughts and the wideness of behavior are playing any real role in the causal-explanatory relation involved, or whether they merely ride piggyback, so to speak, on an underlying causal-explanatory relationship between the neural states, or narrow content states, and physical behavior. (The issues discussed in an earlier section, "The Metaphysics of Wide Content States," are directly relevant to these causal-explanatory questions about wide content. The reader is encouraged to think about whether the third option described in that section could help the content externalist to formulate a better response.)

Wide Content and Self-Knowledge

How do we know that Mary believes that water is wet and that Mary's twin on Twin Earth believes that twater is wet? Because we know that Mary's environment contains water and that Mary's twin's environment contains twater. Now consider the matter from Mary's point of view: How does she know that she believes that water is wet? How does she know the content of her own thoughts?

We believe that a subject has special, direct access to her own mental states (see chapters 1 and 9). Perhaps the access is not infallible and does not extend to all mental states, but it is uncontroversial that there is special first-person authority in regard to one's own occurrent thoughts. When you reflect on what you are thinking, you apparently know directly, without further evidence or reasoning, what you think; the content of your thought is immediately and directly accessible to you, and the question of having evidence or doing research does not arise. If you think that the shuttle bus is late and you might miss your flight, you know, in the very act of thinking, that that is what you are thinking. First-person knowledge of the contents of one's own current thoughts is direct and immediate and carries a special sort of authority.

Return now to Mary and her knowledge of the content of her belief that water is wet. It seems plausible to think that in order for her to know that her thought is about water, not about twater, she is in the same epistemic situation that we are in with respect to the content of her thought. We know that her thought is about water, not twater, because we know, from observation, that her environment is water-inclusive, not twater-inclusive. But why doesn't she too have to know that if she is to know that her thought is about water, not twater, and how can she know something like that without observation or evidence? It looks like she may very well lose her specially privileged epistemic ac-

cess to the content of her own thought, because her knowledge of her thought content is now put on the same footing as third-person knowledge of it.

To make this more vivid, suppose that Twin Earth exists in a nearby planetary system and we can travel between Earth and Twin Earth. It is plausible to suppose that if one spends a sufficient amount of time on Earth (or Twin Earth), one's word "water" becomes locally acclimatized and begins to refer to the local stuff, water or twater, as the case may be. Now, Mary, an inveterate space traveler, forgets on which planet she has been living for the past several years, whether it is Earth or Twin Earth; surely that is something she cannot know directly without evidence or observation. Now ask: Can she know, directly and without further investigation, whether her thoughts (say, the thought she expresses when she mutters to herself, "The tap water in this fancy hotel doesn't taste so good") are about water or twater? It prima facie makes sense to think that just as she cannot know, without additional evidence, whether her present use of the word "water" refers to water or twater, she cannot know, without investigating her environment, whether her thought, on seeing the steaming kettle, has the content that the water is boiling or that the twater is boiling. If something like this is right, then content externalism would seem to have the consequence that most of our knowledge of our own intentional states is not direct and, like most other kinds of knowledge, must be based on evidence. That is to say, content externalism appears to be prima facie incompatible with privileged first-person access to one's own mind. Content externalists are, of course, not without answers, but an examination of these is beyond the scope of this chapter.

* * *

These issues concerning wide and narrow content—especially the second concerning content externalism and self-knowledge—have been vigorously debated and are likely to be with us for some time. Their importance can hardly be exaggerated: Content-carrying states—that is, intentional states like belief, desire, and the rest—constitute the central core of our commonsense ("folk") psychological practices, providing us with a framework for formulating explanations and predictions of what we and our fellow humans do. Without this essential tool for understanding and anticipating human action and behavior, a communal life would be unthinkable. Moreover, the issues go beyond commonsense psychology. There is, for example, this important question about scientific psychology and cognitive science: Should the sciences of human behavior and cognition make use of content-carrying intentional

states like belief and desire, or their more refined and precise scientific analogues, in formulating its laws and explanations? Or should they, or could they, transcend the intentional idiom by couching their theories and explanations in purely nonintentional (perhaps, ultimately neurobiological) terms? These questions concern the centrality of content-bearing, representational states to the explanation of human action and behavior—both in everyday psychological practices and in theory construction in scientific psychology.

For Further Reading

On interpretation theory, see the works by Davidson, Quine, and Lewis cited in footnote 1; see also Daniel C. Dennett, "Intentional Systems" and "True Believers."

On causal-correlational theories of content, see the works cited in footnote 7; see also Robert Cummins, *Meaning and Mental Representation*, especially chapters 4 through 6. Another useful book on issues of mental content, including some not discussed in this chapter, is Lynne Rudder Baker, *Explaining Attitudes*. There are several helpful essays in *Meaning in Mind*, edited by Barry Loewer and Georges Rey.

On teleological accounts of mental content, see Fred Dretske, "Misrepresentation," and Ruth Millikan, "Biosemantics." Karen Neander's "Teleological Theories of Mental Content" is a comprehensive survey and analysis.

On narrow and wide content, the two classic texts that introduced the issues are Hilary Putnam, "The Meaning of 'Meaning,'" and Tyler Burge, "Individualism and the Mental." See also Fodor's *Psychosemantics* and "A Modal Argument for Narrow Content." On narrow content, see Gabriel Segal, *A Slim Book About Narrow Content*. For a discussion of these issues in relation to scientific psychology, see Frances Egan, "Must Psychology Be Individualistic?" Joseph Mendola's *Anti-Externalism* is an extended and helpful analysis and critique of externalism; see chapter 2 for discussion of Putnam's and Burge's thought-experiments in support of externalism.

Concerning content and causation, the reader may wish to consult the following: Colin Allen, "It Isn't What You Think: A New Idea About Intentional Causation"; Lynne Rudder Baker, *Explaining Attitudes*; Tim Crane, "The Causal Efficacy of Content: A Functionalist Theory"; Fred Dretske, *Explaining Behavior* and "Minds, Machines, and Money: What Really Explains Behavior"; Jerry Fodor, *Psychosemantics* and "Making Mind Matter More"; and Pierre Jacob, *What Minds Can Do*.

On wide content and self-knowledge, see Donald Davidson, "Knowing One's Own Mind"; Tyler Burge, "Individualism and Self-Knowledge"; Paul Boghossian, "Content and Self-Knowledge"; and John Heil, *The Nature of True Minds*, chapter 5. Three recent collections of essays on the issue are *Externalism and Self-Knowledge*, edited by Peter Ludlow and Norah Martin; *Knowing Our Own Minds*, edited by Crispin Wright, Barry C. Smith, and Cynthia Macdonald; and *New Essays on Semantic Externalism and Self-Knowledge*, edited by Susan Nuccetelli.

Notes

1. The discussion in this section is based on the works of W. V. Quine and Donald Davidson—especially Davidson's. See Quine on "radical translation" in his *Word and Object*, chapter 2. Davidson's principal essays on interpretation are included in his *Inquiries into Truth and Interpretation*; see, in particular, "Radical Interpretation," "Thought and Talk," and "Belief and the Basis of Meaning." Also see David Lewis, "Radical Translation."

2. Here we are making the plausible assumption that we can determine, on the basis of observation of Karl's behavior, that he affirmatively utters, or holds true, a sentence S, without our knowing what S means or what belief Karl expresses by uttering S. (The account would be circular otherwise.) It can be granted that holding true a sentence is a psychological attitude or event. For further discussion of this point, see Davidson, "Thought and Talk," pp. 161–162.

3. The parenthetical part is often assumed without being explicitly stated. Some writers state it as a separate principle, sometimes called the "requirement of rationality." There are many inequivalent versions of the charity principle in the literature. Some restrictions on the class of beliefs to which charity is to be bestowed are almost certainly necessary. For our examples, all we need is to say that speakers' beliefs about observable features of their immediate environment are generally true; that is, we restrict the application of charity to "occasion sentences" whose utterances are sensitive to the observable change in the environment.

4. Such a position seems implicit in, for example, Daniel Dennett's "True Believers."

5. The following statement from Davidson, who has often avowed himself to be a mental realist, seems to have seemingly irrealist, or possibly relativist, implications: "For until the triangle is completed connecting two creatures [the interpreter and the subject being interpreted], and each creature with

common features of the world, there can be no answer to the question whether a creature, in discriminating between stimuli, is discriminating stimuli at sensory surfaces or somewhere further out, or further in. Without this sharing of reactions to common stimuli, thought and speech would have no particular content—that is, no content at all. It takes two points of view to give a location to the cause of a thought, and thus, to define its content." See Davidson, "Three Varieties of Knowledge," pp. 212–213.

6. To use Robert Stalnaker's term in his *Inquiry*, p. 18. Fred Dretske, too, uses "indicator" and its cognates for similar purposes in his writings on representation and content.

7. This version captures the gist of the correlational approach, which has many diverse versions. Important sources include Fred Dretske, *Knowledge and the Flow of Information* and "Misrepresentation"; Robert Stalnaker, *Inquiry*; and Jerry A. Fodor, *Psychosemantics* and *A Theory of Content and Other Essays*. Dennis Stampe is usually credited with initiating this approach in "Toward a Causal Theory of Linguistic Representation." For discussion and criticisms, see Brian McLaughlin, "What Is Wrong with Correlational Psychosemantics?" (to which I am indebted in this section); and Louise Antony and Joseph Levine, "The Nomic and the Robust"; Lynne Rudder Baker, "Has Content Been Naturalized?"; and Paul Boghossian, "Naturalizing Content" in *Meaning in Mind*, ed. Barry Loewer and Georges Rey.

8. This means that S is entertaining this belief, actively in some sense, at the time.

9. For discussion of this issue, see the works cited in note 7.

10. This is not to say that the teleological approach is necessarily the only solution to the problem of misrepresentation or the disjunction problem. See Jerry A. Fodor, *A Theory of Content and Other Essays*.

11. Hilary Putnam, "The Meaning of 'Meaning'"; Tyler Burge, "Individualism and the Mental." The terms "narrow" and "wide" are due to Putnam.

12. Most notably Descartes and Davidson. See Davidson's "Rational Animals."

13. Hilary Putnam, "The Meaning of 'Meaning,'" p. 227.

14. This idea was first introduced by Paul M. Churchland in "Eliminative Materialism and the Propositional Attitudes." It has been systematically elaborated by Robert Matthews in "The Measure of Mind." However, these authors do not relate this approach to the issues of content externalism. For another perspective on the issues, see Ernest Sosa, "Between Internalism and Externalism."

15. Burge makes this point concerning content sentences in "Individualism and the Mental."

16. Beliefs with wide content will generally not supervene on the internal, intrinsic physical properties of the subjects. That is not surprising; the present case is worth noting because it apparently involves narrow content.

17. In this connection, see Roderick Chisholm's theory in *The First Person*, which does not take beliefs as relations to propositions but construes them as attributions of properties. David Lewis has independently proposed a similar approach in "Attitudes *De Dicto* and *De Se*." On an approach of this kind, both Mary and twin Mary are self-attributing the property of being in pain, and the commonality shared by the two beliefs consists in the self-attribution of the same property, namely that of being in pain.

18. See Stephen White, "Partial Character and the Language of Thought," and Jerry A. Fodor, *Psychosemantics*. See also Gabriel Segal, *A Slim Book About Narrow Content*.