Lines of Thought
Short philosophical books
General editors: Peter Ludlow and Scott Sturgeon
Published in association with the Aristotelian Society

Hume Variations
Jerry A. Fodor

Perfectionism and the Common Good: Themes in the Philosophy of
T. H. Green
David O. Brink

Moral Fictionalism
Mark Eli Kalderon

Knowledge and Practical Interests
Jason Stanley

Thought and Reality
Michael Dummett

Our Knowledge of the Internal World
Robert C. Stalnaker

OUR KNOWLEDGE
OF THE INTERNAL
WORLD

ROBERT C. STALNAKER

CLARENDON PRESS · OXFORD
Contents

Acknowledgements vi
1. Starting in the Middle 1
2. Epistemic Possibilities and the Knowledge Argument 24
3. Locating Ourselves in the World 47
   3A. Notes on Models of Self-Locating Belief 69
4. Phenomenal and Epistemic Indistinguishability 75
5. Acquaintance and Essence 94
6. Knowing What One is Thinking 112
7. After the Fall 132

References 139
Index 145
phenomenal knowledge and self-locating knowledge are, in some sense, essentially subjective, and both raise questions about the relation between an objective conception of the world and a subject's perspective on it. I will come back to the story of Mary in Chapter 4, but first I want to try to make clear exactly how we should understand self-locating attitudes, a problem that is of interest in itself, and that will be relevant to some of the issues about knowledge of content that I want to talk about later, as well as to Jackson's vexing puzzle. So in the next chapter I will sketch the way I think self-locating knowledge and belief should be represented.

3

Locating Ourselves in the World

Consider the case of the two gods. They inhabit a certain possible world, and they know exactly what world it is. Therefore they know every proposition that is true at their world. Insofar as knowledge is a propositional attitude, they are omniscient. Still I can imagine them to suffer ignorance: neither one knows which of the two he is.

David Lewis

John Perry argued that we need to recognize a kind of representational content—reflexive content—that essentially involves the subject who is doing the representing, and her relation to what is being represented. And he argued that distinguishing this kind of content from subject matter content will dissolve the puzzle about Mary. I agree with Perry that it will help, at least to clarify the puzzle, to be explicit about the role of the subject’s perspective in the representation of content, but it remains unclear exactly what reflexive content is. My aim in this chapter is to sketch an account of self-locating belief that I hope will begin to make sense of a notion of informational content that is not detachable from

1 Lewis (1979), 139.
the situation of a subject, or from a context in which the content is ascribed. Perry’s discussion made use of a model of a representational mechanism—a form that internal representations might take. The account I will sketch focuses on the informational content of the representation, trying to be as neutral as possible about the means by which content is stored or expressed. With both ordinary beliefs about the objective world, and self-locating beliefs, we will be concerned with truth conditions: with what conditions a world must meet in order for a believer to have a correct conception of the world. And in keeping with the general anti-Cartesian strategy, the content of a subject’s beliefs will be characterized from an external perspective. Theorists, or attributors, use their own resources to describes the world as the subject takes it to be. That is, they describe the world according to the subject in terms of the things, events, properties and relations that they find in the actual world.

1. CENTERED WORLDS PROPOSITIONS

We start, on this approach, with the classical possible-worlds representation of a state of belief or knowledge as a set of possible worlds, the doxastically or epistemically accessible worlds, those that are compatible with the subject’s beliefs or knowledge. This kind of representation is highly idealized, but the idealization does not avoid or evade the particular problems we want to focus on, and will help to sharpen the issues. The problem is how to generalize or modify the classical model to take account of essentially indexical or self-locating belief and knowledge.

If our question is, exactly when are self-locating statements and beliefs true, then the answer is clear and unproblematic: A statement of the form “I am F” is true, when said or thought by x, if and only if x is F. So if David Kaplan says “my pants are on fire”, what he says is true if and only if David Kaplan’s pants are on fire at the time at which he says it. But this answer does not tell us what the contents of the statements are—what information they convey, what belief they express, what kind of fact, if any, they state. The reason we cannot straightforwardly infer from this account of the truth conditions of the statement or belief to the content is that the facts that determine the truth value may play two different roles: to determine what is said, and to determine whether what is said is true or false. These truth conditions give us the content of the statement only if it is assumed that the statement has the same content on each occasion of use.

One influential answer to the question of content, defended by David Lewis, is this: If the contents of ordinary beliefs about objective facts can be represented by sets of possible worlds, then the contents of self-locating beliefs can be represented by sets of centered possible worlds, where a centered possible world is a pair consisting of a world plus a center, which is a designated time and person. Since “I am sad” is true if and only if it is said or thought by a sad person, its content will be represented by the set of centered possible worlds that have a sad person at their designated center. The content of “the meeting is about to start” will be the set of centered worlds at which the meeting in question takes place soon after the time that is designated as the center.2

This elegant modification of the standard account is a generalization, since as Lewis observed, ordinary beliefs about the objective world can be represented as a special case of self-locating beliefs, beliefs where the centers are irrelevant. Beliefs that might be expressed with eternal sentences, such as my belief that pigs can’t fly, have as their content the set of centered worlds,

2 Lewis’s theory is spelled out in detail in Lewis (1979). In his formulation of the theory, it is properties that are the contents of belief, where properties are identified with sets of possible individuals. To account for the temporal dimension, it is assumed that it is not coextensive individuals, but time-slices of individuals to which beliefs are ascribed. Given the assumptions of Lewis’s general framework, there will be a one-one correspondence between properties in his sense and sets of centered possible worlds.
such that it is false that pigs can fly in $w$.

Lewis’s account modifies the standard possible worlds account by replacing possible worlds, throughout, with centered worlds. A belief state is a set of centered worlds, and the contents ascribed when one ascribes a belief are represented by sets of centered worlds.

There are a number of problems with this account of self-locating content. First, this account identifies contents that ought to be distinguished. What I believe when I believe that I was born in New Jersey is something about myself, something different from what my fellow New Jersey natives believe about themselves. What I tell the waiter when I tell him that I will have the mushroom soufflé is different from what you tell the waiter if you decide to have the same thing. But on the centered worlds account, our respective beliefs and statements have the same content. Now there are different ways of classifying states of belief, and there is nothing wrong with categorizing belief states so that self-ascriptions of the same property count as the same, in one sense. (If Alice believes that she will win the election, and Bert believes that he will win, then in a sense, they both believe the same thing.) But if one thinks of this as classifying by content, the result is that the contents of belief are not true or false in themselves, but only true or false relative to a speaker or thinker and a time. It is natural to allow that sentence or utterance types might be true or false only relative to a situation in which they are used, but this is natural because it is natural to say that the content expressed by a sentence or utterance type may be different from context to context. It seems less natural to say that the content of a belief might be true for one believer and false for another. Second, and more important, Lewis’s account distinguishes contents that ought to be identified. If Rudolf Lingens tells you that he is sad, or that he is Rudolf Lingens, and you understand and accept what he says, then it seems that the information you acquire is the same information that he imparted. But you do not, of course, thereby ascribe the property of being sad, or of being Rudolf Lingens to yourself. This problem with the account is more significant, since we need to be able to compare and contrast the beliefs and other attitudes of different subjects—to represent agreement and disagreement—in order to understand communication. The Lewis centered-worlds account provides no distinction between a difference in perspective and a disagreement. And the fact that the contents of belief are tied to a time (the time of the center) also makes it more difficult to understand the way the beliefs of a single believer change over time. The account provides no distinction between a change in belief that is a change of mind and a change that results from a change in the facts. (I may stop believing that it is raining because it stops raining, or because I learn that I was mistaken. In the former case, it may be that I still believe what I believed before—that it was raining at the time—and one wants an account of the content of tensed beliefs that allows for this.)

A misleading picture sometimes accompanies the Lewis account of self-locating belief: belief about what possible world you are in is like belief about what country you are in, while beliefs about where in the world you are is like a more specific belief about where, in the country you are (what village, street corner, or mountain top). But ordinary belief about where you are in the world is always also belief about what possible world you are in (what possible state of the world is actual). If I am not sure, as I drive along the highway toward New York, whether I am still in Massachusetts, then I am not sure whether I am in a possible world in which this stretch of highway is located in Massachusetts. If I know that the meeting starts at noon, but not whether it starts now, then I don’t know whether or not I am in a world where I am sitting in my office thinking this thought at noon, or in a world where I am thinking it at some earlier time. The misleading picture is encouraged by the imagery of Lewis’s modal realism (according to which possible worlds are literally places where people are
located), but also by the character of some of the examples used to make the point that beliefs can be irreducibly indexical. Often, to nail the point down, the example will be a case where there are two scenarios within a single possible world involving different people, places or times (two amnesiacs lost in different libraries, or the two omniscient gods). The people in the scenarios know (in each situation) all of the relevant objective facts about them, but remain ignorant of which of the actual situations they are in fact in. That is, they know what world they are in, but not where in it they are. But a story of this kind needs to be highly contrived if it is to work: the internal mental perspectives of the two subjects, or of the subject at the two times, need to be indiscernible from each other. But it is important to note that this science fiction element is entirely unnecessary to make the point that self-locating information is irreducible to information about the objective world. Even if no one else in the actual world is, was or will be experiencing the thoughts and feelings that you are now experiencing, and even if the objective description of the world includes a description of which people are having which thoughts and feelings at which times, you still cannot infer from the objective description where you are in the world, and what time it is now. You need to put the objective information together with your knowledge that it is you who are experiencing these thoughts and feelings, and that it is now that you are experiencing them.\footnote{Lewis makes this point. See Lewis (1979), 138–9.}

2. A MODIFIED CENTERED WORLDS ACCOUNT

In the modification of Lewis’s account that I will propose, it will be assumed that ignorance of where one is in the world is always also ignorance of what world is actual. Even if an experience just like this one is taking place at two times in the actual world, and I don’t know which of the two times is now, the world in which this token experience is taking place at a different time is a different (uncentered) world.

To set up the modified account that I think will give a more adequate representation of self-locating content, let me start by distinguishing two questions:

(1) How should a person’s state of belief as a whole be represented so that it includes his or her self-locating beliefs?

(2) What is the content of a self-locating belief?

In the classical formal semantic models for knowledge and belief that ignores the phenomenon of self-locating, a belief state as a whole is represented by a set of possible worlds (the epistemic or doxastic alternatives), and the content of a particular belief is also represented by a set of worlds: \( x \) believes that \( \phi \) if and only if the set of worlds representing \( x \)’s belief state is included in the set of worlds in which the proposition that \( \phi \) is true. The Lewis account tells a parallel story, with centered worlds replacing worlds, leaving the structure of the theory exactly the same. In both cases, the representation of a belief state, and the contents ascribed, are abstracted from the believer. But in my view, the lesson we should learn from the phenomenon of self-locating belief is that we cannot give an adequate representation of a state of belief without connecting the world as the subject takes it to be with the subject who has the beliefs. What we want to represent is the state of belief that a particular individual \( x \) is in at a particular time \( t \) in a particular possible world \( w \). When we represent the way this individual locates himself in the world as he takes it to be, we need to include the information about who it is who is locating himself there, and we need to link the world as the believer takes it to be to the world in which the believer takes the world to be that way.

To model a knowledge or belief state, I will use the same resources that Lewis uses—centered possible worlds—but the
role of the centering will be slightly different than it is in Lewis’s theory. A belief state will be represented by a pair consisting of a centered world (representing the believer and time and world in which the believer is in the belief state) and a set of centered worlds (representing the ways the world might be, according to that believer, the time that, for all he believes, it might be, and the person that, for all he believes, he might be). Call the world that is the first term of this pair “the base (centered) world.” The centered worlds in the set that is the second term are the belief worlds. The role of the centers is to link the believer, and time of belief, to the possible worlds that are the way that the believer takes the world to be at that time, and to represent where, in those worlds, he takes himself to be.

The information about the belief states of a range of believers at various times in different worlds can be encoded by a doxastic accessibility relation, as in a standard Kripke model for a logic of knowledge or belief (of the kind first developed by Jaakko Hintikka). But in a Hintikka-style model, there is an accessibility relation for each subject, and presumably for each time, whereas in the model I am proposing, the subject and time are built into the relata, rather than the relation. Using a single doxastic accessibility relation, with the identity of the believer and time of belief in the relata, not only allows for the representation of essentially self-locating beliefs, but also facilitates the representation of the relationships between the belief states of different believers (what they agree and disagree about, and what one may believe about another’s beliefs), as well as the ways the beliefs of a single believer change over time.

So far, we have addressed only question (1), about the representation of a belief state as a whole. But when we come to question (2) (what is the content ascribed when one ascribes a particular belief?), the answer will be an ordinary proposition, represented by a set of uncentered possible worlds. What makes it possible to describe belief states unambiguously by ascribing propositional belief is the assumption, mentioned above, that belief about where one is in the world is always also belief about what world one is in. This is the main substantive difference between the kind of model of self-locating belief that I am proposing and the Lewisian models, and it requires some explanation and defense. I need to explain, first, why it is reasonable to make this assumption, and second, what the benefits are of making it.

There are lots of realistic cases of self-locating ignorance or error, but all of them will be cases in which the subject’s specific epistemic situation is unique in the actual world. Even though Alice didn’t know, at noon, that it was then noon, it seems reasonable to believe that she never was, and never again will be, in exactly the epistemic situation that she is in then, with exactly the same experiences and memories that she was having at that time. And it is also reasonable to assume that no other actual person will be in exactly that situation. But in some unrealistic cases that have figured prominently in the discussion of self-locating knowledge and belief, it is assumed that there is, within the (fictive) actual world a duplication of epistemic scenarios. Such cases are standardly represented by using two or more actual scenarios to model the epistemic alternatives for the subject or subjects. I want to argue that even in such artificial cases, this way of modeling the situation is unnecessary and misleading. I will consider two notorious cases: first, the two allegedly omniscient gods in Lewis’s story, cited in the heading of this chapter; second, the case of Sleeping Beauty.

3. LEWIS’S TWO GODS

The two gods, Lewis tells us, are not exactly alike. “One lives on top of the tallest mountain and throws down manna. The other
lives on top of the coldest mountain and throws down thunderbolts.” But they are in identical epistemic situations: “Neither one knows whether he lives on the tallest mountain or the coldest mountain, nor whether he throws down manna or thunderbolts.” Since the two gods are omniscient with respect to propositional knowledge, the epistemic possibilities that model their ignorance must be distinct centered worlds that are differently centered on the same possible worlds.

It is not obvious that the coherence of this story will survive close examination (can different agents perform different actions, without realizing, as they act, which one of them is the agent of which action?), but set this problem aside. Assuming it is coherent, how is Lewis’s story to be reconciled with the assumption I am making, that ignorance about where one is in the world is always ignorance about what possible world is actual? I must say that there are two qualitatively indiscernible worlds, and that neither god knows which of them is actual. This is what Lewis calls the haecceitist response, and he rejects it, first, because he rejects the haecceitist metaphysics. (On Lewis’s view, there cannot be distinct but qualitatively indiscernible possible worlds.) But he also denies that the response will work, even given the haecceitist metaphysics. Here is his argument: suppose we assume that there is a world V that is qualitatively indiscernible from the world W of the story, but with the roles of the two gods reversed. This assumption won’t help, according to Lewis, since we can simply stipulate that the gods know which of these two worlds they are in, and this will still leave them ignorant about where they are in the world. “Let the god on the tallest mountain know that his is world W rather than V. Let him be omniscient about all propositions, not only the qualitative ones. How does this help? Never mind V, where he knows he doesn’t live. There are still two different mountains in W where he might, for all he knows, be living.”

Lewis is right, I think, that the haecceitist move does not eliminate the need to link the believer to the worlds compatible with his or her beliefs, and so does not, by itself, provide an account of the states of ignorance of the two gods. The links are necessary to say how the possible states of the world are being used to represent the epistemic situations of our two characters. But we cannot simply stipulate that the gods know, of a certain possible world, that it is the actual world without saying more about what form their cognitive situations take. Suppose we name the gods Castor and Pollux, Castor being the one who is on the tallest mountain in the actual world W, and on the coldest mountain in the counterfactual alternative, V. How do we inform Castor that it is Castor who lives on the tallest mountain (and so that the actual world is W)?

---

5 Lewis (1979), 139.

6 The best I can do to make sense of Lewis’s story is to liken it to Daniel Dennett’s predicament at a certain point in his marvelous memoir, “Where am I?” (Dennett (1978)). In Dennett’s tale, a body, Fortinbras, is connected remotely to two functionally identical brains, Hubert and Yorick (one electronic and one a human brain that is floating in the proverbial vat). Both brains receive the same perceptual input from Fortinbras, by way of radio signals, though only one of them at a time is able to control him. But since the brains are perfectly synchronized, Fortinbras does what both of his potential controllers simultaneously decide that he shall do, and so it seems to each subject that its decisions are efficacious. Hubert and Yorick are each ignorant of which is controlling Fortinbras (and also of which is the human brain, and which the electronic one). Now suppose we modify Dennett’s story by saying that both synchronized brains are under the illusion that they are controlling two bodies, one on the tallest mountain, and one on the coldest mountain, while in fact one brain is controlling just one of the bodies, and the other the other. Put each brain into the head of the body controlled by that brain. Instead of remote radio connections to send the same perceptual inputs to two brains, we assume that each has divine perceptual capacities to see everything at once, from no particular perspective. Both gods decide both to throw down manna from the tallest mountain and to throw down thunderbolts from the coldest mountain. Each god knows that one of his decisions is efficacious, while the other is causally inert, but neither knows which is which.

This may be a way to make sense of the story, but if our gods engage in speech, this may have strange consequences. The god on the tallest mountain might say truly, and with confidence, “I don’t know which mountain I am on, but I am on the tallest mountain.” What both gods simultaneously decide is that these words shall come out of the mouth of that god. Neither god knows whether what he is trying to say is true, but each knows that if he succeeds in saying it, it will be true. So a sincere and true Moore-paradoxical statement is possible in this unusual situation.

7 Lewis (1979), 141.
If we use the name to tell him, he first must be told to whom the name “Castor” refers, and we might do this in either of two ways (or he himself might fix the reference of the name in either of two different ways): we might say: “Let ‘Castor’ be your name,” and then tell him that Castor is the god on the tallest mountain. (Or he might say to himself, “I hereby dub myself ‘Castor’”, and we then tell him where Castor is located.) Alternatively, we (or he) might say: “Let ‘Castor’ be the name of the god on the tallest mountain.” On the first alternative, we have resolved Castor’s doubts about his location in the world. In the second, we have not, but this is because he does not know of Castor that “Castor” is his name, and so does not yet know whether the actual world is W or V (this is a case like Gareth Evans’s case of Julius*).

The haecceitist move does raise delicate questions about the relation between metaphysical issues about the world as it is in itself and issues about a subject’s perspective on the world, just the issues our model is designed to help clarify. We don’t want to rest a metaphysical distinction solely on its utility for the representation of knowledge and belief. But I think the haecceitist move can be justified on metaphysical grounds. Aside from the representation of cognitive states, it seems plausible (or as plausible as things get in this fanciful story) that it might be a contingent fact that our two gods have the properties that distinguish them: a contingent fact that the god who is on the tallest mountain is on that mountain, rather than the coldest mountain, where the other god in fact is, and that the one throws down thunderbolts, rather than manna, etc. The two gods, it seems, might have interchanged their positions and roles. Or, if this is not metaphysically possible because there are facts, perhaps about the origins of each of the two gods that are essential to them, and known by both of the gods, it still seems plausible to say that there might have been other individuals playing just these roles, and this is enough to suggests that there might have been qualitatively indiscernible possible worlds.

Note that this discussion brings out that the issue about self-locating knowledge and belief is not as tightly connected to semantic issues about personal pronouns and demonstratives as is sometimes supposed. If our gods give themselves names, fixing the referents demonstratively, then they can express their knowledge and ignorance using sentences that are semantically context-independent. (“I know that I am Pollux,” says Pollux, “but I still don’t know, in a sense, who I am, since I don’t know which mountain Pollux lives on.”) The introduction of the names does make it possible for others to express the same propositions with the same words, but it does not essentially change the epistemic situation of the relevant subjects. And of course names generally are most often introduced demonstratively. As is emphasized in both Perry’s discussions of reflexive content, and in my uses of the diagonalization strategy, to explain cases of ignorance of the truth of identity statements involving proper names, we need to bring in information, not just about the referent of the names, but also about the facts that connect those names to their referent.

4. SLEEPING BEAUTY

Second, let me consider a case involving temporal self-location in which an individual may (for all she knows) be in essentially the same epistemic situation at two different actual times. Here is the scenario: Sleeping Beauty is to be put to sleep on Sunday night after being told that she will be woken up either once or twice in the next two days, depending on the flip of a fair coin. If heads, she is woken up only once, on Monday, and if tails, she

---

* “Julius” is a proper name, stipulated to be the name of the inventor of the zip, whoever he or she might be. “Julius” is (by stipulation) a rigid designator, so the proposition that Julius invented the zip is contingent, but one may still know a priori (in virtue of the stipulation) that Julius invented the zip, if any one person did. “Julius” made its first appearance in Evans (1979). I discuss the case in Stalnaker (2001).
will be woken up on Monday and again on Tuesday, but only after being given a drug that ensures that she will have no memory of the Monday waking. The question is, to what degree should Sleeping Beauty believe, upon being woken up on Monday, that the coin will, or did, land heads. Adam Elga, who introduced this puzzle to the philosophical literature,\(^9\) defends the answer one third, while David Lewis argued that the rational degree of belief is one half.\(^10\) The argument between Elga and Lewis is carried out within Lewis’s centered-world framework for representing self-locating belief, and there are some presuppositions they share that I think should be questioned: they both assume that two of Sleeping Beauty’s epistemic possibilities (when she wakes up on Monday) should be represented by two situations within the same possible world, and I think this assumption distorts the discussion. While I think Elga gets the right answer to the question posed by the puzzle, I want to use a slightly different argument to defend it.

Let’s begin by describing Sleeping Beauty’s epistemic situation when she wakes up on Monday. (Of course she does not then know it is Monday. We are describing the situation from the theorist’s point of view.) There are three possibilities compatible with her knowledge at that point, which she would describe this way: “Either today is Monday, and the coin will land heads (call this state s1), or today is Monday and the coin will land tails (s2), or today is Tuesday, and the coin landed tails (s3).” Lewis and Elga both assume that (s2) and (s3) are alternative scenarios or predicaments within the same possible world. That is, they assume that the only fact about the world as it is in itself that is relevant is whether the coin lands heads or tails. Were Beauty to learn that fact, she would have all the relevant information about what possible world she was in; her remaining ignorance would be about where, within that world, she was. The first thing to note about this way of modeling the situation is that it requires that Sleeping Beauty be in precisely the same epistemic situation on the two different days (on the assumption that the coin landed tails). It cannot be that it is slightly darker, or lighter, in the room in which she wakes up on one of the two days, or that the exact arrangement of the bedcovers that she sees as she wakes up is slightly different, or that she hears a dog bark in the distance on one day, but not the other. Sometimes a science fiction variant of the story is told in which an exact duplicate of Sleeping Beauty, as she was on Monday, is created on Tuesday, and while it does not seem essential to the problem posed by the story that it have this feature, it is essential to the way of modeling it used by both Elga and Lewis. But as the story is usually told, all that can be assumed is that she receive no information, on waking up on Monday (or Tuesday), that is relevant to the result of the coin toss, beyond the information that she wakes up, and we need only this weaker and more realistic stipulation if we assume that the epistemic alternatives are different possible worlds. (This assumption does not exclude the limiting case in which there are no differences, relevant or irrelevant, between Sleeping Beauty’s situations at two actual times. “Even if the way things seem on the other day is exactly like this, down to the last detail,” she might think (on both days), “it will still be not be this token experience that I was, or will be, having then.”)

The account I am promoting says that the descriptions of the possibilities, (s1), (s2), and (s3), give the relevant information about three different possible worlds, viewed from a certain perspective. On this view, the perspective is essential to the description of the worlds, but not to the world-states described. In two of these possible states of the world ((s2) and (s3)), there is a similar event taking place either the day before or the day after the time that specifies the perspective, but the waking that will take place tomorrow (as Beauty would have put it at the time) in (s2) is a different waking from the one that is taking place now in (s3), and so scenario (s2)

---

\(^9\) Elga (2000). As Elga notes, the problem has its source in a discussion in the game theory literature about games with imperfect recall.

\(^10\) Lewis (2001).
must be a different (uncentered) world from (s3). As we will see, this difference makes a difference to the argument.

There is a fourth possible state of the world that is not compatible with Sleeping Beauty’s knowledge on Monday, but that she might describe, from her perspective, by means of two suppositions, one of which is counterfactual: “Suppose”, she thinks, “that today is in fact Tuesday, but that instead of coming up tails, the coin had landed heads. I know that this is not the actual situation, because if it were, I would not have been woken up today, and I was.” (Even if it is in fact Monday at the time Sleeping Beauty is thinking this particular thought, as we are supposing, she can still characterize this possibility as the one that is like the epistemically possible world in which (as she would put it) “Today is Tuesday”, but in which the coin landed heads, instead of tails.)

The relevance of this fourth possible world to Sleeping Beauty’s epistemic situation is that, while she is on Monday in a position to rule it out, she knows that on Sunday, she was not in position to rule it out. Of course on Sunday she was not in a position to characterize any of these possible situations by fixing the reference of the day in the way that she does fix it (on Monday) as today, or as the day in which this thought is being entertained. But we can still use these four possibilities to characterize Sleeping Beauty’s prior state of knowledge, her epistemic situation on Sunday, doing it in a way that is relevant to connecting her knowledge on Sunday with her knowledge on Monday. Specifically, we can say what information it is that Sleeping Beauty acquires when she wakes up on Monday: what possibility that was previously compatible with her knowledge is now incompatible with it.11

Elga claimed that Sleeping Beauty received no new information upon waking up on Monday, and so that her rational change in belief (from one half to one third) in the proposition that the

11 The point that Sleeping Beauty gains knowledge when she wakes up is argued in Weintraub (2004) and Morgan (2004).

coin would land heads, was a change induced by something other than new information. Elga claimed this because he identified new information with objective information, and in the centered-worlds framework that he was using, Sleeping Beauty learns nothing about what the world is like in itself. Elga’s assumption that there could be rational belief change without new information was the main reason that David Lewis resisted his conclusion. According to Lewis, since Sleeping Beauty receives no new information, her degree of belief in heads must remain the same. But on the analysis I am proposing, one can agree with Lewis that new information is required for a rational belief change, but also with Elga that Sleeping Beauty’s degree of belief in the result of the coin flip should change when she wakes up.

But does Sleeping Beauty really learn something—does she really rule out a possibility previously compatible with her knowledge—simply by being woken up on Monday (something she knew in advance would happen)? Suppose her best friend, Sleeping Ugly, accompanied Beauty on her adventure. He, let us suppose, will be woken up, put to sleep, and given the same amnesia inducing drug as Sleeping Beauty, but Sleeping Ugly will be awakened on Tuesday as well as Monday, whatever happens with the coin (and he will observe, each time, whether or not she is awakened). Surely Sleeping Ugly learns something upon waking up on Monday: he learns that Sleeping Beauty was also awakened. Does he learn anything that she does not learn? If we allow him to tell her what he has learned, is there anything he can tell her, when they both wake up, that she does not already know? Obviously not.

On the analysis I am promoting, the strategy for determining exactly what Sleeping Beauty’s degrees of belief should be, when she wakes up, is to start by determining how her degrees of belief should be apportioned on Sunday between the possibilities that are open to her at that time. All parties to the dispute should be able to agree about this. Then her degrees of belief on Monday (and/or
Tuesday) will be determined simply by conditionalizing on the information that she receives on waking up. What is required for this strategy to work is that the possibilities that are relevant to representing her beliefs on Monday be a subset of those that represent her beliefs on Sunday, or more generally, that one be able to calibrate the informational states that she is in at the different times by characterizing them as subsets of the same set of possibilities. The unreconstructed Hintikka-style models of cognitive states took calibration for granted, but ignored belief change and self-location. The Lewis-centered-worlds models recognized self-location, but provided no resources for representing the relations between informational states across time and across persons, and so no resources for clarifying the dynamics of knowledge and belief, or the communication of information between different subjects. The general framework that I am promoting allows for calibration across time, and across different subjects, but it also recognizes that calibration is a nontrivial problem, and may not be well defined in all cases.

5. COMPARING BELIEFS OVER TIME, AND BETWEEN DIFFERENT SUBJECTS

Sleeping Beauty remembers, on Monday, what she knew on Sunday, and that she knew it. We can represent this fact in a straightforward way by taking the possibilities compatible with her knowledge (on Monday) as full possible worlds, worlds in which she was in one cognitive state on Sunday, and another one on Monday. The information that she had at the earlier time can be preserved at the later time even if it was self-locating information. And by using the same possible worlds to represent the informational states of different subjects, we can say when they agree and disagree, even about self-locating information, and we can represent iterated attitudes—what one person believes about the beliefs of another person. Lingens the amnesiac, for example, can consider, not only who he might be, but also what others might know or believe about who he is, and others may have beliefs about what Lingens does and does not know about his identity. To compare the information, including self-locating information, that is available to different subjects, and to represent what they know and believe about each other, we (as theorists) need to use the same possible worlds to model the various informational states, and the different ways in which we, and they, locate themselves in those possible worlds.

Since different subjects inevitably have different beliefs, and each has different beliefs at different times, comparisons of the contents of informational states across persons and times may require that subjects locate themselves in possible worlds that are not compatible with their present states of knowledge or belief. Self-locating information can exclude a certain possibility only if the subject locates herself in the situation that is excluded. Self-location in a counterfactual possible world was essential to the account of the Sleeping Beauty case, where she excluded (on Monday) the possible world in which it was then Tuesday, and the coin landed heads. Or consider Lingens, who has been doing research in the Stanford Library on various missing persons who he thinks he might be. "I still don't know who I am," he thinks, "but I know that I can't be Gustav Lauben, since if I were, I would have a tattoo of a mermaid on my right arm, and I don't." Here he locates himself in a counterfactual world, and as a result excludes it. Perhaps it was Daniels who suggested that he look into the Lauben possibility. "Daniels thought that I might be Lauben," thinks Lingens. Here he locates himself in a world that is compatible with what he takes Daniels's beliefs to be, but that is not compatible with his own beliefs.

Daniels, like Lingens, has read about Lingens, and just as Lingens does not know whether he himself is Lingens, Daniels doesn’t know whether the famous Stanford amnesiac, the guy he has recently been talking to, the guy he thought might have been Lauben, that guy (as he might put it to himself, recalling the conversation), is Lingens. Lingens’s ignorance of who he himself is is essentially self-locating, but Daniels has no problem with self-location. His ignorance about the identity of that guy is more like a standard Frege case, for example like Ralph’s ignorance about Orcutt in Quine’s example—his failure to know that the pillar of the community that he sees on the beach and the suspicious man in the brown hat, glimpsed in the bar, are one and the same person. Nevertheless, it seems that what Lingens doesn’t know (that he is Lingens) is the same fact as the fact that Daniels doesn’t know. If Daniels found out who the famous Stanford amnesiac was, he could tell Lingens in a straightforward way: When he says to him, “you are Rudolf Lingens” his “you” picks out, in each of the relevant possible worlds, the same person picked out by “I” in the mouth of Lingens (and the person that both pronouns pick out will be, in the possible worlds that this statement excludes, someone other than Lingens). We can capture this identity of content by representing their separate informational states, and their attitudes about each other, with the same possible worlds in which each locates himself and the other.

In cases of ignorance or confusion about identity (as in the case of Ralph and Orcutt), it is not always clear where a subject should locate himself in the possible worlds compatible with the beliefs of someone else. Suppose Orcutt knows about Ralph’s confusions about him. Would he be correctly describing the situation if he said “Ralph believes that I am a spy?” In the world as Ralph takes it to be (and as Orcutt takes Ralph to take it to be), should Orcutt identify himself as the man on the beach, or as the distinct man in the brown hat in the bar? Orcutt’s problem here is the same as ours when we ask whether Ralph believes the singular proposition, about Orcutt, that he is a spy. The right answer, I think, for both Orcutt and for us, is that the attribution of that belief is correct in some contexts, but not in others.14 But if we imagine a context in which Ralph and Orcutt are discussing the situation, face to face, the answer will be clear. If the conversation takes place on the beach, with Ralph taking himself to be talking to the pillar of the community, who is not a spy, then it will clear that this is where Orcutt locates himself in the world according to Ralph. He could sincerely say to Ralph, “you don’t think I am a spy”, and his statement would be unambiguously true.15

We can model a conversational context by a set of possible worlds that represents the common ground of the participants in the conversation—the information that they take to be shared between them. Like common knowledge, or mutual belief, common ground is an infinitely iterated attitude: roughly, a proposition is common ground in a group if each accepts it, each accepts that each accepts it, each accepts that each accepts that each accepts it, etc.16 Since our models of self-locating belief can represent iterated belief, this kind of model of a context can accommodate self-locating communication. So long as the participants have a common way of identifying each other (as participants in the conversation), the iteration involved in defining the common

14 Given our general framework, this means that it is right in some contexts to represent Ralph’s beliefs with a set of possible worlds in all of which Orcutt is a spy, while in other contexts it is wrong to use this set of possible worlds to represent his beliefs. See Stalnaker (1988) and Stalnaker (2008) for discussions of the context-dependence of de re belief ascriptions.

15 Though if we tweak the context a bit, we could make it possible for Orcutt also to say “you don’t realize it, but you know me under a different guise, not realizing that that person is me. In that guise, you think I am a spy.” Cf. Crimmins (1992).

16 See Stalnaker (2002) for a discussion of this representation of context, though the issues about self-locating are ignored in that discussion.
ground will give rise to a model of an informational state with a set of centered worlds with multiple individuals—all of the participants in the conversation—at the centers.\(^{17}\)

The occasion for this digression on the general problem of self-locating attitudes was the suggestion that knowledge of the phenomenal character of experience was in some way like knowledge of who and where we are in the world, and that the problem posed by Mary's predicament might be solved by clarifying the analogy. We will return, in the next chapter, to the puzzle about Mary, exploring this analogy in the context of our general framework. In later chapters, we will use this framework to try to throw light on puzzles about a subject's knowledge of the contents of his or her own thought.

\(^{17}\) See the appendix to this chapter for a few more details.

**Appendix: Notes on models of self-locating belief**

This is a brief sketch of a few details of a formal model, with a little motivation and commentary. The models use exactly the same abstract objects used in David Lewis's theory of de se belief (centered worlds), to characterize belief states, but uses them in a somewhat different way.

A model is a sextuple \(<W, S, T, \geq, E, R>\) where

1. \(W\) is a nonempty set of possible worlds
2. \(S\) is a set of *subjects* or believers
3. \(T\) is a set of times
4. \(\geq\) is a binary transitive connected anti-symmetric relation on \(T\), a relation that determines a linear order of the times.

\(T\) would most naturally have the structure of the points on the real line, but in simple models, we might choose to represent only the beliefs of our subjects at certain selected times, so the linear time order might be discrete, and the number of times might be finite. But it is assumed that the ordering is an objective time ordering, and that times can be identified across possible worlds. That is, a certain date (such as Tuesday, April 3, 2007) might be the date on which it rained in Oxford in certain possible worlds, and was sunny there in others.

Two definitions, before characterizing \(E\) and \(R\):

a. A *center* is a pair, \(<A, t>\), where \(A \in S\) and \(t \in T\).

b. A centered world is a pair \(<c, w>\), where \(c\) is a center and \(w \in W\).

5. \(E\) is the set of centered worlds meeting the condition that the subject of the center exists in the world at the time of the center.

6. \(R\) is a binary relation on \(E\) that is transitive, Euclidean and serial. \(R\) must also satisfy an additional condition, which we will state and explain below.

The interpretation of the fifth and sixth elements, \(E\) and \(R\), is this: subjects may exist at some times at some worlds, and not at others. The set \(E\) of centered worlds is restricted to those that are relevant to representing a subject's beliefs at a time in a world. The relation \(R\) is the doxastic
accessibility relation. To say that \(<<A,t>, x>R<<B,t'>, y>\) is to say that it is compatible with what A believes at time t in world x that she is in world y, that she is person B, and that the time is time t'.

Given R, each centered world in E determines a set of centered worlds—those that are R-related to it. Call a pair consisting of a centered world and its R-related set a belief state, and call the determining centered world the base (centered) world, and the determined set the belief set. The role of the center of the base world is to specify the person whose beliefs are being represented, and the time at which she has those beliefs. The role of the centers of the centered worlds in the belief set is to represent where that subject takes herself to be in the world that, for all she believes, is actual. If Alice thinks, on Sunday, that it might be Monday, and that she might be Clara, rather than Alice, then a world centered on Clara on Monday will be compatible with what she believes (on Sunday).

We impose the following condition on the relation R:

(\(^*\)) For any centers, c, c' and c'', and worlds w and x: if \(<c,w>R<c',x>\) and \(<c,w>R<c'',x>\), then \(c' = c''\).

What this condition requires, intuitively, is that ignorance or uncertainty about where one is in the world is always also ignorance or uncertainty about what world one is in. Even in the highly artificial case where a subject believes that he will, in the actual world, be in two qualitatively indistinguishable situations at different times, \(t_1\) and \(t_2\), without knowing which time it is, it will remain true that (as he would put it at the time) the world where this token thought is occurring at time \(t_1\) (and where another like it will occur at \(t_2\)) is a different (uncentered) possible world from the possible world in which this (token) thought is occurring at time \(t_2\) (and another like it occurred at \(t_1\)).

This crucial condition is the main point at which the proposed model differs formally from Lewis’s account of de se belief, which allows that a case of ignorance might be represented by two centered worlds—two "predicaments", to use Adam Elga's term—centered at different points within the same world. By requiring that ignorance and doubt always be represented by distinctions between possible states of the world, we allow for the calibration of the states of belief of different believers, and of a believer at different times. Even though belief states are represented by sets of centered possible worlds, the contents of belief can be taken to be ordinary propositions—sets of uncentered possible worlds. So in the interpretation of statements of the form "x believes that \(\phi'\)," the "that \(\phi'\)" will denote a set of (uncentered) possible worlds, even though the centers determined by a particular belief state may play a role in determining which proposition is denoted by a that-clause with indexical expressions in it. By taking the contents of belief to be (uncentered) propositions, we can straightforwardly compare the beliefs of different subjects, and we can model the way assertions change the context in a straightforward way. We can also model the dynamics of belief for a single agent—the facts about preservation and change of belief—in a straightforward way. In particular, we can apply a standard belief revision theory to a rational subject with a prior belief state at time t, who then receives some new information at time \(t^*\) while remembering her prior state. Even if some of her prior and posterior information is self-locating (suppose, for example, she didn’t know what time it was at t, or how much time passed between \(t\) and \(t^*\)), she can still revise her beliefs in the standard way. If we want to add to our model probability measures on belief states to represent degrees of belief, this will be as straightforward as in standard belief-logic models, and we could then represent the assumption that rational subjects will revise by conditionalization.

In the standard Hintikka-style semantics for logics of knowledge and belief, ordinary uncentered possible worlds are the relata of the doxastic or epistemic accessibility relations. The identity of the believer, and (implicitly) the time of belief are built into the relation. In a theory of this standard kind with multiple believers, there will be multiple accessibility relations, one for each believer. Our models, in contrast, need only a single doxastic accessibility relation, since the identity of the believer and the time of the belief are determined by the center of the first relatum. By putting the believer and the time of belief into the relata, rather than the relation, we not only provide the resources to represent self-locating belief, but also a more flexible framework for representing the relation between the beliefs of different believers, and of a single believer at different times.

In the standard belief semantics, the representation of iterated belief is a simple matter: If A and B are two believers, and R_A and R_B are their

---

1 I don’t want to rest anything on the assumption that the same token thought might have occurred on a different day. It might be a counterpart token that occurred, in the other possible world, on the other day. What one needs to motivate the assumption that there are two (uncentered) possible worlds here is just that the thought that takes place, in the actual world, at the other time is a different token thought.

---

70 ~ Appendix: Notes on models

Appendix: Notes on models ~ 71
doxastic accessibility relations, then it will be true, in world w, that A believes that B believes that \( \phi \) iff for all worlds x such that \( wR_A x \) and all worlds y such that \( xR_A y \), \( \phi \) is true in y. One can define the set of possibilities compatible with the common beliefs of A and B in terms of the transitive closure of the two relations \( R_A \) and \( R_B \), or more generally, the common beliefs of a set of subjects in terms of the transitive closure of the set of accessibility relations for the subjects in the set. In our models, the representation of iterated belief is a little more complicated, but the complications reflect complexities in the phenomena being modeled, and the increased flexibility in the representational resources of the model.

The first complication comes from the fact that we have made explicit that belief is relative to time, something that is ignored in the standard theory. One might represent A’s beliefs at t about what B believes at some different time \( t' \), but let’s ignore that for now, and just focus on A’s beliefs at some time t about what B believes at the same time. Still, A may not know what time it is, so the actual time at which A has her beliefs may be different from the time she takes it to be. For example, if A mistakenly believes on Tuesday that it is Monday, then there will be a difference between “A believes (on Tuesday) that B now believes that \( \phi \)” and “A believes (on Tuesday) that B believes on Tuesday that \( \phi \).” The truth of the former will depend on what B believes on Monday in the worlds compatible with A’s beliefs, while the latter will depend on what B believes on Tuesday in those same worlds.

A second complication is this: Because of the intentionality of belief, A may have different beliefs about B’s beliefs, relative to different ways of thinking about him. Suppose I am sitting in the bar with a man in a brown hat who is in fact Orcutt, but I am not sure whether he is Orcutt or O’Leary. We are watching the Red Sox on the television, and I believe that the man in the brown hat believes that the Red Sox are losing, since they are losing, and it is evident that the man is paying attention to the game. But I am not sure whether Orcutt believes this, since for all I know, the man at the bar is O’Leary, and Orcutt is somewhere else, blissfully ignorant of the state of the game.

In the simple case, where it is assumed that A knows who B is, we can ignore this, but for the general case, we need to relativize iterated belief, (what A believes about what B believes), to a way that A thinks of B (dare I call it a mode of presentation?), formally represented by a function from worlds to individuals. A function of this kind will represent (in a given world) a possible way of thinking about B if it takes B as its value in that world. In the simple case (where A knows who B is), this function will be the constant function, taking B for all arguments, but in the general case, it might be a variable, or non-rigid, individual concept. (We can assume that the function is everywhere defined within the worlds in A’s belief set, since we can assume that if A identifies B as “the F”, then A believes that there is a unique F.)

To use nonrigid functions, or individual concepts, to characterize the centered worlds does not add any new centered worlds to our model: it just gives us new ways to generalize about them. Suppose \( f \) is the non-rigid function, or individual concept, expressed by “the man in the brown hat”, and that \( f(w) = \) Orcutt. Then the centered world \( << f, t, w >> \) is just the centered world \( << Orcutt, t, w >> \). But when we quantify over centered worlds, \( f \) may take different values for different values of \( w \). For example, consider this generalization:

For all worlds x and y and for all subjects C, if \( << A, t, w >> \) R \( << A, t, x >> \) and \( << f, t, x >> \) R \( << C, t, y >> \), then \( y \in f \).

This says that in world w, A believes that the man in the brown hat believes that \( f \).

Once we have a clear account of iterated belief, we can use it to define a notion of common belief for a group of individuals at a given time and the properties of a common belief state will be generated by the iterative process. It is common belief (among the members of group G) that \( \phi \) iff all believe that \( \phi \), all believe that all believe that \( \phi \), all believe that all believe . . . , etc. To keep things simple, we might assume that everyone in the group knows who everyone else in the group is, but we can also model cases where the members of a group have some common way of identifying each other, even though they may not know who the others, or even themselves, are. So, for example, we might model the common ground (presumed common beliefs) of a conversation between two amnesiacs trying to figure out who and where they are, and what time it is, by pooling the meager information that they each have. In general, the common ground that is determined by the iterative process will generate a representation that parallels the representation of an individual belief state; it will have the same structure, but with centered worlds with multiple individuals at their centers. An individual belief state is a pair consisting of a centered world (the base world) and a set of centered worlds (the belief set). The common ground can also be represented by a base world and a common belief set, but with a sequence of individuals (all
those in the relevant group) at the centers instead of a single individual. The sequences of individuals at the centers of the common belief worlds will represent where the members of the group mutually locate themselves and each other in the possible worlds compatible with their common beliefs.

Our models have an accessibility relation only for belief, but a subject might also have other self-locating attitudes. In some cases, self-location in possible worlds that are not compatible with the subject's belief is derivative from self-location in belief-worlds. (This fact was exploited in the account of the Sleeping Beauty case.) Suppose I don't know whether I am A or B, but I do know that if the coin had landed heads (which I know it did not), then I would have won the bet. What I know is that if I am in fact A, then if the coin had landed heads, A would have won, and if I am in fact B, then if the coin had landed heads B would have won the bet. A second kind of case of derivative self-location in worlds incompatible with the subject's beliefs is iterated belief. To represent my belief that John believes that I am a plumber, I need to locate myself in the possible worlds that, for all I believe, are compatible with what John believes. If John knows me in different guises (or if my beliefs about him allow for this possibility) then my self-location in the worlds as they are according to John will be relativized to one of them.²

² Thanks to Agustin Rayo and Seth Yalcin, both of whom gave me invaluable advice about the ideas developed in these notes.

4

Phenomenal and Epistemic Indistinguishability

In the pure phenomenal case, . . . the referent of the concept is somehow present inside the concept's sense in a way that is much stronger than in the usual case of direct reference.

David Chalmers

My aim in this chapter is to explore the analogy between essentially self-locating knowledge and knowledge of phenomenal experience, but let me begin by reminding you where we are. The debates about Mary and the knowledge argument raised questions about the extent to which features of our representations of certain facts (facts about phenomenal experience, "what it is like") belong to a conception of the world as it is in itself (to what Bernard Williams called an "absolute conception"). The alternative is that the distinctive facts about phenomenal experience should be understood as features of our perspective on the world, facts that essentially involve the relation between a representation and something being represented. Most of the materialist strategies for responding to the knowledge argument aim, in one way or another, to

¹ Chalmers (2003), 233.