The essential contextual

The MIT Faculty has made this article openly available. Please share how this access benefits you. Your story matters.

<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>As Published</td>
<td>10.1093/acprof:oso/9780199573004.003.0006</td>
</tr>
<tr>
<td>Publisher</td>
<td>Oxford University Press</td>
</tr>
<tr>
<td>Version</td>
<td>Author’s final manuscript</td>
</tr>
<tr>
<td>Citable link</td>
<td><a href="https://hdl.handle.net/1721.1/125848">https://hdl.handle.net/1721.1/125848</a></td>
</tr>
<tr>
<td>Terms of Use</td>
<td>Creative Commons Attribution-Noncommercial-Share Alike</td>
</tr>
<tr>
<td>Detailed Terms</td>
<td><a href="http://creativecommons.org/licenses/by-nc-sa/4.0/">http://creativecommons.org/licenses/by-nc-sa/4.0/</a></td>
</tr>
</tbody>
</table>
Linguists have to give an account of sentences like “It is raining”, but they have no professional interest in meteorology. The situation is different with the notions of knowledge, belief, and other propositional attitudes. Semantics must account for sentences that express and describe propositional acts and attitudes, but these notions also play a role in the foundations of semantics, and in a general account of what language is used to do. The semanticist’s job is to give an account of what is said in speech, and how what is said depends on the meanings of the expressions of the language, how they are put together, and on the contexts in which they are used. What is said, it seems reasonable to believe, is information conveyed, and should be the same kind of thing as what it is that is known or believed. And the contexts in which things are said may be represented by informational states – the shared information, or presumed common knowledge of the participants in the conversation. In representing both what is presupposed and what is asserted in a context, we represent propositional attitudes and the way they change in the course of a conversation.

A standard formal model for this idea of context represents context, in one sense, as a set of possible worlds – the context set. (of possible states of the world), interpreted in terms of a certain complex attitude or epistemic state: the presumed common knowledge of the participants, the information that they take themselves to share. The context set is the set of possible states of the world that are compatible with the shared information, the alternative states of the world that the participants mean to distinguish between in their conversation. There is more to be said about exactly what this epistemic state is, and exactly how it relates to the knowledge and beliefs of the individuals, but the key idea is that it is an iterated attitude, with the structure of common knowledge. (It is common knowledge that \( \phi \) between the members of some group if and only if all know that \( \phi \), all know that all know that \( \phi \), all know that all know that all know that all know, . . . etc.) But common ground need not be knowledge, since one may presuppose things that are false. And it need not even be belief, since there may be pretense involved. I’m not going to worry here about exactly what the basic attitude is; the crucial point is the iterated structure. The main problem that I want to focus on in this paper is that the phenomenon of indexical or self-locating belief suggests that the simple idea of belief as a propositional attitude, where propositions are understood in terms of (absolute) truth conditions, fails to give an adequate account of information, not about what the world is like, but about where one is located in the world. This kind of information can be communicated: it can be the content of what is said, or meant, and it can also be presupposed. We need an account of context, and of a context set, that represents the setting in which self-locating communication takes place.

So we will be looking at the problem of self-locating belief, with the aim of finding a

---

1 See the papers in Stalnaker (1999) on presupposition and assertion for my early applications of the notion of a context set. See Stalnaker (2002) for a more recent discussion that is more explicit about the iterated structure.
solution to it that will allow for an account of self-locating contextual information and communication. My plan is this: I will start with a sketch of the classical possible-worlds representation of a state of belief or knowledge as a set of possible worlds, a formal semantic framework first applied to knowledge and belief by Jaakko Hintikka more than forty years ago, and that since then has been extended and widely used in linguistic semantics, computer science, and game theory, as well as in epistemology. This framework provides a straightforward representation of iterated knowledge and belief, and of common knowledge and mutual belief. Second, I will remind you of the problem of self-locating or essentially indexical belief, a kind of attitude that the classical framework has trouble accounting for. I will sketch David Lewis’s modification of the classical analysis, a modification that remains within the possible worlds framework. Lewis’s theory has the resources to represent essentially self-locating attitudes, but it is, in a sense, a static and solipsistic representation, and so does not provide for an account of change of belief over time, or of the interaction of the attitudes of different subjects. So third, I will suggest a modification of Lewis’s account, one that uses formal tools from both the Lewis and the Hintikka frameworks, and show how this modified theory can represent the contexts in which exchange of self-locating information takes place. I will conclude by applying this kind of model to a notorious puzzle case, Mark Richard’s phone booth story, an example that involves multiple identity confusions, self-location, and context-dependence in a discourse situation.

1. The classical account

In the classical theory, a belief or knowledge state is represented by a set of possible worlds. The intuitive idea is to represent a subject’s cognitive situation by the set of possible states of a world that from the subject’s perspective might be the way the actual world is. The point is not to provide a substantive analysis of knowledge and belief, but to represent their abstract structure in a revealing way. One does not say what it is to be a doxastic or epistemic alternative possible world; one simply assumes it by specifying, as a primitive component of the model, a binary relation between possible worlds that determines a set of alternative possible worlds as a function of each world. What is being modeled is the beliefs or knowledge that a given subject has (at a given time, which usually goes unmentioned) in a given possible world. The relation holds between possible worlds x and y if and only if the subject’s state of knowledge in world x, at the relevant time, is compatible with possible world y being the actual world. That is, if xRy, then for all the subject knows or believes in world x, she is in world y. The theory will put various constraints on the relation of doxastic or epistemic accessibility, which will correspond to assumptions about the logic of knowledge or belief.

Hintikka focused on models of the attitudes of single subjects, but much of the interest in this kind of model comes from its account of the interaction of the attitudes of different subjects: what I know or don’t know about what you know, including about what I know about what you know about what I know, etc. Even though Hintikka did not talk about

---

the interaction of the attitudes of different subjects, his models generalize without further assumptions or problems: one just has an epistemic accessibility relation with the appropriate properties indexed to each knower. A model will yield, for each proposition $\phi$ and knower $A$, a proposition that $A$ knows that $\phi$, and so claims about knowledge straightforwardly iterate, both for the same subject, and for different subjects. For any given proposition $\phi$, we get, for example, the propositions that $A$ knows that $A$ knows that $\phi$, that $A$ knows that $B$ does not know that $\phi$, that $A$ knows that either $B$ knows that $\phi$ or else $C$ does, but doesn’t know which. And one can define a notion of common knowledge for a pair or group of knowers, in terms of the transitive closure of the epistemic accessibility relations for the different members of the pair or group. It is common knowledge that has received the most attention, but an infinitely iterated extension of belief, or any acceptance concept can be defined in an exactly analogous way, with a semantics in terms of the transitive closure of a set of accessibility relations of the appropriate kind.

This kind of representation is highly idealized, but it has proved useful for clarifying issues in epistemology, and in applications. The idealization does not avoid or evade the particular problems we want to focus on, and will help to sharpen the issues. Our problem will be how to generalize or modify the classical account to allow for essentially indexical or self-locating belief and knowledge.

2. Essentially indexical attitudes

Worries about essentially indexical attitudes – beliefs, not about what the world is like in itself, but about where one is located in the world – go back at least to Hector-Neri Casteneda’s work in the 1960’s, but they were brought back onto the philosophical stage by John Perry’s classic papers published in the late 1970’s.\(^3\) The thesis was that knowledge and belief about who one is, and what time it is now are not reducible to knowledge and belief about the impersonal and timeless features of the world. The point was made with a series of stories, some mundane, others fanciful: a person in a grocery story chasing a trail of sugar on the floor that was obviously leaking from a sack in someone’s cart, only to discover that the sugar spiller was he himself; a man getting on a trolley who sees a “shabby pedagogue” at the other end, but then realizes that it is he himself in a mirror; an amnesiac, lost in the Stanford library, reading a biography that is in fact of himself, and so learning a lot about himself without knowing that what he is learning is about himself; two omniscient gods who know everything about what the world is like in itself, but that are each ignorant of who he himself is. These stories were all cases where objective knowledge about the subject came apart from first personal knowledge of the subject, and formed the basis for arguments that the one kind of knowledge could not be reduced to the other. There were also stories of a different kind that made the case for essentially indexical attitudes in a different way. These were stories in which two agents had all the same relevant (objective) beliefs and desires, but nevertheless were rationally motivated to act in different ways because of their different

---

\(^3\) Perry (1977) and (1979).
perspectives on the world. John Perry’s classic example: you are being chased by a bear; I run for help, while you curl up in a ball. We both have exactly the same beliefs about the situation, and about the most effective way for each of us to respond to it. We also have exactly the same motivation: to save you from the bear. The only differences in our attitudes that might explain our different actions are differences in self-locating attitudes.

David Lewis developed a very elegant formal semantic theory – a generalization of the standard possible worlds representation of a state of belief – to allow for this distinctive kind of attitude. The idea was to use a different and more fine-grained object to represent the content of an attitude. Rather than propositions, represented by sets of possible worlds (those in which the proposition is true), we are to use sets of centered possible worlds, which are pairs consisting of a possible world and a designated time and person within that world. The world component of a centered world represents an objective possible situation that is compatible with the subject’s conception of the way the objective world is. The person at the center represents the (objectively identified) person that the subject thinks she might be, in a world of that kind. The (objectively identified) time of the center represents a time that, for all the subject believes, might be the time she is in the belief state, in that possible world. So, for example, since the amnesiac Lingens knows (from reading the biography) that Lingens was born in 1953, the possible worlds compatible with his beliefs will all be worlds in which Lingens was born in 1953. But since he does not know that he himself is Lingens, there will be centered worlds compatible with his beliefs that are centered on a different person than Lingens, perhaps one that was born at a different time. The possible worlds compatible with the knowledge of the person who knows that the meeting starts at noon, but not that it is starting now will all be worlds in which the meeting starts at noon, but the center component of some of his epistemic alternatives will be another time: in those possible scenarios, the time at which he is in the given belief state is some time before the noon meeting.

The Lewisian model of a state of belief is a straightforward generalization of the classical model, simply replacing the possible worlds with more fine-grained objects – centered possible worlds. They are more fine-grained in that a single possible world will correspond to multiple centered worlds. And as Lewis emphasized, his account is a generalization in the sense that contents of belief in the classical model (sets of uncentered worlds, representing timeless, impersonal information) will be a special case of contents of belief in his more fine-grained models: each timeless, impersonal proposition will determine a unique centered-worlds proposition, one in which the centers are irrelevant. Suppose X is a set of centered worlds meeting this condition: for all w and

---

4 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 continuant 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.
c, and \( c' <c,w> \in X \) if and only if \( <c',w> \in X \). Then \( X \) is a set of centered worlds that represents exactly the information represented by a simple set of possible worlds: the \( w \) such that for some (or all) \( c, <c,w> \in X \). So sets of centered worlds can represent objective information, as well as self-locating information, and it can represent the logical relations between objective and self-locating contents.

The Lewisian model is like the classical model in that belief states are abstracted from the believer who is in the state. It is the agent’s cognitive situation that determines what belief state he is in, but the state itself is represented by a set of possible worlds (in the classical model) or a set of centered possible worlds (in Lewis’s generalization) in which the believer plays no special role. But I think this kind of representation misconstrues the real message of the phenomenon of self-locating information, which is that it is essential to an adequate representation of a state of knowledge or belief that the information that is the content of a state be linked to the knower or believer who is in the state, and to the situation in which she is in it. It is usually assumed that it is the more fine-grained distinctions between epistemic and doxastic alternatives that is doing the work in explaining the distinctive character of self-locating attitudes, but this is a mistake; the real role that the centers play is to represent the links between a believer’s actual situation (in the world in which he has the beliefs) and the possible worlds that are compatible with his beliefs. Using more fine-grained contents of belief does not help us to represent these links, and it is not necessary to use more fine-grained contents once we have added the structure to represent the links between a subject’s situation and the possible worlds that represent his cognitive state in that situation.

The problem with Lewis’s account of content is that it makes it difficult or impossible to compare beliefs across time, and across persons. One way to see the problem is to consider the representation of iterated knowledge and belief. Lewis does not formulate his theory explicitly in terms of doxastic and epistemic accessibility relations, but he might have, and if he had, it would provide a straightforward semantics for intrapersonal iteration: what any subject knows or believes about what he himself knows or believes. For any \( \phi \) whose value is a set of centered worlds, “I believe that \( \phi \)” will also have a value that is a set of centered worlds: those centered on a person who believes that \( \phi \) at the time of the center. But inter-personal iteration: what A knows about what B knows – is a different matter. In order to make sense of interpersonal iteration, and more generally to make sense of the communication of information, we need an account of the contents of attitudes that is impersonal, or at least interpersonal, and this Lewis’s account of content does not provide. Intuitively, it seems reasonable to say that if Daniels does not know the identity of Lingens, the famous Stanford amnesiac, then in a sense, he is ignorant of the same fact that Lingens is ignorant of. If he finds out, he can tell Lingens. When he says “you are Rudolf Lingens”, he is giving him precisely the self-locating information that Lingens had previously lacked. But Daniels’s knowledge of this fact is, of course, not self-locating knowledge. The property of being Lingens, or the set of centered worlds whose center is Lingens, cannot represent the information that Daniels acquires when he learns who Lingens is, and if this centered worlds proposition is not an adequate representation of what Daniels told Lingens, then it is equally not an adequate representation what Lingens learns While Lewis’s account succeeds in smoothly
integrating one person’s self-locating belief with that same person’s impersonal beliefs, it
fails to integrate the objective and self-locating attitudes of different agents, and that is
what we need in order to explain communication.

As we noted, the Hintikka theory gives a straightforward account of iterated knowledge
and belief, both intra- and interpersonal. Each knower has his or her own epistemic
accessibility relation. A claim such as “A knows that B doesn’t know that \( \phi \)” can be
formalized as \( K_A \neg K_B \phi \), and the semantics will say that this iterated knowledge attribution
is true in possible world \( x \) if and only if, for all worlds \( y \) such that \( xR_A y \), there is a world
\( z \) such that \( yR_B z \) and \( \phi \) is false in \( z \). (Where ‘\( R_A \)’ and ‘\( R_B \)’ are the epistemic accessibility
relations for knowers A and B, respectively) But this representation is perhaps too
straightforward, sweeping some problems under the rug. (Lewis’s account makes iterated
knowledge too hard, but Hintikka’s makes it too easy.) We can compare the contents of
belief and knowledge for different agents, but doing so is not always straightforward.
The modified centered-worlds account that I will propose aims for a middle ground that
allows for the interpersonal interaction of attitudes, but that also brings out the problems
that arise for this kind of interaction.

3. A Modified centered worlds account

The modified account I will sketch uses exactly the apparatus that Lewis introduced
(centered possible worlds) to represent cognitive states, but it will be using them in a
slightly different way. In particular, I will not appeal to the finer-grained distinctions
between possibilities that the notion of a centered world permits, or at least not directly.
Belief states will be modeled by sets of centered possible worlds, but the job of the center
will be to link the believer as he is in the world in which he has the beliefs to the person
he takes himself to be in the world as he takes it to be. It will be an assumption of the
model that the centered worlds that are epistemic or doxastic alternatives will have
different centers only if they are also different worlds. The assumption is that if you
don’t know where you are in the world, then (in all cases) you also do not know what
world you are in. It is this assumption that will allow us to take ordinary impersonal and
timeless propositions as the contents of belief, and so to allow for the comparison and
communication of the information of different subjects, while at the same time
accounting for self-location. But the assumption needs defense. Let me use an extended
example to motivate it:

It is Monday afternoon. After shopping in the mall, I take the elevator down to level B of
the parking garage. I had gone up a different elevator, one in the center of the garage. The
one I came down is either at the east or the west end, I am not sure which – there is an
elevator at each end. I know my car is about in the middle along the northern edge, but
is that to the right or to the left? I have a clear mental map of level B, but it has no “you
are here” marker, so I don’t know how to orient myself on it. The garage is pretty
symmetrical, so it is hard to tell by looking around just where I am. I do know that there

---

5 The framework I will sketch and apply here was introduced in Chapter 3 of Stalnaker
(2008), and an appendix to that chapter has a few formal details.
is a pale green Prius with Massachusetts license plate 374- BJ8 to my right as I come out of the elevator, but knowing that does not help, since of course my mental map of level B does not tell me what cars are parked in what places.

Clearly, it is self-locating knowledge that I lack: I don’t know where to place myself in an environment of which I have a pretty clear objective mental representation. But the knowledge I lack is nevertheless knowledge of what possible world I am in. Presumably, there is not, in the actual world, a pale green Prius with Massachusetts license plate 374-BJ8 in the symmetrical place at the other end of level B of the garage, and there is actually no person, who might, for all either he or I know, be me, looking at it. Assume that I am actually at the east end. Then there is a counterfactual possible world in which I am (at the present time) at the west end. In that counterfactual world, the pale green Prius with that license number is parked at the west end, to my right as I emerge from the elevator. My car is to the north, in this counterfactual world, as it is in the actual world, but this means it is to my left, rather than to my right as it is in the actual world. My ignorance of where I am in the parking garage, and which way I should go to find my car, is represented by the fact that counterfactual worlds like this are compatible with my knowledge.

Now one might tell a science fiction story in which two events of the kind I have described take place, one on Monday at the east end, and one on Tuesday, at about the same time at the west end. The person (perhaps me) who emerges from the elevator at the west end on Tuesday glances at the same pale green Prius with the same license plate. On Tuesday, that car is parked at the west end, in the corresponding spot. The Tuesday person’s experiences are, from the inside, indiscernible from mine. In fact all of that person’s memories and experiences at the time are indiscernible from mine. (Perhaps we have led parallel lives, or perhaps we are both amnesiacs, or perhaps he is me, but was given a drug that snipped out all memory between today and tomorrow, without affecting the rest.) In this story, one might be tempted to say that these two actual scenarios, the one taking place at the east end on Monday, and the other taking place at the west end on Tuesday, are each epistemic alternatives for both me and for my counterpart. Even if I became omniscient about what world I am in, one might think, I might remain ignorant of which of these two people I am (or if my counterpart is actually me, having forgotten the Monday event on Tuesday, I might remain ignorant of what time it is now.) One could say this, but one should not for several reasons. First, one does not need the science fiction story to make the point that some information is essentially indexical. The original, quite mundane story accomplishes this. (If the phenomenon of self-locating belief did depend on such science fiction scenarios, we could safely ignore it.) In the simple story, my belief is essentially indexical in the following sense: I cannot infer from a purely objective description of the world that I am at the east end. My objective description might tell me that there is a Prius at the east end, and no Prius at the west end, but I can use this information to orient myself only by putting it together with the information that there is a Prius here. More directly, of course, my objective description may tell me that Bob Stalnaker is at the east end, and not the west end, on Monday afternoon, but it is only because I know that I am Bob Stalnaker, and that it is Monday afternoon, that I can use this information. In normal cases, the indexical information we
need to locate ourselves in the world is obvious enough to go unnoticed, but it is always
essential. Stories about amnesiacs are not essential to the point; their job is just to make
more prominent the role of the kind of information that usually gets taken for granted.
Second point: even in the science fiction stories with two actual scenarios that are
indiscernible, one can still assume that different epistemic alternatives are scenarios in
different possible worlds. Even if, in this world, another event indiscernible from this one
will take place tomorrow, or did take place yesterday (I don’t know which) that event is
not this one. Since I don’t know whether it is now Monday or Tuesday, I don’t know
whether this token thought is taking place on Monday, or on Tuesday, but I do know that
which ever it is, the (token) thought that I am having on the other day (yesterday or
tomorrow) is a different one. So one can assume that distinct epistemically possible
scenarios are always different possible worlds without excluding the fanciful cases. But
the third point is that one must take doxastic and epistemic alternatives to be different
possible worlds, even in the fanciful cases, if one is to give a proper account of the role of
belief and knowledge in action. Suppose I know that I will be in a similar situation –
perhaps an absolutely indiscernible situation – on both Monday and Tuesday without
knowing, on either day, which day it is. (I will be given an amnesia-inducing drug that
ensures that, on Tuesday, I have no memory of the Monday situation.)

On each day, I must make a decision, perhaps to go left or right to find my car, or to accept or reject a
bet about the result of a coin flip. In deciding what to do, I am making it true that this is
what I do in all of the possible situations that are epistemically possible for me. Deciding
is (at least normally) a way of coming to know. But it would distort the deliberative
situation to think that, on Tuesday (or Monday) I was deciding what to do on the other
day. I might be giving myself evidence about what I will or did do on the other day (if I
have reason to think that my situation will be similar enough), but that is different from
making a choice that decides it.

So we follow Lewis in using sets of centered possible worlds to represent states of belief,
and we can use a doxastic accessibility relation on centered worlds to represent the
beliefs of various believers at various times in a range of different possible worlds. If

\(<c,x> \text{ and } <c',y>\) are centered worlds, then \(<c,x>R<\cdot',y>\) holds just in case it is
compatible with the beliefs of the individual at the center c at the time of that center in
world x that she is the person at the center \(c'\), that the world is world y, and that the time
is the time of \(c'\). The assumption that I have been trying to motivate – that different
doxastic or epistemic alternatives (for a given believer at a given time) have different
centers only if the possible world component is also different, is formally expressed as
follows:

\[
\text{if } <c,x>R<\cdot',y> \text{ and } <c,x>R<\cdot'',y>, \text{ then } c' = c''.
\]

What this assumption does is to allow us to take the contents of belief to be sets of
ordinary, centerless possible worlds, which allows us to compare the beliefs of different

---

6 I am alluding here to the notorious Sleeping Beauty problem that Adam Elga introduced
to the philosophical community in 2000, and that has been extensively discussed in the
literature since then. I discuss this problem, applying the modified centered worlds
framework to it, in Stalnaker (2008), chapter 3.
subjects (to say what they agree and disagree about) and to represent what one subject believes about what another believes. Essentially self-locating beliefs will have, as their contents, ordinary impersonal propositions; their distinctive self-locating character will be a feature of the subject’s relation to that content, and not a feature of the content itself. The belief that Daniels expresses when, after discovering who Lingens is, he tells Lingens, “you are Rudolf Lingens” is not a self-locating belief (for Daniels), but it has exactly the same content as the newly acquired belief that O’Leary expresses when he echoes Daniels, “so, at last I know: I am Rudolf Lingens.” But for Lingens, the belief with that content is a self-locating belief.

As we have seen, in the classical Hintikka models of knowledge and belief, the epistemic and doxastic accessibility relations are indexed to the subject. In the centered-worlds generalization, we have just one epistemic or doxastic accessibility relation: the subjects are determined by the relata, rather than the relation. Instead of saying that y is compatible with what A knows in x if and only if xR_A y, we say that y is compatible with A’s knowledge if and only if for some C, <A,x>R<C,y>. Two subjects A and B have conflicting beliefs if the set of worlds compatible with A’s beliefs is disjoint from the set compatible with B’s. Agreement and disagreement is straightforward (whether the beliefs are self-locating or not), and the theory also provides the resources to represent iterated knowledge and belief, but here things are not quite so straightforward. Let me sketch another simple example to bring out two complications.

I am talking with John Perry at an APA meeting, but he is not wearing his nametag, and I am not sure who he is. I know Perry’s work, but (let’s suppose) I had never before met him. I am pretty sure the guy I am talking with is either John Perry or Fred Dretske, but I am not sure which. He is telling me what a fantastic book Knowledge and the Flow of Information is, and I am wondering whether he is bragging or praising the work of a colleague. I believe that the person with whom I am talking thinks that Knowledge and the Flow of Information is an excellent book, and I also of course believe that he believes that he is telling this to me (though he may not know who I am, since I am not wearing my nametag either). To represent these iterated belief (My beliefs about what John (or Fred) believes), we need to locate both the primary believer (me, in this case) and the person whose beliefs I have beliefs about (the person I am talking with) in the possible worlds that are compatible, according to what I believe, with what he believes.

The first and more obvious point here is that knowledge and belief are intensional: My knowledge and belief about what the person I am talking with believes are not the same as my knowledge and belief about what John believes, even though John is the person I am talking with. The classical Hintikka models, with their indexed accessibility relations ignore this complication. But the centered world models, with the identity of the believer in the relata, can make the required distinctions. Suppose ‘f’ is an individual concept, picking out a person as a function of a possible world. Then the pair <f(x), x> is a centered world, the one with the value of f for world x at the center. If f is a function

---

7 To simplify and avoid clutter, I am going to ignore, from now on, the time at the center, assuming a fixed time, and taking centers to be just individual subjects.
whose value is the person I am talking with in each of the possible worlds compatible with my beliefs, then one can generalize about my beliefs about the beliefs of the person I am talking with (in one sense), and distinguish them from my beliefs about the beliefs of John, who is that person.

The second point is that to calibrate the beliefs of different subjects, we need to locate them, not only in worlds compatible with their own beliefs, but also in worlds compatible with the beliefs of those whose beliefs they have beliefs about. To represent John’s beliefs about what I believe about him, he needs to locate himself, not only in the worlds that are the way he takes things to be, but also in the worlds that are the way he thinks that I take them to be. Suppose he comes to realize that I am not sure whether he is Perry or Dretske. Then his “I” will pick out Dretske in some of the worlds that he takes to be compatible with my beliefs. (“This guy thinks I might be Fred Dretske”, he thinks to himself. In this case, his “I” tracks my concept “the person I am talking with”) The “I” in this belief attribution has no special status for the proposition it is used to pick out, nor does it have a special status for the subject of the belief being attributed (there is nothing self-locating about my beliefs about John.) But the iterated belief is self-locating for the person (John) who is attributing the belief to me. So there will be two different individuals that have a special status with respect to a possible world that is used to represent what one person believes about another’s beliefs.

Say that an individual concept, f, is an I-concept, with respect to a possible world x if and only if for all worlds y and subjects B, if <f(x),x>R<B,y>, then B = f(y). An I-concept f is an individual concept that picks out the individual that f(x) takes himself to be in each of the possible worlds that are compatible with what he believes in world x. Any two I-concepts (relative to world x) that pick out the same individual in world x will agree with respect to all worlds that are compatible with what that individual believes in world x, but they may differ with respect to possible worlds that are not compatible with that individual’s belief state in x. So I-concepts represent possible extensions of a subject’s self-location to epistemically inaccessible possible worlds: where an individual locates herself in worlds that, from her perspective, are counterfactual (including those she takes to be compatible with the beliefs of other subjects). Normally, when we can assume that everyone knows who everyone else is, we take the relevant I-concept to be the one that always picks out the individual herself. But when a person is attributing beliefs to someone who is confused about or ignorant of that person’s identity, more than one I-concept may be used, even within the same context. After our conversation, John might say to someone else, “I was talking with Bob Stalnaker, but he didn’t realize that it was me that he was talking with. He thought I might be Fred Dretske.” There is a shift here: consider the world that John correctly takes to be compatible with my beliefs in which the person I was talking with is Fred Dretske. The “me” in John’s remark picks out Perry in that world, while the “I” picks out Dretske.

Finally, let’s look back, from the perspective of this theory, at the idea of common ground: an infinitely iterated attitude with the structure of common knowledge that is our representation of a context in which a discourse takes place. We have seen, first, that iterated attitude must be defined in terms of a way of identifying the individual whose
attitudes one has attitudes about. So to define something like the common knowledge of a group of subjects, we need to specify, not just the subjects, but the ways they identify each other. In the context of a face-to-face conversation, this will be straightforward. Even if it is a bunch of amnesiacs discussing together who each of them might be, they will have a shared way of identifying each other – a basis for fixing the referents of the “I”s and “you”s in their conversation. And we have seen, second, that in the iterated case, we get, in a sense, multiple centering: in worlds compatible with A’s beliefs about what B believes, we need to locate both A and B (One center to represent who A takes herself to be in the world as she thinks B takes it to be, and another to represent who she thinks B takes himself to be in that same world.) So in a representation of the common ground – the information shared in common between a group of n individuals – there will be n individuals at the center: the individuals that they all presuppose themselves to be.

Here is a quick sketch of the apparatus with which common knowledge is represented:

For any individual concept f, we can define a relativized epistemic accessibility relation between (uncentered) possible worlds in terms of the epistemic accessibility relation between centered worlds as follows:

\[
\text{For any worlds } x \text{ and } y, \ xR_f y \iff f(x), x \text{R}(f(y), y). 
\]

Provided that f is an I-concept, relative to world x, the set \{y: xR_f y\} will be the set of worlds compatible with what f(x) knows in world x.

Now for any two individual concepts, f and g, we can define a binary relation \(R_{fg}\) as the transitive closure of \(R_f\) and \(R_g\), and this relation will determine a common knowledge set, relative to a world x, provided that both f and g are I-concepts, relative to all possible worlds that are \(R_{fg}\) related to x.

That is, the set \{y: xR_{fg} y\} will be the set of worlds compatible with the common knowledge, in world x, of f(x) and g(x) (relative to those ways of identifying each other) if and only if for all y in this set, and for all z and B, if \(f(y), y \text{R}(B, z)\), then B = f(z) and if \(g(y), y \text{R}(B, z)\), then B = g(z).

Intuitively, the idea is that two subjects have common knowledge only relative to a certain pair of ways they have of identifying each other. The same two subjects might have different ways of identifying each other that give rise to different states of common knowledge.

I will conclude by looking at another notorious example that I think this apparatus helps to clarify: the most Byzantine of the many Frege cases that have been discussed in the literature, a case that involves multiple identity confusion, indexicality, and contextual variation: Mark Richard’s example of the phone booth.

---

8 I will characterize common knowledge for the two-person case, but it generalizes to n persons in the obvious way.
4. Mark Richard’s phone booth story

A woman in a phone booth is talking to a man. She is also watching the man, who is waving at her, but she does not realize that it is the same man. The man also does not realize that the woman he is talking to is the same woman as the one he is waving at. The woman tells the man about the man waving at her. Then she says, “the man waving at me thinks I am in danger. But you don’t think I am in danger, do you?” The man replies, “no, I don’t think you are in danger.”

Both the man and the woman are sincere, and it seems that what each says is true. That is, the woman’s statements

“The man waving at me thinks I am in danger”

and

“You don’t think I am in danger”

are both true.

But the singular terms, “the man waving at me” and “you”, both refer to the same person, and the terms occur outside the scope of the attitude verb. So how can both statements be true?

We can model the essentials of the situation with three possible worlds: World $\alpha$ is the actual world in which there are two subjects A (the woman) and B (the man). They are talking to each other on the phone, and B is waving at A from across the street. A is in danger. World $\beta$ is the world as the woman A takes it to be. There is, in $\beta$, a woman, A, who is not in danger, and two different men, B1 and B2. She is talking to B1, and B2 is waving at her. A (in $\alpha$) centers herself at A (in $\beta$, the world as she takes it to be). That is, $\langle A,\alpha \rangle R \langle A,\beta \rangle$. World $\gamma$ is the world as B takes it to be. There are two relevant women in this world, A1 and A2, and two men, B1 and B2. B (in world $\alpha$) centers himself (in $\gamma$) at B1. (That is, $\langle B,\alpha \rangle R \langle B1,\gamma \rangle$.) B1 is talking to A1 in world $\gamma$, and is waving at A2. A1 is not in danger, but A2 is in danger.

Now the woman is aware (in the actual world $\alpha$) of what the man she is talking to believes, and so she takes world $\gamma$ to be compatible with what the man she is talking to believes. That is, $\langle B1,\beta \rangle R \langle B1,\gamma \rangle$. She is also aware of what the man who is waving at her believes: that he is waving at a woman who is in danger. The man (B1) is waving at a woman in danger in world $\gamma$, and so that world is compatible with what she believes the man waving at her believes.

We might distinguish two different I-concepts for A (relative to world $\alpha$): both take world $\beta$ to A (that is what makes them I-concepts for A in $\alpha$), but one takes world $\gamma$ to A1 and the other takes world $\gamma$ to A2. The first I-concept is relevant to the context of the phone conversation with the man, where her ‘you’ picks out B1 (in $\beta$), while the second

---

9 The story was first presented in Richard (1983).
is relevant to her attribution of belief to the man who is waving at her, who is B2, in world \( \beta \), since in the context of this attribution, she identifies herself, in \( \gamma \), with the woman who is in danger. Relative to the first I-concept, the clause “that I am in danger” refers to a proposition that is true in \( \gamma \), while relative to the second I-concept, it refers to a proposition that is false in world \( \gamma \). The man believes the first proposition, but not the second, and that is why “The man waving at me believes am in danger” and “you don’t believe I am in danger” said by the woman, are both true.

It has been clear all along that the compatibility of the woman’s two statements needs to be explained in terms of some kind of context shift. That was Richard’s point. The framework I have sketched by gives a precise account of the way that the referent of “I” in the scope of a second or third person belief attribution will vary with context, and so provides one way of pinning down just what kind of context shift is involved.


