A Formal Account of Self-Knowledge and Action

Yves Lesperance
Dept. of Computer Science
University of Toronto
Toronto, Canada M5S 1A4

Abstract

In this paper, we propose a formal theory of knowledge and action that accommodates indexical knowledge and reflects the dependence of action upon it. The model-theoretic semantics of belief given characterizes an agent's belief state not only in terms of the ways the world might be if his beliefs are true, but also in terms of how the world must look from his perspective. We then propose a specification of the conditions under which an agent is able to achieve a goal by doing an action that does not require the agent to know in an absolute sense who he is or what time it is. Finally, we show through an example how actions can be characterized in an agent-relative way within the theory, so as to avoid requiring the agent to know which objects are acted upon, and how such a characterization can be used to prove that an agent is able to achieve a goal by doing an action if he knows certain facts.

1 Introduction

Doing all but the most basic actions requires knowledge. For example, in order to call somebody, one needs to know his phone number. Doing an action may have the effect that one acquires some significant new knowledge. The most penetrating theory of the relationship between knowledge and action that addresses artificial intelligence (AI) concerns is that of Moore [1985]. His framework is essentially an encoding in ordinary first-order logic of the possible-world semantics of the combination of first-order dynamic logic with an S4 modal logic of knowledge. The central result of Moore's work is a definition of the conditions under which *an agent is able to achieve a goal by doing an action* (CAN), where ability is understood as requiring the agent to know enough to achieve the goal, rather than it being just physically possible. The definition makes use of the fact that actions are. represented by a type of term and that, since we are in a modal framework, we can distinguish between the action description or concept corresponding to an action term and the sequence of primitive actions denoted by the term in a particular possible world. In particular, it makes sense to talk about an agent knowing what an action description is or an agent having *de re* knowledge of an action.[1] This notion is important because an agent may well know that an action described in a given way achieves his goal without being able to do it because he does not know what primitive actions the description denotes.

The possible-world semantics' account of *de re* knowledge is that an agent *a* knows of *x* that it is □ in world *w* iff in all worlds compatible with what *a* knows in w_y *x* is . Following Hintikka [1962], knowing who/what *6* is is taken to be equivalent to knowing of some *x* that it is *9.* This does not however say much about what is actually required for an agent to have *de re* knowledge; the question is recast into that of what is required for the same individual to exist in distinct epistemically possible worlds. Actually, Moore uses a substitutional interpretation of quantifiers, but since rigid designators are substituted for variables, this has little effect on the issue. There has been much philosophical debate on this question but the common answer in AI circles has been that knowing who someone is requires having a *standard name* for that person and similarly for knowing-what [Konolidge, 1986, Levesque, 1984].[2] Moore [1985J is not very clear as to whether he actually holds this view, but his assertion that "in describing standard identifiers we assumed that everyone knew what they referred to" seems to indicate that this is the case. The account works well for some kinds of objects (e.g. phones) and some tasks (e.g. question answering) but requires some serious stretching in other contexts; for example, what are standard names for a mobile vision-equipped robot?

Moore's definition of CAN is recursive and goes as follows: an agent is able to achieve a goal by doing an action (e.g. calling a person) iff either the agent knows what the action is (e.g. knows what dialing the person's number amounts to) and knows that doing the action would result in the goal being satisfied, or the agent knows what an initial action is (e.g. knows what looking up the person's number in the phone book amounts to) and knows

Moore states that "knowing what action is referred to by an action description means having a rigid designator for the action described ... a rigid designator for an action must be an *executable description* of the action".

[2]1 think that formal study of the relationship between knowledge and action holds potential for improving our understanding of this and other epistemological questions.

that his doing this initial action results in him being able to achieve the goal by doing some subsequent action (e.g. dialing the person's number). Note that the agent is required to know who he is and what the initial action is. The recursive case of the definition captures the fact that the agent need not initially know all the actions that make up a successful plan as long as he knows that he will know what to do next at each step of his plan. The only way to talk about time is through the dynamic logic operators. Statements that are not in the scope of a dynamic logic operator may be thought to be about the current time.[3]

One problem with Moore's definition of ability is that it ignores the fact that much of the knowledge required for action is indexical or *de se* rather than simply *de re.*[4] For example, if I am at a party, I can go and talk to the person standing by the punch bowl even if I don't know who that person is; nor do I need to know her latitude and longitude or any other objective specification of where she is; all I need to know is roughly where that person is *relative to me.* If you are tempted to think that one knows who a person is as soon as one knows what she looks like, imagine it's a costume party. Similarly, one can go for lunch now without knowing what time it is. Moreover, one can do both of these actions without knowing who one is — for instance if one is amnesiac. The fact that perception also yields mostly indexical knowledge is an additional reason for dealing with it. It has also been argued that indexical representations carry significant efficiency advantages [Agre and Chapman, 1987].

It is arguable that all aspects of the indexicality of belief may be reduced to two primitives: 'he himself or T, which stands for the agent of the belief, and 'the current time' or 'now', which stands for the time of the belief (this seems to be the view of Lewis [1979]). 'Here' can be taken to stand for 'the location of he himself now'.[5] Demonstratives, such as 'this person', may be viewed as abbreviations for descriptions involving 'he himself and 'now', such as 'the person at such and such relative location from he himself now'. This is the approach we adopt in developing our account.

While it is difficult to deny that one may ignore what time it is or where one is, one may be more skeptical about the possibility of ignoring who one is. It may be argued that one needs not have any objective way of referring to oneself in order to know who one is. But the subjective sense of knowing-who in which this may be the case seems quite distinct from that appealed to in the party example and commonly explained in terms of knowing a standard name. Imagine an agent whose *only* knowledge was a list associating each person, represented by a standard name, with his salary. Would such an agent know how much he himself was making? It does not seem so. Or putting it another way, if one were to claim against common intuition that 'oneself is a standard name, would it be legitimate to expect that all other coreferential standard names would be known to be coreferential? In developing our theory, we will adopt an interpretation of knowing-who and *de re* belief that requires the object to be known under an appropriate *objective* description.[6]

There has been disagreement among philosophers on the adequacy of standard possible-world semantics to handle *de se* attitudes. Our analysis of these issues has lead us to conclude that while standard possible-world semantics is able to distinguish between *de se* and non-*de se* belief (as argued by Stalnaker [I98I]), it fails to characterize the purely internal aspects of a belief state — what two agents have in common when the world appears the same when viewed from their distinct perspective. Since it is these internal aspects of belief states that play a causal role in determining action, a semantics which provides such a characterization would definitely be superior.[7]

In the next section, we propose a logic that handles *de se* knowledge and belief together with a semantics that characterizes the action-determining internal aspects of beliefstat.es. We take the belief accessibility relation *B* to hold over belief indices, which are world-agent-time triples. $\langle\langle w, a, t\rangle, \langle w', t', a'\rangle\rangle \in B$ if world *w'* is compatible with what agent *a* believes at time *t* in world *w* if he assumes that he is a' and the current time is *t'.* Thus, the belief state of an agent at a time in a world is modeled by a set of belief indices, each of which contains an agent that he thinks he might be, a time that he thinks might be current, and a way the world might be if he is that agent at that time. A belief index characterizes both a world and the perspective from which it is viewed. This treatment of *de se* belief was inspired by informal proposals by Perry [I979] and especially Lewis [1979].[8]

The logic includes temporal operators that allow one to assert that a state of affairs holds at a given time instant (this was implicit in the discussion of belief) and that an agent does an action from one instant to another (adapted from [Reseller and Urquhart, 1971]). But any characterization of the ability of agents to achieve goals by doing actions requires consideration of possible

[3]Morgenstern [1987] extends Moore's account to more general actions and plans involving several agents. She gives a consistent account of belief as a syntactic predicate, which results in a more complex but more expressive language. Self-referential beliefs are expressible, which gives rises to paradoxical statements such as "This sentence is known to be false". But these extensions seem largely orthogonal to the issues that concern us here.

Failure to appreciate this has also lead to the neglect of indexical knowledge in AI theories of knowledge and belief.

' It is possible to take 'here' as a primitive instead and define 'he himself as 'the agent located here now'. We prefer the first alternative because it seems more natural to take beliefs to be properties of agent-states rather than locations where an agent is and times.

Our appeal to the standard name account of knowing-who in the above argument should not be taken to imply unreserved acceptance of the account or necessary dependence of our theory upon it.

[7] See [Lesperance, 1989] for a more detailed discussion of these issues.

[8]The indexicality of belief discussed here is a very distinct phenomenon from the self-referentiality allowed by Morgenstern [1987]. We are not aware of any account that would reduce the former to the latter.

courses of events that differ from the actual course of events in the way the future unfolds.[9] In the dynamic logic framework adopted by Moore, the modal character in this kind of consideration is combined with the temporal aspect, as is made explicit by the 'branching time' semantics of the framework. This limits the expressive power of the formalism: one cannot really say that an object has a property at a given time, only that the object has the property in the state that results from doing some sequence of actions; there is no way to say that two actions would be completed at the same time or to say that something *actually* occurred. Our semantic account of *de se* belief requires a distinction between time and possible worlds and this distinction must be preserved when one allows consideration of different possible courses of actions. Due to this, in addition to the temporal operators, modal operators for 'historical' necessity and possibility are included (this aspect of the formalism is adapted from [Thomason, 1984]).[10] Using these operators one can state that an action is possible for an agent at a time, in the sense that all its physical prerequisites are satisfied,[11] or that some effects necessarily hold at the conclusion of an action.

We describe our framework more precisely in the next section. A revised definition of CAN is then proposed. This definition requires the agent to have *de se* knowledge instead of knowing who he is. Finally, we illustrate through an example how actions should be formalized in the framework to avoid making unrealistic demands upon the agent's knowledge and how the theory can be used to prove that an agent can achieve a goal by doing an action if he knows certain facts.

2 Elements of a formal theory

The following is an abbreviated description of the theory of self-knowledge and action. Further details can be found in [Lesperance, 1989].

2.1 Syntax

Our language is an extension of a first-order language with equality. There are four sorts of terms: individual terms, agent terms, which are a subclass of individual terms, action terms, which denote simple actions, and temporal terms, which denote instants. Compound terms of the first three sorts may be formed; the arguments must be individual terms, self, now, and then are distinguished indexical terms, which denote the current agent, the current time, and a contextually determined future time respectively. Variables are written in lower case, while function symbols and predicate constants are written in upper case.

One asserts that a simple action ac is done by an agent ag with the atomic formula $ag \odot ac$ ($\odot$ is a distinguished predicate constant). Such an event is assumed to occur from now to then (temporal operators that shift these contextual parameters are introduced below). Other predicates are taken to denote static relations between their arguments, which are assumed to hold now. We assume a linear ordering on times and the usual relational operators can be used.

Temporal operators can be used to shift the context against which a formula is interpreted, i.e. the times referred to by now and then: $\phi@t$ means that ϕ holds at or occurs from t, now being coreferential with t in the scope of the @ operator; $\phi@@t$ means that ϕ occurs up to time t, then being coreferential with t in the scope of the @@ operator. We take the temporal operators to have higher precedence than -. Complex actions are represented using indexical formulas (action formulas) where self, now, and then stand for the agent, starting time, and completion time parameters respectively. These parameters can then be bound by shifting the context. All the dynamic logic modes of composition except repetition (sequential composition, nondeterministic choice, and test actions) can be encoded using the constructs already in the logic. Shorthand forms are provided for readability. Concurrent actions and plans involving multiple agents are expressible in the logic.

$\Box\phi$ means that ϕ is necessary given everything that has already happened; possibility ($\Diamond$) is defined in the usual way. $BEL(ag, \phi)$ means that ag believes ϕ; if ϕ contains an occurrence of self that is not in the scope of a nested BEL operator or an occurrence of now not in the scope of a BEL or @ operator, the formula is taken to attribute a *de se* belief to the agent. We define $KNOW(ag, \phi)$ to stand for $BEL(\mathbf{ag}, \phi) \wedge \exists \mathbf{ag}'(\mathbf{ag}' = \mathbf{ag} \wedge \phi')$, where ag' is a variable variable that does not occur free in ϕ and ϕ' is the result of substituting ag' for all occurrences of self in ϕ that are not in the scope of a BEL operator.[12] $CAN(ag, \alpha, \beta)$ means that ag is able to achieve the goal β by doing action α, where *a* is an action formula and β is a goal formula. A goal formula is an indexical formula where the now contextual parameter stands for the time at which the goal is achieved.

2.2 Semantics

A semantic structure *M* is a tuple $\langle Agents, Objects, Times, Actions, W, \prec, B, \approx, \Phi, H \rangle$. The first four components arc (non-empty) domains for the appropriate sorts; the domain *Individuals* is $Agents \cup Objects$. *W* is a set of temporally extended possible worlds. $\prec$ is a strict total order on *Times*. $B \subseteq (W \times Agents \times Times)^2$ is the belief accessibility relation. The rationale behind this formulation was explained in the introduction. *B* must be euclidean, transitive and serial; this corresponds to the modal system known as weak-S5 or KD45.

$\approx$ is a family of accessibility relations — one for each time instant — that is used to interpret the 'historical' necessity operator $\Box$. Intuitively, $w \approx_t w'$ if *w* and

No metaphysical commitment is implied here; there should be some naturalistic explanation of such 'ability talk'.

This may not be a good name for the brand of necessity under consideration; Thomason seems to take it in a narrower sense than ours — he seems to take only what has already happened to be necessary; perhaps 'causal-historical' necessity is better.

[11] This should not be confused with ability, which requires the agent to know of the action that it is possible and has the goal as an effect.

[12] This treatment of knowledge is adopted for its simplicity in spite of its well known inadequacies.

w' differ only in what happens after t. $\approx$ must satisfy the following constraints: firstly, for all $t \in$ *Times*, $\approx_t$ must be an equivalence relation — this implies that at any given instant, $\Box$ and $\Diamond$ obey the principles of the modal system S5; and secondly, for all $w, w' \in W$ and $t, t' \in$ *Times*, if $w \approx_t w$ and $t' \preceq t$, then $w \approx_{t'} w'$, i.e. possibilities do not increase as time passes.

The denotation of terms and the satisfaction of formulas are defined relative to indices. An evaluation index is a 4-tuple $\in W \times$ *Agents* $\times$ *Times*, where the agent component is the denotation of the indexical self, the first time component is the denotation of now, and the second time component is the denotation of then.

Φ gives the extension of predicate constants and function symbols at an index. We require that Φ satisfy the following constraints: firstly, for any predicate or function ϵ, $\Phi(\epsilon, w, a, t_1, t_2) = \Phi(\epsilon, w, a', t_1, t_2')$, the denotation assigned by Φ to any symbol is constant with respect to the agent and second time components of indices; and secondly, for any predicate or function ϵ, $w, w' \in W, a \in$ *Agents*, $t_1, t_1', t_2 \in$ *Times*, if $w \approx_{t_1} w'$ and $t_1' \preceq t_1$, then $\Phi(\epsilon, w, a, t_1', t_2) = \Phi(\epsilon, w', a, t_1', t_2)$, i.e. Φ must ensure that historical alternatives up to t_1 differ only in what happens after t_1.

$H \subseteq$ *Actions* $\times$ *Indices* determines which actions are done at which indices; $\langle c, w, a, t_1, t_2\rangle \in H$ iff action c is done by agent a from times t_1 to t_2 in world w. We require that starting times always be prior to or equal to ending times, i.e. for all $\langle c, w, a, t_1, t_2\rangle \in H$, $t_1 \preceq t_2$. Finally, for all $c \in$ *Actions*, $w, w' \in W, a \in$ *Agents*, $t_1, t_2, t_2' \in$ *Times*, if $w \approx_{t_2} w'$ and $t_2' \preceq t_2$, then $H(c, w, a, t_1, t_2')$ iff $H(c, w', a, t_1, t_2')$, again to ensure that historical alternatives up to t_2 differ only in what happens after t_2.

An assignment is a function that maps variables into elements of the domain appropriate to them. $g[v/c]$ is the assignment that is identical to g except that it maps variable v into the entity c.

The *denotation* of a term θ in a structure M at an index $\iota = \langle w, a, t_1, t_2\rangle$ under an assignment g, written $[\![\theta]\!]^M_{\iota,g}$ is defined in the standard way for variables and recursively for compound terms; for the indexicals, we have $[\![\mathsf{self}]\!]^M_{\iota,g} = a$, $[\![\mathsf{now}]\!]^M_{\iota,g} = t_1$, and $[\![\mathsf{then}]\!]^M_{\iota,g} = t_2$.

We can now define what it means for a formula ϕ to be *satisfied* by a structure M, an index $\iota = \langle w, a, t_1, t_2\rangle$, and an assignment g, which we write $M, i, g \models \phi$. For conciseness, we omit the standard part of the definition that deals with constructs of first-order logic with equality; for the rest of the language, we have:

- $M, \iota, g \models \mathsf{ag} \odot \mathsf{ac}$ iff $H([\![\mathsf{ac}]\!]^M_{\iota,g}, w, [\![\mathsf{ag}]\!]^M_{\iota,g}, t_1, t_2)$
- $M, \iota, g \models \mathsf{t}_1 < \mathsf{t}_2$ iff $[\![\mathsf{t}_1]\!]^M_{\iota,g} \prec [\![\mathsf{t}_2]\!]^M_{\iota,g}$
- $M, \iota, g \models \phi@\mathsf{t}$ iff $M, \langle w, a, [\![\mathsf{t}]\!]^M_{\iota,g}, t_2\rangle, g \models \phi$
- $M, \iota, g \models \phi@@\mathsf{t}$ iff $M, \langle w, a, t_1, [\![\mathsf{t}]\!]^M_{\iota,g}\rangle, g \models \phi$
- $M, \iota, g \models \Box\phi$ iff for all $w' \in W$ such that $w \approx_{t_1} w'$, $M, \langle w', a, t_1, t_2\rangle, g \models \phi$
- $M, \iota, g \models \mathsf{BEL}(\mathsf{ag}, \phi)$ iff
  for all $a' \in$ *Agents*, $t' \in$ *Times*, $w' \in W$,
  such that $\langle\langle w, [\![\mathsf{ag}]\!]^M_{\iota,g}, t_1\rangle, \langle w', a', t'\rangle\rangle \in B$,
  $M, \langle w', a', t', t_2\rangle, g \models \phi$

Note that the definition of BEL makes then epistemically rigid.[13]

A formula ϕ is *satisfiable* iff there is at least one structure M, index ι, and assignment g, such that $M, \iota, g \models \phi$. A formula ϕ is *valid* (written $\models \phi$) iff it is satisfied by all structures, indices, and assignments.

2.3 A definition of CAN

Our definition of satisfaction for the CAN operator goes as follows:

$M, \iota, g \models \mathsf{CAN}(\mathsf{ag}, \alpha, \gamma)$ iff
$M, \iota, g \models \exists \mathsf{ac}\, \mathsf{KNOW}(\mathsf{ag}, \forall \mathsf{t} \Box((\mathsf{self} \odot \mathsf{ac})@@\mathsf{t} \supset \alpha@@\mathsf{t})$
$\wedge \exists \mathsf{t} \Diamond (\mathsf{self} \odot \mathsf{ac})@@\mathsf{t} \wedge$
$\forall \mathsf{t} \Box((\mathsf{self} \odot \mathsf{ac})@@\mathsf{t} \supset \gamma@\mathsf{t}))$
or, there exists an action formula α' such that
$M, \iota, g \models \exists \mathsf{ac}\, \mathsf{KNOW}(\mathsf{ag},$
$\forall \mathsf{t} \forall \mathsf{t_i} \Box((\mathsf{self} \odot \mathsf{ac})@@\mathsf{t_i} \wedge \alpha'@\mathsf{t_i}@@\mathsf{t} \supset \alpha@@\mathsf{t}) \wedge$
$\exists \mathsf{t_i} \Diamond (\mathsf{self} \odot \mathsf{ac})@@\mathsf{t_i} \wedge$
$\forall \mathsf{t_i} \Box((\mathsf{self} \odot \mathsf{ac})@@\mathsf{t_i} \supset \mathsf{CAN}(\mathsf{self}, \alpha', \gamma)@\mathsf{t_i}))$

This definition is recursive, and similar in structure to that of Moore [1985]. The first disjunct handles the base case: an agent ag can achieve a goal γ by doing an action α in a structure *M,* at an index ι, under an assignment g, if he knows of some simple action ac that whenever he himself does ac starting now,[14] he also necessarily does α starting now (1st. conjunct), that it is possible for him to do ac starting now (2nd. conjunct), and that whenever he himself does ac starting now, the goal γ is necessarily achieved at the completion time of ac (3rd. conjunct).

The second disjunct handles the recursive case: ag can achieve γ by doing action α at *M,L,* and *g* if he knows of some simple action ac that whenever he himself does ac from now to some intermediate time $\mathsf{t_i}$ followed by some other actions described by α' from $\mathsf{t_i}$ to t, he also necessarily does α from now to t (1st. conjunct), that it is possible for him to do ac starting now (2nd. conjunct,), and that whenever he himself does ac starting now, it is necessary that, he can achieve 7 by doing α' at, the completion time of ac (3rd. conjunct,).[15] Note that only action formulas that encode dynamic logic forms are presently handled by the definition.

The definition differs from that of Moore primarily in requiring that, the agent have *dc se* knowledge of the

'This means that in spite of our intuitions, $\exists \mathsf{t}\mathsf{BEL}(\mathsf{AG}, \mathsf{t} = \mathsf{then})$ is valid. This will be rectified in future versions of the theory. Usually, it does not cause problems because we quantify existentially over the completion time of actions.

[14]Remember that formulas that do not mention time refer implicitly to the current time.

[15]This formvlation, like that of Moore, may be criticized on the following two grounds. Firstly, it fails to characterize the knowledge that the agent must have about what to do after the initial action in terms of his knowing what sequences of simple actions lead to states where the goal is achieved; this knowledge is only characterized in terms of a sentence (α') that describes these action sequences. One shouldn't have to appeal to such syntactic objects in characterizing ability scmantically. Secondly, there is nothing that requires the definition to bottom out; an agent that indefinitely does an action known to be always possible is judged to he able to achieve any goal by repetitively doing the action.

agent (self) and starting time (now) of the actions involved. Other discrepancies arise from differences in our treatment of time, necessity, and complex actions. The definition still requires the agent to know what the next action is and what its arguments are if it has any. In section 3, we show that if actions are correctly specified, this does not require agents to know more than they actually need to.

2.4 Some properties of the theory

We have an axiomatization of the theory that is sufficient for proving most results of practical interest, such as the example given in the next section. We plan to investigate completeness issues. Due to space limitations, we will only discuss a few properties of the part of the theory that deals with belief.

Since we take knowing-who to mean having a standard name, and standard names are- context independent, it seems reasonable to have an agent know that he is himself without knowing who he is. He may even believe that he is someone else (e.g. a lunatic who thinks he is Napoleon). The following proposition shows that the theory respects these intuitions.

Proposition
$\models \forall ag\ \mathrm{BEL}(ag, self = self)$
$\exists ag \neg \mathrm{BEL}(ag, ag = self)$ is satisfiable
$\exists ag\ \mathrm{BEL}(ag, \neg ag = self)$ is satisfiable

Thus, self and now behave essentially like constants.

This affects the way introspection should be characterized in the theory. One must say that if an agent believes ϕ, then he must believes that he himself believes ϕ, and similarly for negative introspection. One cannot replace 'self in the following proposition by a quantified-in variable unless the agent knows who he is.

Proposition
$\models \forall ag[\mathrm{BEL}(ag, \phi) \supset \mathrm{BEL}(ag, \mathrm{BEL}(self, \phi))]$
$\models \forall ag[\neg \mathrm{BEL}(ag, \phi) \supset \mathrm{BEL}(ag, \neg \mathrm{BEL}(self, \phi))]$

The theory accounts properly for indexical beliefs involving several agents. It distinguishes between alternative interpretations of 'John believes that Paul believes that he himself is tall',

BEL(JOHN,BEL(PAUL,TALL(self)))
and
BEL(JOHN, $\exists ag(ag = self \wedge$ BEL(PAUL, TALL(ag)))),

the first involving John's attribution of a self-belief to Paul and the second involving John's belief about a belief of Paul that is about himself.

3 An example

We will now illustrate through an example how the theory can be used to formalize a simple situation and prove that an agent is able to achieve a goal by doing an action given that he knows certain facts. The example involves the ability of an agent to make a phone call. A high-level view of this action takes calling to be a procedure involving a source phone and a destination phone. A direct formalization of this would require an agent to know what phones were involved. It is not obvious what *de re* knowledge of a phone should amount to, perhaps knowing the complete number of the phone including area code, or knowing its absolute location. But no matter what interpretation we take, the requirement seems clearly too strong. Firstly, if the agent knows that the call is local, i.e. his own area code is the same as that of the phone he wants to call, he does not need to know what either area code is in order to be able to make the call. Secondly, the agent only needs to know where the source phone is *relative to himself.* We will show how the calling action can be defined in terms of lower-level actions within our framework in order to avoid making unrealistic demands upon the agent's knowledge.

For simplicity, we will assume that any call to a phone in the same area is a local call (i.e. no need to dial T) and ignore international calls. We start by making various assumptions[16] about the types of action involved. The action DIALLOC consists of an agent dialing a local call on a phone that is within reach; the action takes the number to be dialed and the location of the phone relative to the agent as parameters. The prerequisites of DIALLOC are as follows: it is possible for any agent ag to DIALLOC any phone number n at a location RLOC(p.ag) relative to ag between now and some future time t if some unused phone p is at location RLOC(p.ag) relative to ag and p is within reach.[17]

Formally, this becomes:

Assumption 1: Prerequisites of DIALLOC
$\models \forall ag, n, p[\mathrm{ISPHNO}(n) \wedge \mathrm{PHONE}(p) \wedge \neg \mathrm{INUSE}(p) \wedge \mathrm{INREACH}(p, ag) \supset \exists t \Diamond ag \odot \mathrm{DIALLOC}(n, \mathrm{RLOC}(p, ag))@@t]$

Now for the effects of DIALLOC: if any agent ag makes a local dialing of the number of a destination phone p^ at a relative location RLOC(ps,ag) from now to time t, and an unused reachable source phone p_s is located at RLOC(ps,ag) relative to ag, and the area code of the location of ps is the same as that of p_d, this necessarily results in ps being connected to p_d at time t (this may cause ringing or a busy signal).

Assumption 2: Effects of DIALLOC
$\models \forall ag, p_s, p_d, t \Box[\mathrm{PHONE}(p_s) \wedge \mathrm{PHONE}(p_d) \wedge \neg \mathrm{INUSE}(p_s) \wedge \mathrm{INREACH}(p_s, ag) \wedge \mathrm{AREACODE}(\mathrm{LOC}(p_s)) = \mathrm{AREACODE}(\mathrm{LOC}(p_d)) \wedge ag \odot \mathrm{DIALLOC}(\mathrm{NOOF}(p_d), \mathrm{RLOC}(p_s, ag))@@t \supset \mathrm{CONNECTED}(p_s, p_d)@t]$

We must also assume that any agent knows what the action of making a local dialing of a number n at a reachable relative location l is if he knows what n and l are.

[16]That is, we will only deal with semantic structures where the assumptions come out true. Note that due to this, assumptions not only hold now, but at all times, and it is common knowledge that this is the case.

[17]RLOC(x,ag) can be defined as ROTATE(ALOC(x)-ALOC(ag), FACING(ag)) in a two-dimensional world, i.e. the vector difference between the absolute locations of x and ag rotated by the angle between the direction ag is facing and the absolute frame of reference; INREACH(x.ag) can be defined as LENGTH(RLOC(x, ag)) $\leq$ REACH(ag), i.e. p is within reach of agent ag if the length as a vector of the location of p relative to ag is less or equal to the reach of ag.

Assumption 3: DIALLOC is epistemically rigid
$\models \forall ag, n, l[KNOW(ag, ISPHNO(n) \wedge LENGTH(l) \leq REACH(self)) \supset \exists ac\ KNOW(ag, ac = DIALLOC(n, l))]$

We make similar assumptions concerning the prerequisites, effects, and epistemic rigidity of the action of making a long-distance dialing (DIALLD). This action takes the area code of the location of the destination phone as an argument in addition those of DIALLOC ISAC(c) means that c is an area code.

Assumption 4: Prerequisites of DIALLD
$\models \forall ag, c, n, p[ISAC(c) \wedge ISPHNO(n) \wedge PHONE(p) \wedge \neg INUSE(p) \wedge INREACH(p, ag) \supset \exists t \Diamond ag \odot DIALLD(c, n, RLOC(p, ag))@@t]$

Assumption 5: Effects of DIALLD
$\models \forall ag, p_s, p_d, t\ \Box[\ PHONE(p_s) \wedge PHONE(p_d) \wedge \neg INUSE(p_s) \wedge INREACH(p_s, ag) \wedge \neg AREACODE(LOC(p_s)) = AREACODE(LOC(p_d)) \wedge ag \odot DIALLD(AREACODE(LOC(p_d)), NOOF(p_d), RLOC(p_s, ag))@@t \supset CONNECTED(p_s, p_d)@t]$

Assumption 6: DIALLD is epistemically rigid
$\models \forall ag, c, n, l[KNOW(ag, ISAC(c) \wedge ISPHNO(n) \wedge LENGTH(l) \leq REACH(self)) \supset \exists ac\ KNOW(ag, ac = DIALLD(c, n, l))]$

We can now define calling (CALL) as a conditional action in terms of DIALLOC and DIALLD. Note that self, now and then respectively stand for the agent, starting time, and completion time of the action in such complex action definitions. An agent self calls a destination phone p_d from a source phone p$ (from now to then) when he makes a local dialing of the number of p_d at the relative location of ps if the area codes of the locations of ps and the destination phone p_d are the same, and makes a long-distance dialing of the area code of the location of p_d and number of p_d at the relative location of ps otherwise.[18]

Definition 1: Action CALL
$self \odot CALL(p_s, p_d) =$
$IF(AREACODE(LOC(p_s)) = AREACODE(LOC(p_d)),$
$\quad self \odot DIALLOC(NOOF(p_d), RLOC(p_s, self)),$
$\quad self \odot DIALLD(AREACODE(LOC(p_d)), NOOF(p_d), RLOC(p_s, self)))$

We also assume that the area code of the location of a phone that is within reach of an agent is the same as that of the location of the agent. Finally, we assume that n is a phone number if it is the number of some phone and that c is an area code if it is the area code of the location of something.

Assumption 7
$\models \forall p, ag[PHONE(p) \wedge INREACH(p, ag) \supset AREACODE(LOC(p)) = AREACODE(LOC(ag))]$

Assumption 8
$\models \forall n[\exists p(PHONE(p) \wedge n = NOOF(p)) \supset ISPHNO(n)] \wedge \forall c[\exists x\ c = AREACODE(LOC(x)) \supset ISAC(c)]$

[18] $IF(\phi, \alpha_1, \alpha_2)$ can be defined as $(\phi \supset \alpha_1) \wedge (\neg\phi \supset \alpha_2)$.

Given these assumptions, the theory allows us to prove proposition E1, which says that if an agent AG now knows what the location of an unused reachable phone Ps is relative to himself, and knows what the number of another phone P_d is, and either knows that his own area code is the same as that of P_d or knows that they are different and knows what the area code of P_d is, then he can achieve the goal of establishing a connection between P_s and P_d (at some future time) by calling P_d from Ps at the current time (see [Lesperance, 1989] for the proof).

Proposition E1
$\models \exists l, n\ KNOW(AG, PHONE(P_s) \wedge \neg INUSE(P_s) \wedge l = RLOC(P_s, self) \wedge INREACH(P_s, self) \wedge PHONE(P_d) \wedge n = NOOF(P_d)) \wedge$
$[KNOW(AG, AREACODE(LOC(self)) = AREACODE(LOC(P_d))) \vee \exists c\ KNOW(AG, AREACODE(LOC(self)) \neq AREACODE(LOC(P_d)) \wedge c = AREACODE(LOC(P_d)))] \supset$
$CAN(AG, self \odot CALL(P_s, P_d), CONNECTED(P_s, P_d))$

Note that the agent only needs to have *de re* knowledge of the relative location of the source phone and the number and possibly area code of the destination phone; he needs not have *de re* knowledge of either phone. The phones are arguments of the high-level action CALL, but these absolute references disappear when CALL is mapped into the lower-level actions DIALLOC and DIALLD in definition 1. It is important that this be done if the requirement for an agent to know what the initial action is in the definition of CAN is not to be overly restrictive. Note also how no requirement is made that the agent know who he is; all the knowledge that the agent is required to have about himself, such as whether his area code is the same as that of the destination phone and where the source phone is relative to him, is *de se* rather than *de re.*

4 Discussion

Agents act upon and perceive the world from a particular perspective. It is important to recognize this perspectival relativity if one is not to be overly demanding in specifying what they need to know in order to be able to achieve goals through action. They need not know very much about the objects they act upon because these objects are close at hand. And perception yields just the kind of agent-relative knowledge that is needed for action. We have developed a formal theory of knowledge and action that takes this relativity into account. The logic of belief proposed accommodates the indexical knowledge required for action and supplied by perception. The semantic account of belief given classifies agents not only in terms of what worlds are compatible with their beliefs, but also in terms of what their perspective upon these worlds must be. We have proposed a definition of ability that does not require the agent to know in an absolute sense who he is and what time it is. Through an example, it was shown how actions can be formalized so that an agent does not have to know in an absolute sense which objects are acted upon,

but only such facts as where the objects are relative to him. It was also shown how the theory together with such a formalization can be used to prove that an agent can achieve a goal by doing an action if he knows certain facts.

We are investigating the adequacy of the theory for characterizing agents with different kinds of action repertories at various levels of abstraction. We have advocated a strategy calling for the replacement of the requirements for agents to know what the objects they act upon are by requirements to "know where these objects are relative to them" or "know how to access these objects". We have shown how this strategy can be followed in the example of the previous section. This suggests that it might at least in principle be possible to systematically reduce explanations of ability involving knowledge of the objects acted upon to explanations involving only knowledge of primitive actions and other entities whose characterization as mental representations is unproblematic, such as relative locations, joint rotation angles, phone numbers, etc. If this characterization of the requirements action puts upon knowledge is correct, it should be possible to produce a revised definition of ability that embodies it. Preliminary examinations indicate that this is the case for various simple domains involving agents with limited action repertories, for example, a robot navigating in a simple world and manipulating objects.

However, this approach has some counter intuitive consequences: there is no requirement to even appeal to knowledge of relatively high-level entities such as relative locations; ability to do high-level actions, such as sending an electronic message to someone giving a talk at a remote institution, can be reduced to knowledge of what primitive actions to do in what circumstances, such as knowing what finger movements to do. But a real person would be prepared to deal with a huge number of unforeseen eventualities arising during the execution of his plan to send a message; he might have to track down the speaker who has moved, etc. It seems that for agents with open-ended action repertories, the flexibility of behavior cannot be explained without appealing to knowledge of at least some of the objects acted upon, for example, without the agent at some point knowing who the speaker was. The reduction does not capture the right generalizations. But if *de re* knowledge of objects is required in these cases, what does it really amount to and how is it related to the kind of agent-relative knowledge described earlier, which seems equally required for action. We will describe the result of our investigations of these issues in [Lesperance, 1990].

Acknowledgements

Hector Levesque has been an invaluable source of advice and support. Graeme Hirst, Joe Nunes, Jim des Rivieres, Calvin Ostrum, David Israel, and John Perry have provided useful discussion and/or advice. The referees' comments were also helpful.

References

[Agre and Chapman, 1987] Philip E. Agre and David Chapman. Pengi: An implementation of a theory of activity. In *Proceedings of the Sixth National Conference on Artificial Intelligence,* pages 268-272, Seattle, WA, July 1987. American Association for Artificial Intelligence, Morgan Kaufman Publishing.

[Hintikka, 1962] Jaakko Hintikka. *Knowledge and Belief* Cornell Univ. Press, Ithaca, NY, 1962.

[Konolidge, 1986] Kurt Konolidge. *A Dedudtion Model of Belief* Pitman Publishing, 1986.

[Lesperance, 1989] Yves Lesperance. A formal account of self-knowledge and action (extended version). Technical report, Department of Computer Science, University of Toronto, 1989. To appear.

[Lesperance, 1990] Yves Lesperance. *A Formal Theory of Indexical Knowledge and Action.* PhD thesis, Department of Computer Science, University of Toronto, 1990. To appear.

[Levesque, 1984] Hector J. Levesque. Foundations of a functional approach to knowledge representation. *Artificial Intelligence,* 23:155 212, 1984.

[Lewis, 1979] David I jewis. Attitudes *de dicto* and *de se. The Philosophical Review,* 88(4):513 543, 1979.

[M oore, 1985] Robert C. M oore. A formal theory of knowledge and action. In J. R. Hobbs and Robert C. Moore, editors, *Formal Theories of the Commonsense World.* Ablex Publishing, Norwood, NJ, 1985.

[Morgenstern, 1987] Leora Morgenstern. Knowledge preconditions for actions and plans. In *Proceedings of the Tenth International Joint Conference on Artificial Intelligence,* pages 867-874, Milan, August 1987. Morgan Kaufman Publishing.

[Perry, 1979] John Perry. The problem of the essential indexical. *Nous,* 13:3-21, 1979.

[Rescher and Urquhart, 1971] Nicholas Rescher and Alasdair Urquhart. *Temporal Logic.* Springer-Verlag, Vienna, 1971.

[Stalnaker, 1981] Robert C. Stalnaker. Indexical belief. *Synthese,* 49:129-151, 1981.

[Thomason, 1984] Rich mond C. Thomason. Combinations of tense and modality. In D. Gabbay and F. Gucnthner, editors, *Handbook of Philosophical Logic,* volume 2, pages 135- 165. D. Reidel Publishing, Dordrecht, Holland, 1984.