Naming and Identity in Epistemic Logics
Part I: The Propositional Case*

Adam J. Grove
Department of Computer Science
Stanford University, Stanford CA 94305, USA†

Joseph Y. Halpern
IBM Almaden Research Center
650 Harry Road,
San Jose, CA 95120-6099, USA

Abstract

Modal epistemic logics for many agents often assume a fixed one-to-one correspondence between *agents* and the *names* for agents that occur in the language. This assumption restricts the applicability of any logic because it prohibits, for instance, anonymous agents, agents with many names, named groups of agents, and relative (indexical) reference. Here we examine the principles involved in such cases, and give simple propositional logics that are expressive enough to cope with them all.

1 Introduction

It is much harder to represent an agent's knowledge about the world when the world contains many agents than it is when there is only one agent. Not only must the agent reason about the state of the world, he must also reason about what other agents know about the world, and what these other agents know about other agents' knowledge, and

*Research sponsored in part by the Air Force Office of Scientific Research (AFSC), under Contract F49620-91-C-0080. The United States Government is authorized to reproduce and distribute reprints for governmental purposes and distribute reprints for governmental purposes. The first author was also partially supported by an IBM graduate fellowship while this paper was being written. This paper is essentially identical to one that appears in *Journal of Logic and Computation* **3**:4, 1993, pp. 345–378. Portions of it appeared in preliminary form in a paper entitled "Naming and identity in a multi-agent epistemic logic" in *Principles of Knowledge Representation: Proceedings of the Second International Conference* (J. A. Allen, R. Fikes, and E. Sandewall, eds.), 1991, pp. 301–312.

†Present address: NEC Research Institute, 4 Independence Way, Princeton NJ 08540, USA

so on. One of the most subtle issues in such reasoning—and the focus of this paper—is that of *naming*. That is, how does one agent refer to others?

Many treatments of multi-agent epistemic logic make several simplifying, and therefore restrictive, assumptions about agents and their names. In particular it is often assumed that:

- There is a fixed and known collection of agents.
- Each agent has just one name (say agents $1,\ldots,n$, or agents Alice, Bob, Charlie, ...) and an agent can reason about other agents only in terms of these names ("Alice knows that Bob knows..." but not "Alice knows that someone with red hair knows...").
- Every name denotes just one agent.
- The composition of the system and the names of the agents are common knowledge, so that every agent knows them, and knows that every agent knows them, and knows that every agent knows that every agent knows them, and so on.

In other words, such logics make no practical distinction between the individual agents themselves, and the terms these agents use to refer to each other in reasoning (i.e., what we call *names*). In the propositional modal logics which are the basis for our investigations in this paper, this is implicit both in the syntax (where there are modal operators $K_1,\ldots,K_n$, one corresponding to each of the agents $1,\ldots,n$) and in the semantics (where there are binary relations $\mathcal{K}_1,\ldots,\mathcal{K}_n$ that describe the worlds that each of the agents considers possible).

There are many applications where these assumptions are quite reasonable, particularly those involving interactions among a fixed set of agents. For these applications, epistemic logic has provided a useful tool for formally modeling the interaction of agents. It has been successful in the analysis of distributed protocols (see [CM86, DM90, HM90, HZ87, Maz88, MT88, NT87] for some examples, and [Hal87] for an overview) and in artificial intelligence (for example, [RK86]). However, there are many situations where these assumptions are inappropriate. This is certainly true of the reasoning people do in everyday life. Consider that:

- There are very many people. No one knows just how many and, in any case, the number changes every second. It would certainly be impossible to know everyone's name. Even in small groups individuals can join or leave, and so the composition of the group might change frequently.
- Sometimes we do not know or care about another's "proper name". A customer and salesman in a shop might not bother to ask each other their name. Nevertheless, each will do considerable reasoning about the knowledge and goals of the other. Presumably, each refers to the the other using a description or role-name, such as "the salesman".

- The descriptions we use to refer to others sometimes refer to groups, not single individuals. Further, we frequently reason about the knowledge held by groups of people. This happens even if we are unable to list the names of the people in the group.
- We refer to others in many ways. Sometimes it can be truly surprising to discover that two descriptions or names we have for people actually refer to the same person (consider a costume party as an extreme example of this).
- When we think of names as descriptions, they are clearly not common knowledge. I may not know who "the salesman" really is. Or perhaps I do know this, but he does not know that I know.

The same issues arise even more forcefully in computer science, particularly in distributed systems and artificial intelligence research. All later examples will be taken from these areas, and will demonstrate some aspect of the general problem of "naming". The goal of this work is to find logics that are expressive enough for these examples. In particular, we want logics that make a real distinction between agents and names, because so many applications require this.

We have divided our results into two parts. The first part, contained in this paper, deals with propositional logics only. Propositional logics are of interest because of their simplicity; for instance, validity is decidable for all the logics we introduce here. Furthermore, all of the issues mentioned above have a simple solution even within a propositional framework. However, there are some problems that seem to require a stronger, first-order, logic and such a logic is the subject of part II ([Gro]).

Our approach is to start with one well known, but restrictive, logic called $S5_n$ and then generalize it as we identify weaknesses in expressivity. For instance, it is not hard to modify the associated possible-worlds semantics of knowledge, which we review in Section 2, to allow different sets of agents at different worlds. We can also enlarge the class of names to allow multiple names for agents and to allow a name to denote several agents. Finally, we can allow for *non-rigid* names, as has already been done in [DM90, MT88], to deal with names that denote different agents in different situations. Because the denotation of a name can vary, some agents may be uncertain who the name actually denotes. In particular, the denotation of a non-rigid name is not necessarily common knowledge.

In Section 3 we examine a logic that includes all these extensions. This logic, and the variations of it that we consider, go quite far towards solving the problems with $S5_n$. We give complete axiomatizations, and show how different assumptions about naming can be captured semantically and axiomatically. For example, we can give simple logics which make any subset of the assumptions mentioned at the start of this paper.

However, and perhaps surprisingly, the logics of Section 3 turn out to have several more subtle weaknesses. One difficulty is that, in anonymous or symmetric systems, an agent might not have any name at all. And even when he does have a name, it is possible that the agent does not know what this name is (indeed, we will see later that it might not even know *who* he is). This possibility arises as soon as we allow names to be non-rigid.

To see the problems this could cause, imagine a distributed computer network in which a process a broadcasts a message m. Informally, we could say a knows that all his neighbors (with respect to the network topology) will, soon, know the content of m. This is because a knows that he sent the message. Further, a knows that he knows this.

Although these statements seem simple enough, it is hard to express them satisfactorily in standard epistemic logic. The obvious approach would be to translate a statement like "a knows that he knows" as "a knows that a knows". But these two assertions, which relate to the introspective abilities of a, are equivalent only if *a knows that he is a*, that is, if a knows his name. This is an unrealistic assumption in anonymous systems. Suppose the system is highly symmetric, so that many processes can be in identical states (they run the same program, have the same values for all variables, and so on; in particular, we assume that they do not have a unique name as part of their state). There is just no information available to a that might not be available to someone else also (process c say). But then a really doesn't know that he is not c! Or consider two identical robots, which we—as external observers—call $R2$ and $D2$. A command, "$R2$, come here!" will be obeyed only if one of these agents has been programmed to respond to such orders: in effect, he must know his name. But it is possible that the agents were not programmed this way. How can we design a logic to deal with situations like these?

Part of the answer is that the language must be extended so that it includes the equivalent of pronouns like *I* or *he*. Also very useful is the ability to refer to other agents relatively (process a referring to *his neighbors,* for example). In Section 4 we present a propositional logic that includes these features, and demonstrate a modification of the possible-worlds approach that captures these notions semantically.

Finally, in Section 5 we note that there are some more difficult problems with names that seem to require a first-order logic of some sort for their solution. A complete discussion of these problems, and a logic that addresses them, can be found in part II of this work ([Gro]). Nevertheless, an important conclusion from this first paper is that many interesting and useful models of "naming" can be captured well in propositional logic. We present a range of related logics that do this.

Several of the issues we discuss in this paper have been looked at before. We briefly mention some of the related work here; more discussion appears later in the paper (in particular, in Section 4.5).

The idea of using non-rigid names for groups of agents in propositional epistemic logic can already be found in [DM90, MT88]. There are significant technical differences between this work and ours, which we discuss in Section 3. An even more important contrast is that [DM90, MT88] each address one particular application, and so do not examine any general theory of how nonrigid names work, nor do they beyond defining the semantics and an appropriate language for the case of interest to them. We examine a range of very general logics, and also provide complete axiomatizations and some complexity results.

The main part of our paper is Section 4, which begins by discussing some of the weaknesses inherent in the logics of Section 3 and [DM90, MT88]. Our concerns with the use of the pronoun I, and some of the issues concerning relative names and anonymity, are closely related to earlier philosophical work such as [Cas68, Per79, Lew79]. With regards to our proposed solution to the problems we find, Lewis's work is certainly most relevant. Lewis develops and argues for a semantic account of such *de se* knowledge (this is Lewis's term) which is substantially equivalent to the semantics we adopt. Lewis's paper gives convincing philosophical arguments for the semantics he proposes. Nevertheless, Lewis does not present any formal system that incorporates these ideas, whereas we examine several. We also argue that these ideas may be important in practical applications. The only formal system that we know about, aside from our own, that is based on these semantics is Lespérance's work [Les89, Les91]. Lespérance's logic is very different to ours, reflecting the fact that it was developed to address different goals; we discuss the differences in Section 4.5. Lespérance's work is also important because he argues that *de se* knowledge (in Lewis's terminology) is necessary in certain applications of modal logic to robotics and artificial intelligence. His arguments are particularly interesting because they are more concrete, and in some cases quite different in character, from Lewis's and from our own. Finally, a recent paper by Seager [Sea90] looks at a multiple-agent logic for belief, and addresses the issue of agents that must refer to themselves indexically. We also compare this work with ours in some detail later.

2 Possible-world semantics: a review

We base our investigations on a standard possible-worlds approach to epistemic logic. We provide a brief review here; the reader can find more details in [Che80, HC84, HM85].

The logic is used to model the knowledge of a group of n agents, $1, \ldots, n$, who reason about a world described using a set Φ of primitive propositions. A formula in the language can be any propositional symbol from Φ, or a Boolean combination (formed using $\neg$ and $\wedge$) of other formulas. We use other Boolean connectives such as $\vee$, $\Rightarrow$, and $\Leftrightarrow$ occasionally in formulas; they can be defined in terms of $\wedge$ and $\neg$ in the usual way. In addition, we have modal operators $K_1, \ldots, K_n$, one for each agent. If φ is a formula, then so are $K_1\varphi, K_2\varphi, \ldots, K_n\varphi$. We read $K_i\varphi$ as *(agent) i knows* φ.

Semantically, a possible-worlds structure (or *Kripke* structure) over Φ for n agents is a tuple $M = (W, \pi, \mathcal{K}_1, \ldots, \mathcal{K}_n)$. W is a set of states, or possible worlds, and π associates with each possible world a truth assignment to the propositions Φ. That is, $\pi : W \rightarrow (\Phi \rightarrow \{\texttt{true}, \texttt{false}\})$. Each $\mathcal{K}_i$ is a binary relation on W, i.e., $\mathcal{K}_i \subseteq W \times W$.

The set $\{w' \in W : (w,w') \in \mathcal{K}_i\}$ can be viewed as *the set of worlds agent i considers possible from w*. $\mathcal{K}_i$ is intended to model knowledge by capturing the ignorance an agent has about what the world is like: if the world is in fact w, agent i considers it possible that the world is one of $\{w' : (w,w') \in \mathcal{K}_i\}$. The $\mathcal{K}_i$ are sometimes called (epistemic) *accessibility* relations.

A formula φ is either true or false at a pair (M,w) consisting of a structure M and a world w in M. We define what it means for φ to be true at world w in structure M, written $(M,w) \models \varphi$, by induction on the structure of φ:

$(M,w) \models p$ (for $p \in \Phi$) if $\pi(w)(p) = \mathtt{true}$

$(M,w) \models \neg\varphi$ if not $(M,w) \models \varphi$

$(M,w) \models \varphi \wedge \psi$ if both $(M,w) \models \varphi$ and $(M,w) \models \psi$

$(M,w) \models K_i\varphi$ if $(M,w') \models \varphi$ for all w' such that $(w,w') \in \mathcal{K}_i$.

We often omit the structure M, writing $w \models \varphi$ rather than $(M,w) \models \varphi$, when M is not relevant or is clear from context.

Let $\mathcal{M}$ be the class of all possible-worlds structures. A formula φ is said to be valid in a structure M, written $M \models \varphi$, if $(M,w) \models \varphi$ for all worlds w in M; φ is said to be valid with respect to $\mathcal{M}$ if $M \models \varphi$ for all $M \in \mathcal{M}$.

We will be interested in characterizing the properties of the logics we study by providing sound and complete axiomatizations for them. It is well known that the following system, called K_n, with axiom schemes A1, A2 and inference rules R1, R2, is sound and complete with respect to $\mathcal{M}$; that is, a formula is provable in this system iff it is valid with respect to $\mathcal{M}$ (see [HM85] for a proof):

A1. All instances of propositional tautologies

A2. $K_i\varphi \wedge K_i(\varphi \Rightarrow \psi) \Rightarrow K_i\psi$

R1. From φ and $\varphi \Rightarrow \psi$ infer ψ (*modus ponens*)

R2. From φ, infer $K_i\varphi$ (*knowledge generalization*)

In many cases of interest, the binary relations $\mathcal{K}_i$, $i = 1,\ldots,n$, satisfy some additional conditions. We will be most interested in the case where the $\mathcal{K}_i$'s are equivalence relations, that is, they are reflexive, symmetric, and transitive.[1] In this case, we need to add the following three axioms to K_n to get a complete axiomatization; the resulting system is called $\mathrm{S5}_n$ (see [HM85] for a proof):

[1]It is not always realistic to look at accessibility relations that are equivalence relations. In particular, logics for *belief* usually omit the reflexivity requirement because an agent might (mistakenly) think that the real world is impossible. Of course, issues of naming are important in these logics as well. Essentially all of the results in this paper can be restated for logic of belief; see [Gro92].

A3. $K_i\varphi \Rightarrow \varphi$

A4. $K_i\varphi \Rightarrow K_iK_i\varphi$

A5. $\neg K_i\varphi \Rightarrow K_i\neg K_i\varphi$

A4 and A5 are called axioms of *positive introspection* and *negative introspection*, respectively. They say that an agent knows what he knows and what he does not know.

One important application area for the logics of knowledge we have been considering is in terms of understanding, reasoning about, and analyzing distributed systems. We briefly review the model here; the interested reader should consult [Hal87, HM90] for more details.

As before there are n agents $1, \ldots, n$ (the processes in the system), and a collection Φ of primitive propositions, that are typically intended to denote events of interest in the system (such as "the value of variable x is 0" or "process 1 has just received a message from process 3"). We characterize the system at a given point in time in terms of a *global state*; this is a tuple $(s_1, \ldots, s_n)$ where s_i is the *local state* of process i.[2] The local states of an agent intuitively encode all the information that the process has available at a given point in time. The details will be application dependent. In typical distributed systems applications, it will include the values of variables and a history of messages received. On the other hand, if we are modeling a group of agents playing a poker game, the local state will include the cards that the agent holds, and the bets that have been made thus far.

A system is not a static entity; it is constantly changing over time. In order to capture this, we define a *run* of the system to be a function from time to global states. Intuitively, a run is a complete description of what happens over time in one possible execution of the system. A *point* is a pair (r, m) consisting of a run r and a time m. The global state $r(m)$ describes the state of the system at the point (r, m). Formally, we take a *system* to consist of a set of runs. Intuitively, the system includes all the possible executions of the system, that is, all the different ways it could evolve through time.

Assume we are given a truth assignment π that tells us, for each primitive proposition $p \in \Phi$, whether p is true or false at the point (r, m) in a system $\mathcal{R}$. Typically, the truth value of p at (r, m) will be easily determined from the global state $r(m)$. For example, if p is "process 1 has just received a message from process 3", we would simply look at process 1's local state in $r(m)$ to see if it has indeed received such a message (this is assuming that 1's state reflects the messages that it has just received.) Given a truth assignment π, we can view a system as a possible-worlds structure in a straightforward way. The possible worlds are simply the points. The binary relations

[2] It is often conceptually useful to add one more component to the global state; this is the state of the *environment*, which intuitively describes everything relevant to the system not included in the local states of the processes. For ease of exposition, we have chosen to omit the environment state here. Since we can just view the environment as another agent, this omission does not affect any of our discussion.

$\mathcal{K}_i$ are determined by the local states: we take $((r,m),(r',m')) \in \mathcal{K}_i$ iff i is in the same the local state in both the global states $r(m)$ and $r'(m')$. Thus, $\mathcal{K}_i$ is an equivalence relation on points. (This concrete model is the reason for our interest in systems where $\mathcal{K}_i$ is an equivalence relation; in philosophical discussion or in application to real—human—knowledge this assumption is generally accepted to be unrealistic.) Intuitively, we are saying that agent i can tell two possible global states apart exactly when its local state is different in the two. If the local state really captures all the features of the system that are visible to the agent, then the agent would not simultaneously consider possible two worlds where its state differed (because—so long as we do not take computational considerations into account—the agent *can* tell them apart). Conversely, an agent cannot tell apart two points where it has the same local state. Given this way of viewing a distributed system as a possible-worlds structure, we can ascribe knowledge to the processes using the same definitions of $\models$ used above. Thus, it makes perfect sense to say that, at the point (r,m), process 1 knows that process 2 knows that process 1 received a message from process 3.

There is one feature of these definitions that deserves further comment, since it will be important later. One intuition we have about the use of epistemic logic as a modeling tool, such as in the theory of distributed systems, is that the knowledge ascribed to an agent should be determined solely by his local state. However, the definition of $\mathcal{K}_i$ given above seems to violate this principle in certain cases. To see the problem, notice that nothing we have said so far rules out the possibility that agent i's local state at point (r,m) could be be the same as agent j's local state at some other point (r',m'). Say this is state s. Although both agents are in the same state, it is not necessarily the case that i's knowledge at (r,m) is the same as j's knowledge at (r',m'). That is, given our definitions, the set of points these two agents consider possible can be different (for instance, if i would not be in state s in (r',m') then he will not consider (r',m') to be possible, whereas j obviously does). This seems wrong: if i and j are both in state s and so have access to the same information, how can their knowledge be different? Later we discuss new models for knowledge that do not have this problem. However, one way of understanding the model above is that it implicitly assumes that agent i *knows who he is*: intuitively, he doesn't consider (r',m') possible because at (r',m') it is j, not i, who is in state s. More concretely, this means that that each agent has access to a unique identifer of some sort, in addition to the rest of its local state. Under this assumption, although it may seem that two agents are in the same state s, this will not really be the case if we look at their *complete* state (which also includes an identifier that distinguishes any one agent from the others). Although the assumption that an agent knows its own name is often a reasonable one, it is not necessarily one we always want to make. As we shall see, it is intimately connected with the introspection axioms. This issue is quite subtle, but surprisingly important. We examine in in much more detail later in the paper, and in particular in Section 4.1.

3 Adding non-rigid names

In the next two subsections, we remove the assumptions regarding agents and names that have been made in propositional epistemic logics like $S5_n$. We do this in two stages, to give the reader a feeling for the issues involved. Here, we work in the context of possible-worlds structures, but we can easily translate our remarks to the distributed systems framework (from where much of our motivation has come).

3.1 Allowing a different set of agents in each world

Our first step is to remove the assumption that there is a fixed set of agents, $1, \ldots, n$. This is fairly easy to do. We assume that in each world w, there is a set A_w of agents that exist in w. Of course, a very special case is where A_w is identical for all worlds w. In this circumstance, an agent always *knows* who else is present because it is the same set in all the worlds he considers possible. But if the collection of agents is subject to change (as is the case in a dynamically evolving system) or if the agent in question has been designed to operate in varied environments, then it might not possess this knowledge.

As before, agents' knowledge is encoded in terms of binary relations on worlds: for every agent a in A $(= \cup_w A_w)$ there is a corresponding relation $\mathcal{K}_a$. Let $W_a = \{w : a \in A_w\}$; intuitively, W_a consists of the worlds where agent a exists. In this section, we assume that $\mathcal{K}$ satisfies three properties. First, we require that when $(w, w') \in \mathcal{K}_a$, then $w \in W_a$. Intuitively, this restriction is reasonable because we do not want to ascribe knowledge to an agent at any world where he is not present. Nevertheless, given just this first condition, an agent a could consider possible some world w' where it doesn't actually exist (i.e, $a \notin A_{w'}$). This is not completely implausible: the intuition is that this could happen if the agent "doesn't know who he is". But even in this case, it is not reasonable that the agent could consider any w' at all to be possible. If he is not actually present in w', there should at least be some other agent present in w' which is very similar to him (in the distributed systems model, this other agent must be in the same state). Otherwise, the agent would surely be able to eliminate w' as a possibility. This issue is subtle, and in fact is best approached with the more general framework we develop in Section 4. For this section, we assume that whenever $(w, w') \in \mathcal{K}_a$, then $w' \in W_a$ (an agent cannot consider a world possible where he doesn't exist).[3] Finally, because of our interest in the distributed systems model, we restrict attention to the case where $\mathcal{K}_a$ is an equivalence relation on W_a, although it would be straightforward to extend all our work to other binary relations.

It is easy to modify the distributed systems model presented in the previous section

[3]It turns out that this restriction is inessential. Neither this requirement about the range of the accessibility relation nor the requirements of reflexivity, transitivity and symmetry that we make subsequently have any impact on the properties of our first logic. They aid intuition, however, and become important in Section 3.4, so we have decided to adopt them from the outset.

in order to capture the possibility of there being different agents at different points in the system. The global state at point (r, m) now consists of a collection of agents $A_{(r,m)}$ (the agents that exist there), and a function from this set to to local states. Before, the set of agents was fixed as $\{1, \ldots, n\}$ in all global states, so we could characterize the associated function using an n-tuple of local states. In the current, more general, model, the domain of this function can vary and so needs to be given individually for each global state. We define a binary relation corresponding to each agent $a \in A$ just as before: $((r, m), (r', m')) \in \mathcal{K}_a$ iff a exists in both (r, m) and (r', m') and has the same local state at each point.

3.2 More general names

We should modify the syntax of the language to reflect this change in the semantics. We could, of course, simply include a modal operator K_a corresponding to each agent $a \in A$. However, it turns out to be more useful to change the language even further. For example, if we consider a *token-passing* system where, at any point in time, exactly one process holds a token (and thus is able to carry out actions), we may want to reason about the token-holder's knowledge, and make a statement of the form "process 1 knows that the token-holder knows ...". This statement makes perfect sense even if process 1 does not know which process is the token-holder. Thus, it seems useful to extend the language with a *name* representing the token-holder, and to permit reasoning about the knowledge of the agent with that name. Even more generally, we may want to allow a name to denote a set of agents. In [DM90, MT88], there are modal operators that refer to the set of correct processes (i.e., those processes that are still functioning correctly, of which there may conceivably be none, one, or many).

So, in general, we assume that we have a collection $\mathcal{N}$ of *names*, each of which denotes a (possibly empty) set of agents at each world. The appropriate choice of names depends on the application, just as the choice of primitive propositions does. We then have a semantic function μ that associates with each world w and name $\mathbf{n}$, the set $\mu(w, \mathbf{n})$ of agents with name $\mathbf{n}$ in world w.

In the special case where there is a fixed set of agents $\{1, \ldots, n\}$, we can let the symbols $1, \ldots, n$ denote both the agents and their names. Thus, we would have, for example, $\mu(w, 1) = \{1\}$ for all w. This collection of names corresponds to the syntax of the logic we reviewed in Section 2, but we must note that these names have several special properties. A name like 1 is an example of a *rigid* name, since it denotes the same agent in each world. But more generally, we may want *non-rigid* names. For example, the process who holds the token, or the set of correct processes, will vary from point to point, so rigidity is inappropriate. Second, names $1 \ldots n$ each denote just *one* agent (in any world). While this assumption allows the logic to have a substantially simpler syntax than the general case where we allow groups of agents, we have seen that it is just too restrictive for us. On the other hand, we wish to retain the simplicity of propositional logics as far as is possible; here we do not want to consider a full first (or higher) order logic where we can speak about sets of agents directly. Finally, a somewhat more subtle point, we note that this simple logic gives every agent a name. For instance, in any world agent 1 is referred to by some name (in fact, the name 1). Thus, it seems that agents cannot be anonymous. Our more general semantics (with $\mathcal{N}$ and μ) does not have this requirement: instead, there may simply be no $\mathbf{n} \in \mathcal{N}$ with $1 \in \mu(w, \mathbf{n})$, or perhaps there are such names $\mathbf{n}$ but they all refer to large groups of agents. Anonymity can be modeled just as a lack of individual names. This suggestion, that anonymity is best captured with a weak language (few names) may seem unusual. After all, one could always suppose that $\mathcal{N}$ is extended to include a name for every agent, and the resulting language would seem to be more powerful. Our point is that adding names is not necessarily helpful if no one *knows* who the names denote; we return to this issue several times later. Complete anonymity can be regarded as either the lack of individual names, or as a situation where such names exist but no one knows about them. The latter viewpoint is still consistent with our logic, but so is the former, and the former is often simpler.

So, to summarize, our semantics for names are more general because they allow *non-rigid names*, they allow names to denote *groups* of agents, and we allow direct expression of anonymity because an agent *need not have any name.*

Formally, we take a *possible-worlds structure for naming* over Φ and $\mathcal{N}$ to be a tuple $M = (W, A, \alpha, \mathcal{K}, \pi, \mu)$. W and A are just sets, respectively the *worlds* and *agents.* The function $\alpha : W \rightarrow \mathcal{P}(A)$ selects the agents that exist at a given possible world (where $\mathcal{P}(A)$ is the set of subsets of A). We usually write A_w rather than $\alpha(w)$ for $w \in W$. The truth assignment π remains as before, a mapping that associates with each world a truth assignment to the primitive propositions in Φ. $\mathcal{K}$ is a function from agents to binary relations on W: as discussed earlier, we require that the domain and range of $\mathcal{K}(a)$ be contained in W_a, and further, that $K(a)$ be an equivalence relation (on its domain). Instead of $\mathcal{K}(a)$, we often write $\mathcal{K}_a$. Finally, μ interprets names; it is a function mapping a name $\mathbf{n}$ and a world w to some set of agents $\mu(w, \mathbf{n})$ (in fact, some subset of A_w). We take $\mathcal{M}_{\mathcal{N}}$ to be the class of all possible-worlds structures for naming.

What syntax should correspond to this semantics? We clearly would like modal operators that allow us to refer to the knowledge of agents with a particular name. As observed in [HM90], once we have names corresponding to a group of agents rather than just a single agent, we have a number of different ways of capturing the knowledge of the group. In [HM90], it is assumed that there is a fixed set of agents $\{1, \ldots, n\}$. For each subset $G \subseteq \{1, \ldots, n\}$, the modal operators D_G, S_G, E_G, and C_G are introduced, read as *it is distributed knowledge among the agents in G*, *someone in G knows*, *everyone in G knows*, and *it is common knowledge among the agents in G*, respectively. Roughly speaking, a fact φ is distributed knowledge among the agents in G if φ is a consequence of the pooled knowledge of the agents in G. For example, if agent 1 knows φ and agent 2 knows $\varphi \Rightarrow \psi$, then ψ is distributed knowledge among $\{1, 2\}$. A fact φ is common knowledge among the agents in G if everyone in G knows φ, everyone in G knows that everyone in G knows φ, everyone in G knows that everyone in G knows that everyone in G knows φ, and so on.

For simplicity, we focus here on only two of these operators, the ones corresponding to someone knows and to everyone knows. Although the other operators are certainly of interest, many of the points we want to bring out already become clear with these two. We remark that these operators are also useful in practice. It is particularly useful to say "everyone knows" when specifying initial conditions on a system (perhaps, at time 0, every *correct* process knows some fact, such as that at most half of the other processes have failed). On the other hand, *someone* is good for modeling knowledge acquired as the system evolves. For if a receives a message, it knows that *some* other agent sent the message (so, presumably, also knows its content). The manner in which the message was received constrains the set of possible senders; for example, if the message reaches a over some channel c say, then *someone* who is connected to a by c sent it.

So, corresponding to each name $\mathbf{n}$, we have two modal operators, $E_{\mathbf{n}}$ and $S_{\mathbf{n}}$. Intuitively, these say "everyone with name $\mathbf{n}$ knows" and "someone with name $\mathbf{n}$ knows". Let us call the new language (closed under the new operators, rather than the K_i as before) $\mathcal{L}_{\mathcal{N}}$.

We extend the definition of $\models$ to $\mathcal{L}_{\mathcal{N}}$ formulas as follows:

$(M, w) \models E_{\mathbf{n}}\varphi$ if, for all $a \in \mu(w, \mathbf{n})$, we have $w' \models \varphi$ for all w' with $(w, w') \in \mathcal{K}_a$

$(M, w) \models S_{\mathbf{n}}\varphi$ if, for some $a \in \mu(w, \mathbf{n})$, we have $w' \models \varphi$ for all w' with $(w, w') \in \mathcal{K}_a$.

Our definitions imply that $(M, w) \models E_{\mathbf{n}}\varphi$ if $\mu(w, \mathbf{n}) = \emptyset$, while $(M, w) \not\models S_{\mathbf{n}}\varphi$ if $\mu(w, \mathbf{n}) = \emptyset$. $E_{\mathbf{n}}$ essentially acts as a universal, while $S_{\mathbf{n}}$ essentially acts as an existential. However, $E_{\mathbf{n}}$ and $S_{\mathbf{n}}$ are not dual, although they may appear to be; generally, one cannot be defined in terms of the other. We could have $\neg S_{\mathbf{n}}\neg\varphi$—no one with name $\mathbf{n}$ knows $\neg\varphi$—even though not everyone with name $\mathbf{n}$ knows φ: some of the agents with name $\mathbf{n}$ may not know either. We also remark that our definition of $E_{\mathbf{n}}$ is slightly different from that given in [DM90, MT88]; we return to this point later.

Notice that if a name $\mathbf{n}$ denotes a unique agent in all worlds, then $E_{\mathbf{n}}$ and $S_{\mathbf{n}}$ are identical. In such a case, we write the more intuitive $K_{\mathbf{n}}$: "the (unique) agent denoted by $\mathbf{n}$ knows".

3.3 Properties of epistemic logic with names

It is easy to see that $E_{\mathbf{n}}$ satisfies an analogue of A2: $(E_{\mathbf{n}}\varphi \wedge E_{\mathbf{n}}(\varphi \Rightarrow \psi)) \Rightarrow E_{\mathbf{n}}\psi$ is valid. And although we have assumed that the binary relations corresponding to each agent are equivalence relations, none of the properties of S5—that is, the analogues of A3, A4, and A5—hold for $E_{\mathbf{n}}$. For example, $E_{\mathbf{n}}\varphi \Rightarrow \varphi$ does not necessarily hold at a world where there is no agent with name $\mathbf{n}$, since in that case $E_{\mathbf{n}}\varphi$ is vacuously true. (However, the weaker principle $\neg E_{\mathbf{n}} \mathit{false} \wedge E_{\mathbf{n}}\varphi \Rightarrow \varphi$ will hold always, because $\neg E_{\mathbf{n}} \mathit{false}$

is true just when at least one agent is named $\mathbf{n}$.[4]) It is also not hard to show that the introspection axioms A4 and A5 do not hold. Using standard techniques for proving completeness in modal logics, it can be shown that $E_{\mathbf{n}}$ by itself satisfies precisely the axioms of K together with the axiom we have just seen, $\neg E_{\mathbf{n}} \mathit{false} \wedge E_{\mathbf{n}}\varphi \Rightarrow \varphi$.

This is the same logic as for the knowledge of one agent, with the single constraint placed on the $\mathcal{K}$ relation that when the agent considers any worlds at all possible from some world w, it must consider w itself possible (the relation $\mathcal{K}$ is reflexive on its domain). In retrospect, this may not be too surprising. Instead of viewing $E_{\mathbf{n}}$ as the knowledge of a group of agents, we could equally well consider it to model the knowledge of one quite ignorant agent (intuitively, an agent who knows only what everyone in the group knows).

Turning to $S_{\mathbf{n}}$, we can see that it does not even satisfy an analogue of A2. The reason is that while some agent with name $\mathbf{n}$ may know φ and another may know $\varphi \Rightarrow \psi$, there may be no agent with name $\mathbf{n}$ that knows ψ (unless there is a unique agent with name $\mathbf{n}$). While $S_{\mathbf{n}}\varphi \Rightarrow \varphi$ is valid, the analogues of the axioms A4 and A5 do not hold for $S_{\mathbf{n}}$. The logic of $S_{\mathbf{n}}$ alone is what has been called *monotonic* ([Che80, ch. 7–9]).[5]

Besides the properties of $E_{\mathbf{n}}$ and $S_{\mathbf{n}}$ individually, there is some interaction between these operators. For example, although an analogue of A2 does not hold for $S_{\mathbf{n}}$, a modified version does hold: if someone with name $\mathbf{n}$ knows φ and *everyone* with name $\mathbf{n}$ knows that φ implies ψ, then someone with name $\mathbf{n}$ knows ψ. In addition, if $\neg E_{\mathbf{n}} \mathit{false}$ holds, then there is at least one agent with name $\mathbf{n}$, so that $S_{\mathbf{n}} \mathit{true}$ holds.

It turns out that these axioms summarize the interaction between $S_{\mathbf{n}}$ and $E_{\mathbf{n}}$. Consider the following axiom system, which we denote $AX_{\mathcal{N}}$. We group the axioms and rules into axioms and rules for propositional reasoning (A1 and R1), for reasoning about $S_{\mathbf{n}}$ (S1), for reasoning about $E_{\mathbf{n}}$ (E1 and E2), and for combined reasoning about $E_{\mathbf{n}}$ and $S_{\mathbf{n}}$ (C1 and C2). All these axioms and rules hold for every $\mathbf{n} \in \mathcal{N}$.

A1. All instances of propositional tautologies

R1. From φ and $\varphi \Rightarrow \psi$ infer ψ (modus ponens)

S1. $S_{\mathbf{n}}\varphi \Rightarrow \varphi$

E1. $E_{\mathbf{n}}\varphi \wedge E_{\mathbf{n}}(\varphi \Rightarrow \psi) \Rightarrow E_{\mathbf{n}}\psi$

E2. From φ infer $E_{\mathbf{n}}\varphi$

C1. $S_{\mathbf{n}}\varphi \wedge E_{\mathbf{n}}(\varphi \Rightarrow \psi) \Rightarrow S_{\mathbf{n}}\psi$

[4] Throughout, we use *false* as an abbreviation for some fixed contradiction (such as $p \wedge \neg p$) and *true* to stand for some fixed tautology.

[5] Monotonic modal logics have a semantic theory where a possible world is related to *sets of* sets of possible worlds (not just one such set of alternatives as in the more familiar normal logics). The intuition here is similar to that discussed in [FH88]: we can view each member of $\mathbf{n}$ as corresponding to some "frame of mind" of an agent.

C2. $\neg E_{\mathbf{n}} \mathit{false} \Rightarrow S_{\mathbf{n}} \mathit{true}$

Theorem 3.1: *$AX_{\mathcal{N}}$ is a sound and complete axiomatization with respect to $\mathcal{M}_{\mathcal{N}}$.*

Proof: Soundness is straightforward. We defer the details of the completeness proof to Appendix A. ∎

This theorem shows that we can enumerate all formulas that are valid for our semantics. Of course, this does not prove that the validity problem—deciding whether a given formula is valid in the logic—is decidable. However, it turns out that this is the case. In fact, we have the following theorem:

Theorem 3.2: *The problem of deciding whether a formula in $\mathcal{L}_{\mathcal{N}}$ is valid with respect to $\mathcal{M}_{\mathcal{N}}$ is* PSPACE*-complete.*

Proof: This theorem can be proved using similar techniques to those found in [HM85]. In Appendix B we discuss further how these techniques apply here. ∎

Just as with standard epistemic logics, we can explore the effect of additional semantic conditions, and look at the corresponding axioms required. One significant class of conditions arises from considering the relationships that might hold *between* names. For example, it could be that one name always denotes a set of agents contained in another. If it is known that some process (which has name 1, say) never fails then, somewhat informally, we might say $1 \subseteq \mathit{correct}$. In the general case, whenever two names $\mathbf{n}, \mathbf{n}'$ are such that $\mathbf{n} \subseteq \mathbf{n}'$ in this sense, it is easy to see that the following two axioms are sound: $E_{\mathbf{n}'} p \Rightarrow E_{\mathbf{n}} p$ and $S_{\mathbf{n}} p \Rightarrow S_{\mathbf{n}'} p$. In fact, adding these axioms gives a logic complete for such situations.

Further work along these lines would be to look at the name denoting the union (or intersection, etc.) of two other names: indeed, the natural generalization is to consider an algebra of names We later look briefly at how this might be done. Another recent work that looks at algebraic structure on sets of agents, although not in the context of epistemic logic, is [ABLP91].

There is also another, different, direction we can take when looking at variants of our basic logic, which involves placing further constraints on the nature of one particular name. For example, our logic was general in that names might occasionally (i.e., at some possible worlds) denote no agents at all. Yet sometimes we can be certain that this is impossible. Adding $E_{\mathbf{n}}\varphi \Rightarrow S_{\mathbf{n}}\varphi$ gives a logic sound and complete for structures where the name $\mathbf{n}$ always denotes at least one agent. We can also consider the converse, namely, that $\mathbf{n}$ always denotes at most one agent. If $\mathbf{n}$ denotes at most one name, then the axiom $S_{\mathbf{n}}\varphi \Rightarrow E_{\mathbf{n}}\varphi$ is sound. In fact, this axiom characterizes the constraint that $\mathbf{n}$ satisfies at most one name.

Putting these observations together, it follows that $S_{\mathbf{n}}\varphi \Leftrightarrow E_{\mathbf{n}}\varphi$ characterizes the situation where $\mathbf{n}$ is a unique identifier, always denoting exactly one agent. In this case, we write $K_{\mathbf{n}}\varphi$ instead of either $S_{\mathbf{n}}\varphi$ or $E_{\mathbf{n}}\varphi$. However, even if names $\mathbf{n}$ are always unique identifiers, and the underlying binary relations are equivalence relations, we still do not recover the familiar introspection axioms of S5. The problem is that the non-rigidity of names implies that agents might not know their names. For example, suppose that agent a has won a prize (which has a unique winner), but has not yet been informed of this fact. Then if $\mathbf{n}$ denotes "the prize-winner", and a knows a fact φ, then $K_{\mathbf{n}}\varphi$ holds, although $K_{\mathbf{n}}K_{\mathbf{n}}\varphi$ does not. In the next subsection we consider the impact of knowing one's name.

3.4 Knowing one's name(s)

We say that *agents know their names* in a structure M, if, whenever an agent has name $\mathbf{n}$ in some world, it also has name $\mathbf{n}$ in all the worlds it considers possible. Formally, M is a structure where agents know their names if $a \in \mu(w, \mathbf{n})$ implies $a \in \mu(w', \mathbf{n})$ for all w' such that $(w, w') \in K_a$. Let $\mathcal{M}'_{\mathcal{N}}$ be the subclass of $\mathcal{M}_{\mathcal{N}}$ where agents know their names.

Note that if $\mathbf{n}$ is a unique identifier, then we can show that the modal operator $K_{\mathbf{n}}$ satisfies the axioms of S5 in all structures in $\mathcal{M}'_{\mathcal{N}}$. This reinforces the intuition that the axioms A4 and A5 do not just have to do with introspection, but are intimately bound up with knowing one's name.

More generally, what are the properties of $E_{\mathbf{n}}$ and $S_{\mathbf{n}}$ in structures where agents know their names? It is not hard to see that we get a positive introspection axiom for $S_{\mathbf{n}}$:

S2. $S_{\mathbf{n}}\varphi \Rightarrow S_{\mathbf{n}}S_{\mathbf{n}}\varphi$

Consider any name $\mathbf{n}$ denoting a nonempty set of agents. It is easy to verify that either some agent named $\mathbf{n}$ knows ψ, or else some agent named $\mathbf{n}$ knows that that not everyone named $\mathbf{n}$ knows ψ. A slight generalization of this principle is expressed in the following axiom:

C3. $S_{\mathbf{n}}\varphi \Rightarrow S_{\mathbf{n}}(\varphi \wedge \psi) \vee S_{\mathbf{n}}(\varphi \wedge \neg E_{\mathbf{n}}\psi)$

We also get a number of other mixed introspection axioms that hold for various combinations of $S_{\mathbf{n}}$ and $E_{\mathbf{n}}$, such as:

- $\neg E_{\mathbf{n}}\varphi \Rightarrow S_{\mathbf{n}}\neg E_{\mathbf{n}}\varphi$
- $\neg S_{\mathbf{n}}\varphi \Rightarrow E_{\mathbf{n}}\neg E_{\mathbf{n}}\varphi$
- $E_{\mathbf{n}}\varphi \Rightarrow E_{\mathbf{n}}S_{\mathbf{n}}\varphi$.

Unfortunately, no simple combination of these axioms seems to be complete for structures where agents know their names. None of them seem quite to express a property which says, intuitively, every agent with name $\mathbf{n}$ either knows some fact φ, or else knows that not every agent with name $\mathbf{n}$ knows φ (for he, himself, does not). It seems very difficult to express this property directly in our restricted language. We can say that every agent with name $\mathbf{n}$ knows φ and that every agent with name $\mathbf{n}$ knows that some agent with name $\mathbf{n}$ does not know φ; however, we cannot quite say that every agent with name $\mathbf{n}$ knows either that *he himself* knows φ, or that some agent with name $\mathbf{n}$ does not know φ. Axiom C3 has some of this flavor, but it is not quite enough. We can come even closer with the following axiom.

C4. Let $\varphi_1, \ldots, \varphi_l$ be arbitrary formulas. Let $\psi_1, \ldots, \psi_{2^l}$ be all the formulas of the form $\varphi'_1 \wedge \varphi'_2 \wedge \ldots \wedge \varphi'_l$, where φ'_i is either φ_i or $\neg E_{\mathbf{n}}\varphi_i$. Then for any $A \subseteq \{1, 2, \ldots, 2^l\}$:
$$\bigwedge_{i \in A} \neg S_{\mathbf{n}}\psi_i \Rightarrow E_{\mathbf{n}}(\bigvee_{i \notin A} S_{\mathbf{n}}\psi_i)$$

Although C4 may appear somewhat complex, it is not too hard to see that it is sound. Since each agent with name $\mathbf{n}$ either knows φ_i or knows that it is not the case that everyone with $\mathbf{n}$ knows φ_i, it follows that each agent with name $\mathbf{n}$ knows at least one formula of the form ψ_j, $j = 1, \ldots, 2^l$. Suppose no one with name $\mathbf{n}$ knows ψ_i, for $i \in A$ (that is, $\bigwedge_{i \in A} \neg S_{\mathbf{n}}\psi_i$ holds). Thus everyone with name $\mathbf{n}$ knows that somebody with name $\mathbf{n}$ (namely it itself) knows ψ_i for some $i \notin A$. Thus, $E_{\mathbf{n}}(\bigvee_{i \notin A} S_{\mathbf{n}}\psi_i)$ holds. This relatively complicated axiom, together with all the others we have mentioned, turns out to be enough to characterize validity in $\mathcal{M}'_{\mathcal{N}}$, the class of models where agents know their names. The various mixed introspection axioms mentioned above can be easily shown to follow from these axioms. Let $AX'_{\mathcal{N}}$ be the result of adding S2, C3, and C4 to $AX_{\mathcal{N}}$.

Theorem 3.3: *The system $AX'_{\mathcal{N}}$ is sound and complete with respect to $\mathcal{M}'_{\mathcal{N}}$.*

Proof: See Appendix C. ∎

Although it is possible that we can get a complete axiomatization with an axiom simpler than C4, we conjecture that we cannot do much better. This suggests that the semantic condition we are trying to capture—knowing one's name—is very nearly beyond the expressive ability of the logics that we are considering.

We have noted that the original possible-worlds semantics (Section 2) is characterized by names which denote exactly one agent, and agents that know their name(s). It is interesting, but not unexpected, to see this reflected in the logic: when we add the axiom $E_{\mathbf{n}} \Leftrightarrow S_{\mathbf{n}}$ to $AX'_{\mathcal{N}}$, the result is equivalent to the axiomatic system S5. (We saw in the previous section that axiom $E_{\mathbf{n}} \Leftrightarrow S_{\mathbf{n}}$ corresponds to the situation where $\mathbf{n}$ refers to a single agent.) Our first system $AX_{\mathcal{N}}$ is strictly weaker than the logic S5, and can now say precisely what features (i.e., axioms) it lacks.

"Knowing one's names" has another interesting application. Earlier, when we introduced our operator $E_{\mathbf{n}}$, we noted that in [DM90, MT88] slightly different semantics are given to formulas of the form $E_{\mathbf{n}}\varphi$ for non-rigid names like $\mathbf{n}$. Intuitively, under the semantic conditions of [DM90, MT88], $E_{\mathbf{n}}\varphi$ is taken to mean that for all agents a

with name $\mathbf{n}$, agent a knows that *if* it has name $\mathbf{n}$, then φ holds. More formally, using $E^*_{\mathbf{n}}$ to distinguish this modality from the one we have defined, we have

$(M,w) \models E^*_{\mathbf{n}}\varphi$ if, for all $a \in \mu(w,\mathbf{n})$ and all w' with $(w,w') \in \mathcal{K}_a$ and $a \in \mu(w',\mathbf{n})$, we have $(M,w') \models \varphi$

The constraint "$a \in \mu(w',\mathbf{n})$" is what distinguishes $E^*_{\mathbf{n}}$ from $E_{\mathbf{n}}$. We could also define $S^*_{\mathbf{n}}$ as a variant of $S_{\mathbf{n}}$ by including this same constraint. The modal operator $E^*_{\mathbf{n}}$ is used in [DM90, MT88] rather than $E_{\mathbf{n}}$ for quite pragmatic reasons. They are interested in reasoning about correct processes that do not necessarily know that they are correct. But they do know that *if* they are correct, then they are bound to perform certain actions. Thus, it is becomes appropriate to make statements like "all the correct processes know that if they are correct then ...". This is exactly what the operator $E^*_{\mathbf{n}}$ lets us do. In the following, it will be useful to consider a mapping τ on the formulas in $\mathcal{L}_{\mathcal{N}}$ that replaces all occurrences of $E_{\mathbf{n}}$ by $E^*_{\mathbf{n}}$ and $S_{\mathbf{n}}$ by $S^*_{\mathbf{n}}$. We call the class of all formulas so obtainable (i.e., the range of τ) $\mathcal{L}^*_{\mathcal{N}}$.

Although there is no restriction made in [DM90, MT88] to structures where agents know their own name (indeed, it would be inappropriate to make this restriction, precisely because correct processes do not necessarily know they are correct), it should be clear that the modal operator $E^*_{\mathbf{n}}$ is somewhat related to the idea of agents knowing their names. For one thing, it is easy to see that $E_{\mathbf{n}}$ and $E^*_{\mathbf{n}}$ are equivalent if agents know their own names.

Lemma 3.4: *For all structures $M \in \mathcal{M}'_{\mathcal{N}}$ and $\varphi \in \mathcal{L}_{\mathcal{N}}$, we have $(M,w) \models \varphi$ iff $(M,w) \models \tau(\varphi)$.*

Moreover, in the language that we are considering, it turns out that the properties satisfied by $E^*_{\mathbf{n}}$ (and $S^*_{\mathbf{n}}$) in all structures are identical to the properties satisfied by $E_{\mathbf{n}}$ (and $S_{\mathbf{n}}$) in structures where agents know their names. More precisely, so long as we restrict attention to $\mathcal{L}^*_{\mathcal{N}}$, the following is true:

Theorem 3.5: *The axiom system $AX^*_{\mathcal{N}}$, obtained from $AX'_{\mathcal{N}}$ by replacing $E_{\mathbf{n}}$ and $S_{\mathbf{n}}$ by $E^*_{\mathbf{n}}$ and $S^*_{\mathbf{n}}$ everywhere, is sound and complete for the class of all structures $\mathcal{M}_{\mathcal{N}}$ (with respect to the language $\mathcal{L}^*_{\mathcal{N}}$).*

Proof (Outline only): Completeness is easy by the previous lemma, Theorem 3.3, and the observation that $\mathcal{M}'_{\mathcal{N}} \subseteq \mathcal{M}_{\mathcal{N}}$.

The key to the soundness proof is the construction, for any $M \in \mathcal{M}_{\mathcal{N}}$, of another structure $M' \in M'_{\mathcal{N}}$ which validates the same formulas. This can be done because, given the definition of $E^*_{\mathbf{n}}$ and $S^*_{\mathbf{n}}$, it never matters what worlds an agent with name $\mathbf{n}$ considers possible except for those other worlds where it also has name $\mathbf{n}$. This suggests that we redefine the knowledge of an agent so that it considers only such worlds as possibilities. The resulting model is in $\mathcal{M}'_{\mathcal{N}}$ but validates the same formulas. Formally, given $M = (W, A, \alpha, \mathcal{K}, \pi, \mu)$, let $M' = (W, A \times \mathcal{N}, \alpha', \mathcal{K}', \pi, \mu')$, where

$\alpha'(w) = \alpha(w) \times \mathcal{N}$, $\mathcal{K}'_{(a,\mathbf{n})} = (\mathcal{K}_a \cap \{w' : a \in \mu(w', \mathbf{n})\}^2)$, and $\mu'(w, \mathbf{n}) = \mu(w, \mathbf{n}) \times \{\mathbf{n}\}$. A straightforward argument by induction on the structure of formulas show that for all formulas $\varphi \in \mathcal{L}^*_{\mathcal{N}}$ and all worlds $w \in W$, we have $(M, w) \models \varphi$ iff $(M', w) \models \varphi$. ∎

This result is based on the observation that, for an agent a with name $\mathbf{n}$ at w, only the worlds that a considers possible *and where a has name* $\mathbf{n}$ are relevant to the evaluation of $E^*_{\mathbf{n}}$ and $S^*_{\mathbf{n}}$. But if the language was richer (in particular, had other modalities without this special property) the theorem and the construction on which it is based would fail. As a very simple illustration of this, note that if we had considered a language with both $E_{\mathbf{n}}$ and $E^*_{\mathbf{n}}$ together, then $E_{\mathbf{n}}\varphi \Leftrightarrow E^*_{\mathbf{n}}\varphi$ is clearly not sound in all of $\mathcal{M}_{\mathcal{N}}$ (only the left to right implication holds in general) although it is valid in all $M \in \mathcal{M}'_{\mathcal{N}}$. This shows that there really is a difference between the semantic requirement of "knowing one's name" and the alternative knowledge semantics of [DM90, MT88]. Nevertheless, these concepts are very similar and a reasonably rich language is required to demonstrate the distinction. We return to this issue later, where we give a more satisfying account of operators like $E^*_{\mathbf{n}}$.

4 Relative names and knowledge about self-identity

4.1 The problem

We have just seen how the logic $\mathrm{S5}_n$ can be relaxed, so as to deal with non-rigid names (names whose denotation is not common knowledge) and groups of agents. However, another issue arises which our logic does not seem to handle adequately.

Consider these examples:

- Suppose we wish to design a knowledge-based programming language [HF85]. It will contain commands of the form:

  `if <condition on computer's knowledge> then <perform action>.`

  We might hope to express the condition on the computer's knowledge using our previous logic; after all, it was intended to be expressive enough to represent knowledge. Unfortunately, it is not that easy. Unless we give each processor running this program a unique name, and then modify the program given to that processor to refer to this name, we will not be able to refer to that *particular* agent's knowledge. A far simpler solution to this problem is to be able to say "if *you* know φ" (or, from the point of view of the computer, "if *I* know ...").

- Processes are connected in a network, by a collection of point-to-point communication channels. This system is anonymous in that no global (commonly-known) naming scheme exists; each process can, however, distinguish among its various incoming and outgoing channels. At some point, each process "tells" the process at the other end of its first output channel some fact, say φ. What is the state of knowledge after this has occurred? Clearly it is something like "everyone knows that the process(es) on the end of their first channel knows φ"; it would be nice to express this as $E_{all}E_{\#1}\varphi$. (This notation, as well as this example, is based on [MR89, Rot89]. We assume the name *all* is interpreted as referring to every agent.) Unfortunately, we cannot give semantics to an operator such as $E_{\#1}$ in the framework discussed in the previous section. Which set of processes should $\#1$ denote in world w? Intuitively, the set of processes denoted by $\#1$ should be different in $K_1E_{\#1}\varphi$ and $K_2E_{\#1}\varphi$. In the first case it should be the processes at the end of p_1's first channel; in the second case, it should be those at the send of p_2's first channel. Clearly no choice that depends only on the world w will work.

- As a final example, consider a network of n processes that have just completed execution of a *leader-election* protocol, that is, a special protocol designed to select one of the processes to play some special role in subsequent computation. Further, suppose that the processors here are anonymous: all the non-leaders are in identical states (i.e., they are running the same program with the same input, and they do not have any unique identifier as part of their state). However, recall that in the semantics for distributed systems in Section 2, we assumed that all processes knew their name (i.e., we implicitly assumed that some unique identifier was part of its state). This assumption is clearly inappropriate. It is also easy to see that it leads to problems. For example, if agent p_i is not the leader then, using the original definition of the possibility relations in distributed systems, it follows that p_i considers possible just the worlds where it (i.e., p_i itself) is in a non-leader state. Thus, p_i knows that p_i is not the leader. However, according to these semantics, another non-leader, p_j, will *not* know that p_i isn't the leader. But this is surely wrong: in a truly anonymous system, p_i and p_j should have identical knowledge because they are in identical states. The semantics we have seen does not model this: rather, it models the situation that would arise if agents did indeed know what their names were.

  There seems to be a simple solution to this problem, which ensures that agents do not know their names. We can change the model so that an agent considers all those worlds possible where *anyone* is in his current local state. Then any two agents in the same state (such as p_i and p_j in the example) necessarily have the same knowledge. This is a improvement, but another more subtle difficulty arises. Part of our intuition about the example above is that p_i knows that he isn't the leader. But suppose, for definiteness, that the system described above includes a global clock, and a leader is chosen at time 5. Then at all the points at time 5 (i.e., all the points of the form $(r, 5)$), there will be one leader and a number of non-leaders. Since all the leaders are in the same state (say, a distinguished "leader" state) and all of the non-leaders are in the same state, the revised semantics we have just proposed has the property that every process, whether he is the leader or not, regards all time 5 points as being possible. Every process has the same knowledge! In particular, there can be no sense in this model in which "p_1 knows he is not the leader" holds (because then all others would know this as well, including the agent who *is* the leader). So we need to look harder to find appropriate semantics for systems like this. We want semantics that ascribes the same knowledge to agents who are in the same state, but lets us ascribe different knowledge to agents in different states.

The first example above illustrates the usefulness of being able to say *"I know".* Now I is not a non-rigid name in the same sense as, say, "the leader" or "the correct processes". The agent that it denotes depends not just on the world, but on the agent uttering the assertion. A similar phenomenon arises in the second example; in order to decide what agent(s) are denoted by #1, we must know which agent is making the utterance. It is easy to comes up with other examples of what we call *relative names;* names whose denotation is relative to the agent speaking.

The third example illustrates a different, but closely related, issue. Recall that in Section 3 we looked at a logic so general that an individual agent might not have a name at all; or else might be named but be ignorant about what this name is. But in that section, we quickly moved on to consider the more usual case where names are known by their owners.

But the third example shows that there is more to be said about the former, anonymous agents, type of situation. First, it reminds us that the original semantics for knowledge in distributed systems is simply inappropriate for these applications. Two agents with identical local states could be ascribed different knowledge, but this is incompatible with our understanding of how *knowledge* should work. We have already seen the explanation for this: the original semantics (implicitly) assumed that each local state has a hidden component, which is some identifier unique to that agent. So this assumes agents' states are never *truly* identical. But, as we saw in the example, formulating semantics that do not have such an assumption built in is not trivial.

The real problem is this: up to now, we have modeled uncertainty by describing which worlds the agent considered possible. The source of the uncertainty was lack of knowledge about what the actual world is like. Thus, the fewer worlds an agent considers possible, the greater its knowledge about the actual world. But in the example, the agent has another source of uncertainty. The agent knows perfectly well what the world looks like: there is a leader in a distinguished "leader" state, and there are other processes in a special "non-leader" state. Since the system is anonymous, if he is not the leader, then it does not knows which agent he is (he only knows that he is not the leader). In fact, even if he *is* the leader, it doesn't know which agent he is (it knows that he is the leader, but the identity of the leader might vary from world to world). It is such uncertainty—not about what the world is actually like, but about who the agent is in the world—that explains why we think that the leader and the others have different knowledge.

4.2 The solution

In fact, relative names and *knowledge about self-identity*—essentially, knowing who you are—can be dealt with the same way. Up to now we have taken $\mathcal{K}$, for any agent, to be a binary relation on worlds. But we have just seen the problem with this: two agents can consider the same set of worlds possible (if we view a world as a description of what the system could be like, as we have done in our distributed system model) and still have different knowledge, because of their uncertainty about who they are in these worlds. In order to capture this uncertainty, we modify $\mathcal{K}$ so that it becomes a relation on (world, agent) *pairs*. We then interpret $((w, a), (w', a')) \in \mathcal{K}$ as saying that in world w, agent a thinks it might be a' in w'.

Essentially, by moving to these pairs, we have augmented the notion of possible world so that it explicitly includes the agent from whose viewpoint everything is observed. After all, any assertion has to be made *by someone*. So formulas are not just claims about a world, but really about a world and an agent (the speaker). What we are doing in looking at pairs like (w, a) is explicitly recognizing this.

The first dividend is that relative names, like #1 and I are easy to interpret: at (w, a) they are interpreted relative to a. In particular, I just denotes the agent making the assertion, a.

The new definition of $\mathcal{K}$ also neatly captures an agent's knowledge (or ignorance) about who he is. The agent doesn't just consider other *worlds* possible, but also considers *who it might be* within these other worlds. Two agents could agree perfectly well on what the world is like objectively (the leader and a non-leader in the previous example would agree on this) but differ about who they think they are in these worlds (the non-leader knows, that, in any world that is possible, he is not the agent in the "leader" local state).

Now, let us look at this formally. We take a *possible-worlds structure with knowledge about self-identity* over Φ and $\mathcal{N}$ to be $M = (W, A, \alpha, \mathcal{K}, \pi, \mu)$. W, a set of *worlds*, A, a set of *agents*, and $\alpha : W \rightarrow \mathcal{P}(A)$ are as before. The other components of M are more interesting.

The relation $\mathcal{K}$ and its intuitive interpretation was described above. We repeat the key point: it relates (world, agent) *pairs*, so that given a world w and an agent a, $\mathcal{K}$ will serve to determine the collection of such pairs that a considers possible from w. For the same reasons as before, we require that when $((w, a), (w', a')) \in \mathcal{K}$, then $a \in \alpha(w)$ and $a' \in \alpha(w')$. And, as for our earlier logic, we are going to require that $\mathcal{K}$ be reflexive, symmetric, and transitive. The motivation here is the same; in our application to distributed systems we say that a in w considers that it might be a' in w', just in case the local states of a and a' (in w and w' respectively) are identical. Any $\mathcal{K}$ derived in this way will be an equivalence relation, so we restrict attention to such relations.

Now that formulas are evaluated at (world, agent) pairs, we have more freedom in the definition of π and μ. For example, it turns out to be useful to allow the truth of a primitive proposition in Φ to depend not just on the world (as before), but also on the agent we are considering as a "viewpoint" in the world. For the example of processors in a ring, a proposition *leader* would naturally be defined to be true at (w, a) just when a is in fact the leader in possible world w. Another example would be a proposition $x_I = 1$ with the intended interpretation that my (i.e., the speaker's) local variable x has value 1. Note that this makes sense (and an agent might know whether it is true) even if the agent doesn't know its own name.

Let us call any proposition interpreted in this way *relative* (because it is relative to an agent also). For these propositions, we must redefine π; now $\pi : W \rightarrow (A \rightarrow (\Phi \rightarrow \{\mathtt{true}, \mathtt{false}\}))$.[6] So $\pi(w)$ does not give a truth assignment on Φ immediately; for that we need to specify an agent as well. In the following, we will often write $\pi(w)(a)$ (which *is* a truth assignment on Φ) as $\pi(w, a)$.

In fact, we can interpret all propositions in this more general way. Those primitive propositions whose truth depends only on the possible world (we call these propositions *absolute*) are modeled simply through π being independent of the second (agent) argument. That is, for such p, $\pi(w, a)(p) = \pi(w, a')(p)$ for $a, a' \in A_w$. So absolute propositions can formally be regarded as special cases of relative propositions.

The truth condition for propositions now becomes:

$w, a \models p$ for p in Φ if $\pi(w, a)(p) = \mathtt{true}$.

We can analyze names in a similar way. Names should still, ultimately, refer to a set of agents, but which set can now depend both on the world, and an agent in this world. Formally, we achieve this generality by regarding μ as mapping a world to a binary relation on the agents in that world (rather than just a set of agents, as before). Then name $\mathbf{n}$ at pair (w, a) is taken to refer to the collection of agents which stand in the relation $\mu(w, \mathbf{n})$ to the agent a. This model is useful, because binary relations on agents arise frequently and naturally. Consider the second example in this section, the network of processors linked by numbered communication lines. A channel name like #1 could be easily modeled as a relation on agents where $(a, b) \in \mu(w, \#1)$ if, in world w, agent b is at the end of a's first channel. Many other similar examples of such *relative* names can be found.

In our logic, these relative names are interpreted relative to *oneself* (i.e, relative to the agent of evaluation; the agent whose knowledge we are reasoning about.) For example, if a processor sends a message along his first outgoing line, he may know that every agent on the end of #1 *relative to him* will soon receive it. In the formal semantics the agent a in (w, a) is taken to be the (implicit) reference point for all relative names. This is reflected in the truth conditions:

$w, a \models E_{\mathbf{n}}\varphi$ if, for all b with $(a, b) \in \mu(w, \mathbf{n})$, for all w', b' such that $((w, b), (w', b')) \in \mathcal{K}$, we have $w', b' \models \varphi$.

[6] $\pi(w)(a)$ need only be defined for $a \in A_w$.

$w, a \models S_{\mathbf{n}}\varphi$ if, for some b with $(a, b) \in \mu(w, \mathbf{n})$, for all w', b' such that $((w, b), (w', b')) \in \mathcal{K}$, we have $w', b' \models \varphi$.

We have two further comments to make about relative names. First, just as with propositions, we can retain the notion of *absolute* names (the set of agents referred to depends only on the world), but we can regard these as simply special cases of relative names (formally, where $(a, b) \in \mu(w, \mathbf{n})$ if and only if $(a', b) \in \mu(w, \mathbf{n})$, for all $a, a', b \in A_w$). That is, the reference point (first agent in the pair) is irrelevant. So we have not lost anything by assuming that all names denote relations on agents. Second, we assume that there is one special name, which we call I, that agents use to refer to themselves directly. Formally, in our logic the name I always has a fixed denotation: it denotes the *identity* relation (on agents in a world). This restriction on the logic is useful because, since $(a, a') \in \mu(w, I)$ just if $a = a'$, the name I will always end up referring to exactly one agent (the agent whose knowledge we are reasoning about). So the symbol I provides a way for an agent to directly refer to *itself.*

We now briefly review the three examples of the introduction and confirm that our logic has sufficient expressive power and semantic flexibility to deal with them. The first example is solved, because we now have a formal theory that allows for the symbol I. We can write `if <I know` φ`> then <perform action>`, where the condition is indeed a formula in our logic. The network example is similarly easy: $E_{all}E_{\#1}\varphi$ has the semantic interpretation "everyone knows that everyone on the end of his own first channel knows φ" as we would wish. By moving to (world, agent) pairs we ensure that the #1 is always evaluated relative to the appropriate agent. Finally, we can see that the semantic difficulties apparent in the last case are overcome: even when all processors consider the same world(s) possible, only the agent who actually *is* the leader will consider that he might be the leader. This distinction now receives recognition in our formal model.

It may be helpful to look at this last example in more detail. Recall that the situation consists of a ring of processors where just one (the *leader*) is in some special, distinguished state (and everyone knows that there must be such a processor).

To be more concrete, let us assume there are three machines and that everyone knows this. These are identified using some characteristic (such as their location) as X, Y, Z. Each machine has a local state consisting of the contents of all storage it can access and the readings on all sensors and input devices available to it. Two of the three are in local state s_0, while the other is a different, *leader,* state s_1. While this is an extremely simple picture, it could certainly arise in practice. Our problem is to find logical theories adequate to express some of the properties that the system could possess. Here, we look at how possible-world semantics with knowledge about self-identity helps us achieve this.

In the distributed systems model, every possible world w is associated with a global state (a function from agents to local states) like $\{(X, s_0), (Y, s_1), (Z, s_0)\}$. For such w we take a relative proposition *leader* to be true only for the agent in s_1 at that world.

To make the model slightly richer, we also suppose there are two relative names *left* and *right*. For example, $(X, Y) \in \mu(w, \mathit{left})$ if X is one position counterclockwise in the ring, from Y. Finally, there are names for each of the three agents; say $\mathbf{x}, \mathbf{y}, \mathbf{z}$.

Given just this vocabulary of names and propositions, we need six possible worlds, because there are three different choices for the leader, and two ways of orienting the ring. We will call these worlds $w^1_X, w^2_X, w^1_Y, w^2_Y, w^1_Z, w^2_Z$ (where the superscript is 1 if the orientation is clockwise, otherwise 2; the subscript indicates who is the leader). Note how the possible worlds correspond directly to the description of the system given above.

The final semantic component is the knowledge relation $\mathcal{K}$. Recall that in general an agent a (in w) considers another world-agent pair (w', a') possible just in case the state of a' in w' is the same as that of a in w. For example, $\mathcal{K}$ contains $((w^1_X, X), (w^2_Y, Y))$ because X in w^1_X and Y in w^2_Y are both in state s_1. Because of this, X cannot distinguish the situation where he is X and the world is w^1_X from that where he is in fact Y in w^2_Y. It is not hard to see that $\mathcal{K}$ is an equivalence relation on the collection of pairs, with just two equivalence classes: the pairs (w, a) where agent a in w is in s_0, and the class of pairs where a is in s_1.

What can we express in our logic? We want simple assertions about the system to receive a direct translation as logical formulas.

In w^1_X, X knows that he is the leader. That is, the sentence $K_{\mathbf{x}} \mathit{leader}$ (equivalently, $K_{\mathbf{x}} K_I \mathit{leader}$) holds.[7] This is as it should be: X can simply examine its state to discover this. As a second example, $K_{\mathbf{x}} K_{\mathit{left}} \neg \mathit{leader}$ also holds at w^1_X. This is because X knows there is more than one process. If we had modeled the situation where X did not know the size of the ring, and considered it possible that he was the only node present, this formula would be false because X would think it possible that he is his own left neighbor.

Even though X knows that *it* is the leader, it does not actually know that X is the leader. That is, $w^1_X, X \not\models K_{\mathbf{x}} K_{\mathbf{x}} \mathit{leader}$. This also is expected; it arises because X does not know that he is X. If someone were to tell X, in w^1_X, that X was in fact the leader, it really would have gained information and should then consider fewer alternatives possible, as the model here predicts.

While all the examples we have seen are, admittedly, trivial, it remains true that the semantic concepts of *relative names* and *knowledge about self-identity* play a large role in the simplicity and directness of our solutions. Difficulties arise in theories without these features because there is an inherent conflict between two assumptions frequently made about possible-worlds models of knowledge. First, often our intuitive idea of "possible world" is that such a world is some objective model of the way things really are; objective, in the sense that this world does not depend on the particular agent considering it. This is quite clearly the situation in the distributed systems model, where the structure of a global state is just a set of agents and a function from these

[7] Strictly speaking, sentences are true of world-agent pairs rather than just worlds in our logic. But when, as here, the truth is independent of the *agent* component we omit mention of it.

agents to local states. This intuition is useful because we feel there are some aspects of the real world that are truly independent of the observer. However, this conflicts with a wish to model knowledge as "truth in all possible worlds". Consider again the example of a leader-election protocol on a ring (Section 4.1). There it is apparent that no process in the ring can rule out any time 5 point as being impossible, and so such a notion of knowledge is too weak (it does not even distinguish between the leader and the others). Our response to this conflict to retain the objectivity of worlds, but model knowledge as truth in all the (world, agent) pairs an agent considers possible. The intuition behind the second component of these pairs—who an agent thinks he might be—is plausible and, as we have just seen, leads to an effective theory.

4.3 Properties of our logic

In this section, we look at axioms that are appropriate for the language and semantics just presented. It will become apparent that the name I plays a special role in our logic. This is a reflection of the special status of I in the semantics, where it is given a fixed interpretation as the identity relation on agents in a world. One consequence of this is that, at any pair (w, a), the name I will refer to just a single agent (a). It follows that the axiom

K1. $E_I\varphi \Leftrightarrow S_I\varphi$

is sound. (We saw in Section 3.3 that this was sound for names which refer to exactly one agent.) We can use this special property of the name I to simplify our notation: in the following, we will write K_I rather that E_I or S_I.

In Section 2 we saw that the logic S5 has so-called introspection axioms A4 ($K_{\mathbf{n}}\varphi \Rightarrow K_{\mathbf{n}}K_{\mathbf{n}}\varphi$) and A5 ($\neg K_{\mathbf{n}}\varphi \Rightarrow K_{\mathbf{n}}\neg K_{\mathbf{n}}\varphi$). Under the more general semantics of Section 3 neither of these axioms is sound, because even when there is only one agent with name $\mathbf{n}$ it is not necessarily the case that this agent knows he has this name.[8] This is a slightly curious situation, because the usual intuitive explanation of the axioms—agents know what they know and know about what they do not know—is so appealing. We can understand this issue by noting that axioms A4 and A5 are not simply statements about introspection at all, because introspection is concerned with reflection on one's *own* knowledge. Using the name I and the semantics for knowledge about self-identity we can now do better, because we can speak *directly* of such knowledge. The formulas $K_{\mathbf{n}}\varphi \Rightarrow K_{\mathbf{n}}K_I\varphi$ and $\neg K_{\mathbf{n}}\varphi \Rightarrow K_{\mathbf{n}}\neg K_I\varphi$ seem to describe introspection more accurately. When we generalize these to take account of names that denote groups of agents, we get:

K2. $S_{\mathbf{n}}\varphi \Rightarrow S_{\mathbf{n}}(\varphi \wedge K_I\varphi)$

K3. $\neg S_{\mathbf{n}}\varphi \Rightarrow E_{\mathbf{n}}\neg K_I\varphi$

[8] In essence, this observation goes back to Hintikka [Hin62]. See also [Les89].

These are easily seen to be sound for our semantics, whether or not agents know their names.

This generalizes A4 and A5, so what about the other axioms of the system $S5_n$, A2 $(K_{\mathbf{n}}\varphi \wedge K_{\mathbf{n}}(\varphi \Rightarrow \psi) \Rightarrow K_{\mathbf{n}}\psi)$ and A3 $(K_{\mathbf{n}}\varphi \Rightarrow \varphi)$? Axiom A2 follows from the more general E1. And, in the case where $\mathbf{n}$ is I, A3 is also sound for our semantics:

K4. $K_I\varphi \Rightarrow \varphi$.

However, names other than I behave differently because our semantics allows φ to be interpreted relative to an agent. For example, we would write *someone in* $\mathbf{n}$ *knows that he is correct* as $S_{\mathbf{n}}$*correct*. Here, *correct* is a (relative) proposition. But just because $S_{\mathbf{n}}$*correct* is true at some pair w, a does not guarantee that $w, a \models$ *correct*. After all, a may not be one of the agents with name $\mathbf{n}$, and even if it is, it is not necessarily a *correct* agent. So $S_{\mathbf{n}}$*correct* $\Rightarrow$ *correct* is not valid; we do not get an analogue of the S1 axiom introduced in Section 3.3. In general, all we can say is that no one knows falsity:

S1′. $\neg S_{\mathbf{n}}$*false*.

Let the axiom system $A\mathcal{X}_{\mathcal{N}}^{\mathrm{ksi}}$ consist of A1, R1, E1, E2, C1, C2, as well as the new axioms S1′, K1, K2, K3, K4. Then:

Theorem 4.1: *The system* $A\mathcal{X}_{\mathcal{N}}^{\mathrm{ksi}}$ *is sound and complete with respect to the class of all possible-worlds structures with knowledge about self-identity.*

Proof: See Appendix D. ∎

Furthermore, the following result can be shown, using techniques from [HM85]:

Theorem 4.2: *The problem of deciding if a formula is valid with respect to the class of all possible-worlds structures with knowledge about self-identity is* PSPACE*-complete.*

Proof: See Appendix E for some discussion of the proof of this theorem. ∎

One interesting special case of our logic is where the only name present in the language is I. Axiom E1 reduces to $K_I\varphi \wedge K_I(\varphi \Rightarrow \psi) \Rightarrow K_I\psi$, and K2, K3, K4 reduce to A3, A4, A5 respectively. Other axioms are subsumed by these. That is, the logic which is sound and complete for reasoning about the name I is especially simple—it is just the logic S5 (i.e., $S5_1$).

As in Section 3, it is useful to see formally how this generalizes our previous work, and how the earlier logic (of Section 3.3) can be recovered. Showing how this is done casts further light on the axioms $A\mathcal{X}_{\mathcal{N}}^{\mathrm{ksi}}$.

Our first observation along these lines is the following. It turns out that, if we were to retain the semantics with knowledge about self-identity (including relative names and propositions) but delete the name I (with its special semantics) from the logic, then a sound and complete axiomatization is obtained by discarding K1, K2, K3, K4. Although this result is not completely obvious, it is intuitive: the axioms K1, K2, K3, and K4 talk about properties of the operator K_I and are required only if I is being used. However, even without the name I, this logic is still not the same as that in Section 3.3: we have seen that axiom S1 is not valid, and only the weaker S1$'$ holds.

Suppose, however, that some names and propositions are believed to be *absolute*, in the sense discussed in 4.2. That is, we believe that the semantic interpretation of these names and propositions should be independent of the identity of the agent we are reasoning about. It is possible to reflect this semantic restriction axiomatically, as follows. Let us call an *objective* formula any absolute primitive proposition, $E_{\mathbf{n}}\varphi$ or $S_{\mathbf{n}}\varphi$ where $\mathbf{n}$ is an absolute name, or any boolean combination of objective sentences. The truth of objective formulas depend only on the world (and not on the agent). It turns out that the axiom

S1$''$ $S_{\mathbf{n}}\varphi \Rightarrow \varphi$ *for objective* φ

is sound. And adding this to $A\mathcal{X}_{\mathcal{N}}^{\mathrm{ksi}}$ gives a logic which is complete for the case when some symbols are absolute.

When we do not force any symbols to be absolute, the only objective sentences are, essentially, *true* and *false*. In this case, S1$''$ reduces to S1$'$, which is exactly what we saw in the theorem. On the other hand, suppose we regard all propositions and names as absolute. Then *every* formula is objective, and S1$''$ reduces to S1. In this case we recover the logic of 3.3. This is not surprising, because none of the additional semantic structure we have adopted (such as $\mathcal{K}$ being a relation on pairs and allowing μ to denote relations on agents) plays any role when all sentences are objective.

4.4 Names and propositions: extending the logic

One aspect of the semantics we have presented is that the distinction between propositions and names seems somewhat artificial. We could regard a relative proposition $p \in \Phi$ as denoting a set of agents in each world w (those agents a such that p is true at (w, a)), but this is semantically the same as an absolute name. Both are really just sets of agents in a world; names and proposition symbols are just the syntactic way we refer to them.

A very good illustration of this point is the work of [DM90, MT88]. Recall they were interested in the collection of nonfaulty processors (denoted *correct*) in a network, and required a definition of $E^*_{correct}\varphi$ which (informally) was *all nonfaulty processors know that, if they are nonfaulty, then* φ. In our current logic, we can try to translate this simply as $E_{correct}(correct \Rightarrow \varphi)$, where *correct* is regarded both as a relative proposition and an (absolute) name.

This translation is appealingly direct, but there is a problem. So far, we have implicitly assumed that set of names was disjoint from the atomic propositions. So suppose we extend the language, so that a symbol (like *correct*) can be used either way. This is still not sufficient, because we have to make the semantic connection between a symbol used both as a name and as a proposition. Intuitively, we want it to refer to the same set of agents in each world. Formally, we need to link how π and μ interpret $\mathbf{n}$:

$\pi(w, a)(\mathbf{n}) = \mathit{true}$ iff $(a, a) \in \mu(w, \mathbf{n})$.

To begin with, we restrict attention to absolute names only. Then this condition simply states that $\mathbf{n}$ as a proposition is true at (w, a) just when a is in the range of the interpretation of $\mathbf{n}$ as a name. When this semantic condition holds for all symbols used both as names and propositions, our previous axiomatization is no longer complete and we must add:

N1. $\mathbf{n} \wedge K_I\varphi \Rightarrow S_{\mathbf{n}}\varphi$

N2. $S_{\mathbf{n}}(\mathbf{n} \Rightarrow \psi) \Rightarrow \psi$ for all *objective* formulas ψ.

(Since $S_{\mathbf{n}}\neg\mathbf{n}$ is equivalent to $S_{\mathbf{n}}(\mathbf{n} \Rightarrow \mathit{false})$, N2 can be regarded as a stronger version of the principle that someone in $\mathbf{n}$ can never know $\neg\mathbf{n}$.)

Theorem 4.3: *The system $A\mathcal{X}_{\mathcal{N}}^{\mathrm{ksi}}$ together with axioms N1 and N2 is sound and complete with respect to the class of all possible-worlds structures with knowledge about self-identity in which the absolute name* $\mathbf{n}$ *satisfies the semantic condition given above.*

Proof: See after the proof of Theorem 4.1 in Appendix D. ∎

As an application of this, when $\mathbf{n} = \mathit{correct}$, we now have a complete axiomatization of a logic which can express the modalities defined in [DM90, MT88]. This has emerged as a particular case of our more general theory.

The motivation for this extension to the logic was the observation that propositions and absolute names both "really" refer to sets of agents. Indeed, the worlds in our possible-worlds models now have a lot of structure: agents, sets of agents, relations on agents, and so on. It would seem that such a complex entity is most completely described in a first-order language (at least; arguments could be made for even higher-order logic). However, in this paper we look at propositional logics only because of their simplicity. As we have seen, there are expressive logics even in this framework. But it is true that we sometimes need more. In such cases, there is an alternative to adopting a full quantified logic. Instead, we can extend the propositional languages in various ways to include whatever features we require. The result we just have seen is partially in this spirit. When we take the concept of "set of agents" seriously, it is natural that we should not be restricted as to whether we refer to it by a name (subscript of E, S) or as a proposition. In the remainder of this section, we look at several other ways in which the logic could be extended. Each addition will, in some respect, make the logic look less propositional: this is the price we pay for greater expressive power. Because there are so many variations, we do not present axiomatizations for every one of the suggested ideas.

We continue by making the observation that the languages we have seen so far are quite limited as to how we refer to sets of agents: we can only refer to collections named by a symbol in $\mathcal{N}$ or through a relative proposition. But suppose we wish to describe properties of all the processes that are correct *and* received my last message. A robot might want to reason about all agents that are *not* in the same room as him. It might be the case that *everyone who doesn't know* p knows q. How do we deal with these?

The last example suggests that we may want to talk of about all agents with a certain property (i.e., that satisfy a particular formula). In the language, we want to allow arbitrary formulas as subscript to E and S. Of course, in the language $\mathcal{L}_{\mathcal{N}}$ we have considered until now, only names can appear there. Nevertheless, suppose the language was extended in this way, so that E_ψ and S_ψ are permitted modal operators, for any formula ψ. It is not hard to give appropriate truth conditions, such as:

$w, a \models E_\psi \varphi$ if, for every b with $w, b \models \psi$, for all w', b' such that $((w, b), (w', b')) \in \mathcal{K}$, we have $w', b' \models \varphi$.

Given this addition to the language, we then have the desired expressive power. (In fact, it is not hard to give an axiomatization for the new logic. The new axioms required are variants of N1 and N2, where formula ψ replaces occurrences of $\mathbf{n}$. This works because the extension to the language which would allow E_ψ and S_ψ is not as powerful as it might appear. An equivalent effect is achieved by including a new name, say $\mathbf{p}$, adding $\mathbf{p} \Leftrightarrow \psi$ as an axiom, and using $E_{\mathbf{p}}$ instead of E_ψ everywhere.)

The other examples suggested above concerned the combination of existing names. We can add operators $\cup$, $\cap$, $^-$ to the language to express union, intersection, and complement. Then if $\mathbf{n}$ and $\mathbf{m}$ are two (absolute) names, then so are $\mathbf{n} \cup \mathbf{m}$, $\mathbf{n} \cap \mathbf{m}$ and $\bar{\mathbf{n}}$. The appropriate semantic conditions for these new names should be clear. As before, we can easily give an axiomatization for the logic corresponding to this larger language: simply add $\bar{\mathbf{n}} \Leftrightarrow \neg\mathbf{n}$, $\mathbf{n} \cup \mathbf{m} \Leftrightarrow \mathbf{n} \vee \mathbf{m}$, and $\mathbf{n} \cap \mathbf{m} \Leftrightarrow \mathbf{n} \wedge \mathbf{m}$, as well as N1, N2, to the axiom system $A\mathcal{X}_{\mathcal{N}}^{\mathbf{ksi}}$.

Until now, we have concentrated on names which are (semantically) absolute, because our intuitions are clearer here. But what we have done for these names can be done for the general, relative, case.

First, a correspondence between names and propositions can be extended to this situation. If $\mathbf{n}$ is any name, we can certainly extend the language to allow it to be used as a proposition. The appropriate consistency condition is the same as that which we presented earlier for absolute names. The intuition is that an occurrence of a name $\mathbf{n}$ in the syntactic position of a proposition, is read as "Am I in $\mathbf{n}$?". Formally, we ask at (w, a) whether (a, a) is part of $\mu(w, \mathbf{n})$. This is exactly the same intuition as we had previously, but because names are now relative, the name $\mathbf{n}$ will be interpreted relative to "me" anyway. So what is really being asked is "Am I myself one of the agents that I call $\mathbf{n}$?". Clearly, the ability to use names as propositions is far more intuitive and useful in the case of absolute names (but for some examples using relative names as propositions see [Gro]).

Second, we can explore the idea of combining names further. In the case of absolute names, we can combine names using the basic set-theoretic operations, as discussed above. However, in the case of relative names it also makes sense to look at the (relational) *composition* of two names. We can add a composition operator $\circ$ to the language, so that given names $\mathbf{n}$ and $\mathbf{m}$ we can form a new name $\mathbf{n} \circ \mathbf{m}$. Semantically, we want $(a, b) \in \mu(w, \mathbf{n} \circ \mathbf{m})$ just if there is $a' \in A_w$ with $(a, a') \in \mu(w, \mathbf{m})$ and $(a', b) \in \mu(w, \mathbf{n})$. As an example, if I broadcast a message to all my #1 neighbors, knowing that they will do the same, then eventually $E_{\#1 \circ \#1}\varphi$.[9]

Even with all this, there is a one large and natural class of statements that we still cannot express directly. We would like to do more than combine names; sometimes we want to speak directly of the (set-theoretic) relationships between them: "The class of working machines is disjoint from the class of machines in this building", "Everyone in this room is a student", "I am the only correct process", and so on.

Formally, we could achieve this with one further addition to our language: a symbol $\subseteq$ denoting inclusion. For propositions $\mathbf{m}, \mathbf{n}$:

$$w, a \models \mathbf{m} \subseteq \mathbf{n} \text{ iff } \{b : (a, b) \in \mu(w, \mathbf{m})\} \subseteq \{b : (a, b) \in \mu(w, \mathbf{n})\}.$$

It is clear that the resulting language is extremely expressive. For example, our previous suggestions about treating names as propositions is subsumed by this: $I \subseteq \mathbf{n}$ ("Am I in $\mathbf{n}$?") has the same semantics that we gave to $\mathbf{n}$ when viewing it as a proposition. Another important application of $\subseteq$ is that it can be used to test whether two names denote the same set of agents (for we can define $\mathbf{m} = \mathbf{n}$ as $\mathbf{m} \subseteq \mathbf{n} \wedge \mathbf{n} \subseteq \mathbf{m}$).

We do not have a sound and complete axiomatization for the logic with composition and inclusion, although this might be useful. Once one begins extending the basic logic this far, there are clearly numerous other possibilities and variations to consider as well. A point is soon reached at which additional expressive power is more appropriately achieved by abandoning the propositional framework altogether, and including first-order quantification in the language. This is a complex issue, which we investigate in part II. The point of this section is there is a large range of logics lying between our basic propositional theory on one hand, and a full first-order logic on the other. Many applications may require only a few, limited, special features and in such cases it is sensible to find the simplest logic that is adequate.

4.5 Comparisons with other work

The key to our logic is the proposal that knowledge should be regarded as a relation between pairs which consist of a world *and* a viewpoint (agent) in that world. Essentially equivalent semantics were described by Lewis [Lew79]. [Cas68] and [Per79] also contain good arguments for the main conclusion, that knowledge about oneself, and "indexical" pronouns like *I* or *he*, really are special. A good summary of the associated philosophical debate can be found in Lespérance's thesis [Les91].

[9] It is not necessarily the case that $E_{\#1}E_{\#1}\varphi$; perhaps the first processors in the relay forget the message after forwarding it.

Lewis's work is most relevant to ours, for he explicitly advances the principle that the objects of knowledge extend beyond *propositions* (by which he means sets of possible worlds) to *properties* (sets of individuals, or, equivalently—as his individuals must belong to just one particular world—sets of pairs of a world, and an individual in that world). In our terminology, Lewis is arguing that agents consider (world, agent) *pairs* to be possible, rather than just collections of possible worlds alone. But, unlike our work, Lewis does not provide a formal system to accompany his philosophical arguments. This is, of course, one of the principal contributions we make in this paper.

Lewis's idea has also been adopted by Lespérance [Les89, Les91]. This work applied the idea of *de se* knowledge (this is the term Lespérance and Lewis both use for, essentially, what we are calling knowledge about self-identity) to an analysis of what it means to say an agent *can* do an action. Based on the earlier [Moo85], Lespérance's work includes a good argument for why knowledge about the world relative to oneself permits a more accurate and effective description of how agents behave. Like us, Lespérance develops a formal logic. But, unlike his work, the logic we have looked at in this section is propositional and, furthermore, we have been able to give completeness and complexity results. Propositional S5 (for one or many agents) is a simple, elegant, and well understood logic, and we have set out to show that some quite small modifications are sufficient to deal with non-rigid names and anonymity. As the results of this and the previous sections demonstrate, we can do this to a considerable extent. Certainly, knowledge about self-identity can be incorporated within the propositional framework. In part II we look at a richer first-order logic which is more directly comparable with Lespérance's work, and we discuss this at greater length there.

We also note that there are other logics with a characteristic similar to knowledge about self-identity, in that they supplement the "possible worlds" where formulas are evaluated by an explicit *context* referring to where, when, or by whom an utterance is made. One well-known example is Kaplan's work on demonstratives and indexicals.[10] Perhaps the less important distinction between [Kap89] and our work is that the former does not address epistemic modalities, so does not produce anything resembling a logic of knowledge that can deal with indexical terms. What is more important is that philosophical investigations, like Kaplan's, appear almost entirely concerned with formalizing human reasoning and natural language. The principal question is to ask what the words used in such reasoning "really mean". We emphasize that our work is not intended, in any way, to address these issues. We seek formal logics that are simple, unambiguous, and sufficiently expressive to allow us to talk about the states of simple "agents" in terms of some concept roughly understandable as "knowledge". A more detailed discussion of this difference can be found in part II, because the first-order logic we develop there is more directly comparable with existing philosophical accounts of naming and reference.

[10] *Indexical* is a term frequently used to describe objects, like names, that are relative to one agent's perspective. We have used the word "relative" in preference in "indexical" in this paper. In part this is to avoid confusion; in [DM90, MT88] the word indexical is instead used as a synonym for non-rigid.

Some other recent work relevant to our framework is [MR89, Rot89], which outlines a propositional epistemic logic which includes relative names like #1 and I, specifically intended for an application to distributed computing (investigation of message diffusion in anonymous systems). Because they had such a specific application in mind, Moses and Roth do not look at a more general language or develop a general semantic theory.

An interesting recent paper is Seager's [Sea90]. Seager shares our interest in multiple agent logics of knowledge and belief (he concentrates on the latter). He develops a fairly simple and essentially propositional logic that includes indexicality (in the sense of our knowledge about self-identity). Furthermore, his basic framework is modal logic and possible-worlds semantics. Nevertheless, his theory ends up being very different from ours. Simplifying matters significantly, we note that Seager uses a set of basic (singular) names, say $t, u, v, \ldots$ and an associated set of "quasi-indexicals" $het, heu, hev \ldots$.[11] These latter are similar in purpose to our single name "I"; Seager motivates them as "secret names" individuals refer to themselves with, that are tied to action and perception. This approach seems unnecessarily complex. Most importantly, we do not understand what it means for one individual to reason about the quasi-indexical associated with another. To look at a very simple example of this, both $B_t(u = heu)$ and $B_t(u \neq heu)$ are satisfiable formulas in Seager's logic. That is, t is allowed to have an opinion on whether u, and u's secret name for himself, actually denote the same object or not. We do not see how such an opinion could arise, or even what it really means. This is an important example, because if such applications of multiple quasi-indexicals really are significant then this would be a major gap in our own logic, and show the existence of a deep flaw in our analysis. However, it turns out that none of Seager's motivating examples actually depend on these multiple quasi-indexicals; they can all be captured easily in our logic as well (the basic reason for this is that, in Seager's examples, agents only really need to reason about their "own" quasi-indexical). Furthermore, although Seager's possible-world semantics are considerably complicated by the presence of multiple quasi-indexicals, he does not give enough motivation for some of the most important definitions used.[12] So we are unable to find any good justification for multiple quasi-indexicals within Seager's technical results either. We believe that our logic is substantially simpler than Seager's, and yet because of features such as group names, relative names, and the use of indexicals as subscripts on epistemic operators, far more powerful.

Finally, we should contrast the logic we have just introduced with that of Section 3. The formal distinction between the two, and how the later logic reduces to the earlier

[11] The term "quasi-indexical" is due to Castañeda [Cas68]. Castañeda uses it to draw a distinction between words such as the English "I", that refer to the speaker, and more indirect usages that are better better read as, say, "he, himself" (the latter are the quasi-indexicals). This distinction is certainly important if we are trying to understand natural English usage, but within our logic it does not correspond to any interesting technical or semantic classification.

[12] In particular, the definition of "indirectly t-accessible" in Seager's paper, which is rather complex and one of the most radical departures from standard possible-worlds semantics for belief, surely needs more explanation.

in special cases, was examined in 4.3. Here, we wish to make some more general comments. It seems that a fully expressive epistemic logic generally requires agents to be able to refer to themselves somehow (perhaps by I, although there are surely other possibilities). We have tried to motivate this above, and substantially more discussion is in [Cas68, Lew79, Per79]. Nevertheless, this issue can be virtually disregarded so long as every agent has a unique, individual, commonly known name. In the "standard" logic (Section 2) agent 1 never needs to refer to *himself*; he can equivalently refer to *the agent named* 1. Even when this trick can be used, some minor difficulties remain (for example, our logic could express "everyone knows that he knows φ" as concisely as $E_{all}K_I\varphi$, whereas the standard logic would have to list $K_iK_i\varphi$ for every agent i; and even this only works for a finite and fixed set of agents). But, by and large, the need for knowledge about self-identity is avoided in these special cases. But, once we consider more general situations (group names, names which are non-rigid and so not commonly known, anonymity), we cannot do this, and knowledge about self-identity is required.

5 Concluding Remarks

In this paper we have considered the role of *naming* in propositional modal logics for many agents. We have seen that a practical epistemic logic should make a distinction between individual agents and the names used in reasoning which refer to agents. The questions of which sorts of names are useful and what assumptions are reasonable turn out to be surprisingly complex. Nevertheless, we have shown that many of the possibilities can be captured using very simple propositional logics and uncomplicated possible-worlds semantics.

Some of the specific issues we have raised have been noted elsewhere. However, we know of no other work that treats these questions about naming in a uniform and general framework. It is also the case that previous work has been largely confined to high-level motivation or simple syntactic (language) considerations. We have given a more-or-less complete account of the semantics and axiomatizability of the various logics and the connections between all the variations.

We conclude by noting some significant omissions from this paper. First, in this paper we have been concerned with propositional logics only. But, for some applications, propositional logic is insufficient. Sometimes this is for the usual reasons, to do with the greater expressive power of predicates, functions, equality, and quantification, as well as more realistic semantics of first-order logic. However, there are also some issues specifically related to issues of naming that have no easy propositional solution. For instance, note that in natural language some "names" conventionally get a different interpretation to that given by our logics. For example, in English "John knows that I know φ" asserts that John knows something about me (the speaker of the sentence). Yet in our logics, $K_{John}K_I\varphi$ has John knowing something about himself, perhaps better represented as "John knows that 'I know φ' ". We might describe this situation by saying that the name I can be read using different *scopes*. This is also related to the traditional concern in philosophy about *de re* versus *de dicto* reference, as well as theories of direct reference ([Kap89]). The issues of when different scopes are really useful, and how to capture them properly in a logic, turn out to be far more involved than this simple example might suggest. Propositional logic seems too weak to handle scope and related problems adequately. Scope is one of the problems that is best viewed as part of a general theory of naming in first-order epistemic logic, and we discuss these issues in full detail in part II. One contribution in that paper is a new first-order modal logic that is expressive enough to handle all naming problems we have encountered.

Another omission from this paper concerns the temporal aspects of knowledge. In almost any application for which epistemic logic could be useful, such as in distributed systems analysis, it is important that the knowledge held by the agents changes with time. This paper, as well as part II, is concerned with a general theory of naming only, and so our logics have had no temporal component.[13] Nor have we addressed the question of how an agent should revise his knowledge or beliefs. In general, the machinery for "naming" will only be one aspect of a complete formal description of any system, and the associated logic will also include other features, e.g., temporal modalities. It would be interesting to embed the current work within such a larger context, to see if additional issues arise as well as to find concrete applications of our logics. Some work on these issues can be found in [Gro92].

Acknowledgments

The authors are grateful to Ron Fagin, David Israel, Daphne Koller, Hector Levesque, Karen Myers, Yoav Shoham, Moshe Vardi, and particularly Yves Lésperance, for valuable comments and advice.

A Proof of Theorem 3.1

Theorem 3.1 states that axiom system $AX_{\mathcal{N}}$, consisting of A1, R1, S1, E1, E2, C1 and C2, is sound and complete respect to $\mathcal{M}_{\mathcal{N}}$, the class of all possible-worlds structures for naming.

This proof of completeness follows the standard technique of constructing a canonical model, in which each consistent set of sentences is true (somewhere). For an introductory discussion see, for example, [Che80, HC78, HM85].

We leave verification of soundness to the reader: it is not difficult to show that all instances of the axioms and rules are true in every possible-worlds structure (over Φ and $\mathcal{N}$).

[13] Notice that there are some interesting interaction between names and time. For example, how can an agent refer to *itself* at some future time where it might not even exist?

To prove completeness, we construct a canonical model $M = (W, A, \alpha, \mathcal{K}, \pi, \mu)$. As is usual in such proofs, W consists of all sets of sentences which are consistent (relative to the axiom system) and are maximal in this respect. It is easy to define π by taking $\pi(w)(p) = \texttt{true}$ if and only if $p \in w$. This guarantees that $M, w \models p$ exactly when $p \in w$. Our goal is to construct M so that this property holds for all formulas, not just the primitive propositions. This property, that $M, w \models \varphi$ if and only if $\varphi \in w$, is referred to as the *Truth Lemma*, and is at the heart of all the completeness proofs we present. We achieve the Truth Lemma through our choices for the set of agents, α, the function $\mathcal{K}$, and μ.

The basic idea is that whenever we have $S_{\mathbf{n}}\varphi$ in world w, we would like to define an agent, in $\mu(w, \mathbf{n})$, that knows *just* φ. In this way, we can be certain that someone (the agent just defined) does in fact know φ, whenever such a sentence occurs in w. This tactic works directly if we just seek a result for the logic of $S_{\mathbf{n}}$, but when $E_{\mathbf{n}}$ is in the language we need to define our agent so that it also knows all the sentences that ought to be known by everyone with name $\mathbf{n}$.

For simplicity, we take A to be the collection of all subsets of W; that is, in M we identify an agent with a set of possible worlds. We make the obvious definition for α: $\alpha(w)(= A_w) = \{a : w \in a\}$. $\mathcal{K}$ is defined as: $(w, w') \in \mathcal{K}(a)$ if both $w \in a$ and $w' \in a$. Then, as required, $\mathcal{K}(a)$ (or $\mathcal{K}_a$ as we usually write) is an equivalence on its domain, and this domain is identical with W_a (the set of worlds where a is present).

The remaining task is to define μ. For every sentence $S_{\mathbf{n}}\varphi$ in w, we consider the set $D_{\varphi,w,\mathbf{n}} = \{\varphi\} \cup \{\psi : E_{\mathbf{n}}\psi \in w\}$. Such a set of sentences determines an agent (that is, a set of worlds) $a_{\varphi,w,\mathbf{n}} = \{w' : D_{\varphi,w,\mathbf{n}} \subseteq w'\}$. We simply define $\mu(w, \mathbf{n})$ to be $\{a_{\varphi,w,\mathbf{n}} : S_{\mathbf{n}}\varphi \in w\}$. Of course, it is necessary that this definition of μ satisfy $\mu(w, \mathbf{n}) \subseteq A_w$. We need to show that if $a_{\varphi,w,\mathbf{n}} \in \mu(\mathbf{n}, w)$, then $w \in a_{\varphi,w,\mathbf{n}}$; that is, $D_{\varphi,w,\mathbf{n}} \subseteq w$. However, we know that $S_{\mathbf{n}}\varphi \in w$, and so by axiom S1 we have $\varphi \in w$ as well. Now suppose that $E_{\mathbf{n}}\psi \in w$; we must show $\psi \in w$. By propositional logic, $\psi \Rightarrow (\varphi \Rightarrow \psi)$ is provable, thus (by E2) so is $E_{\mathbf{n}}(\psi \Rightarrow (\varphi \Rightarrow \psi))$. But as $E_{\mathbf{n}}\psi \in w$, we can conclude from E1 that $E_{\mathbf{n}}(\varphi \Rightarrow \psi) \in w$. Using the fact that $S_{\mathbf{n}}\varphi \in w$ and C1, we see that $S_{\mathbf{n}}\psi \in w$; we finish by using S1 again, to deduce that $\psi \in w$.

So the canonical model M is a correct possible-worlds structure. We finish by showing the Truth Lemma that $\mathcal{M}, w \models \varphi$ if and only if $\varphi \in w$, for all formulas φ. Since any consistent set of formulas is contained in some w, this will allow us to conclude that these sentences are also true at some w in $\mathcal{M}$. Demonstrating the satisfiability of any consistent set of sentences amounts to a proof of completeness.

Proof of this property is by induction on the structure of φ, with the base case handled through the definition of π. The inductive steps for the Boolean connectives are easy, so we turn the case of the modal operators.

Suppose $S_{\mathbf{n}}\varphi$ is in w. Consider the agent $a_{\varphi,w,\mathbf{n}}$ (which is in $\mu(w, \mathbf{n})$). The worlds $a_{\varphi,w,\mathbf{n}}$ considers possible from w are just those in $a_{\varphi,w,\mathbf{n}}$ itself, and for any such world w', our construction ensures that $\varphi \in w'$. By our inductive hypothesis, that the Truth Lemma applies for subformulas of $S_{\mathbf{n}}\varphi$ (and in particular, φ), we conclude that $w' \models \varphi$.

So we have found an agent, in $\mu(w, \mathbf{n})$, who knows φ (at w). So $w \models S_{\mathbf{n}}\varphi$ as required.

Suppose $E_{\mathbf{n}}\varphi \in w$, and consider any agent $a \in \mu(w, \mathbf{n})$. Again, the worlds a considers possible from w are just those of a itself, and a was constructed so that all such worlds contain φ. By the inductive hypothesis, φ is actually *true* at all such worlds, and so agent a actually knows φ at w. This being true for all such a, we conclude $w \models E_{\mathbf{n}}\varphi$.

Next, suppose $\neg S_{\mathbf{n}}\varphi \in w$. Assume, for the sake of deriving a contradiction, that some $a_{\varphi',w,\mathbf{n}} \in \mu(w, \mathbf{n})$ knows φ. That is, φ is true in all worlds contained in $a_{\varphi',w,\mathbf{n}}$. By the inductive hypothesis, φ is actually contained in all such worlds. It follows that $\{\varphi', \neg\varphi\} \cup \{\psi : E_{\mathbf{n}}\psi \in w\}$ must be inconsistent. (For if $\{\varphi', \neg\varphi\} \cup \{\psi : E_{\mathbf{n}}\psi \in w\}$ was consistent, it would have a maximal consistent extension, which would be a world in $a_{\varphi',w,\mathbf{n}}$ that does not contain φ.) This inconsistency implies that there must be $\psi_1, \ldots, \psi_n$ such that $E_{\mathbf{n}}\psi_i \in w$ and $\varphi' \wedge \psi_1 \wedge \ldots \wedge \psi_n \Rightarrow \varphi$ is provable. It is not hard to show, from this, that, $S_{\mathbf{n}}\varphi' \wedge E_{\mathbf{n}}\psi_1 \wedge \ldots \wedge E_{\mathbf{n}}\psi_n \Rightarrow S_{\mathbf{n}}\varphi$ is provable also. The proof of this uses only E1, E2 and (especially) C1, but is slightly tedious, so we omit it here. Now, however, we see the contradiction: the antecedent of this latter sentence is surely in w, thus so also must be $S_{\mathbf{n}}\varphi$. But this is contrary to the assumed consistency of w.

The final case to consider is when $\neg E_{\mathbf{n}}\varphi \in w$. Since $\neg E_{\mathbf{n}}\varphi \Rightarrow \neg E_{\mathbf{n}}\mathit{false}$ is provable (propositional logic, and E1, E2), we know $\neg E_{\mathbf{n}}\mathit{false} \in w$. So (by C2) $S_{\mathbf{n}}\mathit{true} \in w$. Given this, let us consider the agent $a_{true,w,\mathbf{n}}$ (in $\mu(w, \mathbf{n})$). Could $a_{true,w,\mathbf{n}}$ know φ? If this was the case then, arguing similarly to the previous paragraph, there are $\psi_1, \ldots, \psi_n$ such that $E_{\mathbf{n}}\psi_i \in w$ and $\psi_1 \wedge \ldots \wedge \psi_n \Rightarrow \varphi$ is provable. From this is follows (using only E1, E2 here) that $E_{\mathbf{n}}\psi_1 \wedge \ldots \wedge E_{\mathbf{n}}\psi_n \Rightarrow E_{\mathbf{n}}\varphi$ is provable. But then $E_{\mathbf{n}}\varphi \in w$, which is a contradiction. We have demonstrated one agent in $\mu(w, \mathbf{n})$ that is ignorant of φ; this is enough to show $w \models \neg E_{\mathbf{n}}\varphi$ as required. ∎

B Proof of Theorem 3.2

Theorem 3.2 states that the problem of deciding whether a $\mathcal{L}_{\mathcal{N}}$ formula is valid, with respect to the class of models $\mathcal{M}_{\mathcal{N}}$, is *PSPACE*-complete. In the following, we look instead at the complementary problem of deciding a formula's satisfiability; since *PSPACE* is a deterministic complexity class, showing *PSPACE*-completeness for satisfiability is equivalent to proving the theorem as stated.

We will not give all details of the proof of this theorem here. Very similar proofs, for the case of more standard logics with single-agent modal operators like K_i, have been given by Ladner [Lad77] and Halpern and Moses [HM85]. Here we restrict ourselves to a general outline of how the techniques in these papers can be modified to suit our requirements.

The easier part of the proof is to show that the satisfiability problem is *PSPACE*-hard. For this, we consider the sublanguage $\mathcal{L}_{\mathcal{N}}^{E} \subset \mathcal{L}_{\mathcal{N}}$ consisting of the formulas where only modalities like E_{name} appear (that is, there are no "someone knows" operators present). A straightforward semantic argument, based on our completeness result, shows that we can restrict attention to models where each names denotes at most one agent in any world. Intuitively, without $S_{\mathbf{n}}$ in the language it is impossible to say whether $E_{\mathbf{n}}$ refers to many agents or to just one quite ignorant agent with name $\mathbf{n}$; see Section 3.3 for more discussion of this. This shows that the satisfiability problem for $\mathcal{L}_{\mathcal{N}}^{E}$ is equivalent to the satisfiability decision problem in a more standard modal logic where each name denotes one agent. Then a very general result of Ladner's, [Lad77, Theorem 3.1], applies to show immediately that this problem is *PSPACE*-hard. The same idea, of looking at that fragment of the logic which only uses E, can be used for every propositional logic we present in this paper, although details differ from case to case. (Where Ladner's theorem does not apply other techniques from [HM85] can be used instead.) All these logics are *PSPACE*-hard.

It is more difficult to show that the satisfiability problem is in *PSPACE*, that is, that there is a polynomial-space algorithm for testing satisfiability. Halpern and Moses develop algorithms for the logics they consider by showing that any satisfiable formula has a model which looks like a tree, with depth which is polynomially bounded in the size of the formula. Further, if such a tree model exists, a tableau-like method can be used to construct it. Because this construction can be carried out using depth-first search, we usually only need space polynomial in the tree's depth.

The tableau technique can be adapted for our logic. The idea is to reduce the problem of determining satisfiability for some formula φ in our logic to the satisfiability problem for some other φ' in a more conventional logic (where modal operators refer to just one agent).

In the following, let m be the length of φ. We claim that if φ is satisfiable in any structure in $\mathcal{M}_{\mathcal{N}}$, it is satisfiable in a structure where (1) names denote disjoint sets of agents, i.e., $\mu(w, \mathbf{n}) \cap \mu(w', \mathbf{n}') = \emptyset$ if $w \neq w'$ or $\mathbf{n} \neq \mathbf{n}'$, and (2) at every world, a name denotes either no agents or else denotes m agents. The most difficult part of proving this is showing that no name needs to denote more than m agents. Suppose φ is true at world w, in model M, and that name $\mathbf{n}$ appears in φ somewhere. In general, it is possible that $\mu(w, \mathbf{n})$ has more than m members. However, look at the (at most m) subformulas of φ which are true at w. Each subformula of the form $S_{\mathbf{n}}\psi$ or $\neg E_{\mathbf{n}}\psi$ can be forced to be true by just a single agent in $\mu(w, \mathbf{n})$. So for each such subformula choose one such agent. An inductive argument shows that M', the structure that is identical to M except that $\mu(w, \mathbf{n})$ is the set of agents just selected, still satisfies φ at world w.

Using this observation, we consider a new logic with modal operators $K_{\{\mathbf{n},1\}}$, ..., $K_{\{\mathbf{n},m\}}$ for every name $\mathbf{n}$ which appears in φ. These are interpreted as single-agent epistemic operators, over models of the special form described above, as follows: $(w, w') \in \mathcal{K}_{\{\mathbf{n},i\}}$ just if the i'th agent in some ordering of $\mu(w, \mathbf{n})$ considers w' possible from w. In these models, $E_{\mathbf{n}}\psi$ is equivalent to $(K_{\{\mathbf{n},1\}}\psi \wedge \ldots \wedge K_{\{\mathbf{n},m\}}\psi)$ and $S_{\mathbf{n}}\psi$ is equivalent to $(K_{\{\mathbf{n},1\}}\psi \vee \ldots \vee K_{\{\mathbf{n},m\}}\psi)$.

It is possible to rewrite φ completely using these equivalences so that no modalities

$E_{\mathbf{n}}$ or $S_{\mathbf{n}}$ appear, and then one of the satisfiability algorithms from [HM85] applies almost directly (only small details differ, to account for the possibility that a name $\mathbf{n}$ might not denote any agents, in which case all $K_{\{\mathbf{n},i\}}$ operators should correspond to empty accessibility relations). Unfortunately, the new sentence φ' can be exponentially longer than φ, and so this procedure is not in *PSPACE*. However, this exponential blowup is easy to avoid. Although φ' can be very long, it is easy to see that the number of distinct subformulas in φ' is at most m^2. The complexity of the Halpern and Moses's algorithms depends on the number of subformulas of φ' rather than the length of φ' itself (assuming that the algorithm is implemented to avoid ever explicitly considering φ' in full, which is easy to do). In this way, we can obtain a *PSPACE* algorithm for satisfiability. ∎

C Proof of Theorem 3.3

Recall, Theorem 3.3 states that axiom system $AX'_{\mathcal{N}}$, consisting of A1, R1, S1, S2, E1, E2, C1, C2, C3, and C4 is sound and complete respect to $\mathcal{M}'_{\mathcal{N}}$, the class of all possible-worlds structures where agents know their names.

Again, soundness is relatively easy, and we omit details. (Note that we briefly discussed reasons for the soundness of C3 and C4, in Section 3.4.)

Completeness is shown with a similar style of proof to that used previously. We begin defining the canonical model in the same way: W is all maximal consistent sets of sentences, and α, π, A, and $\mathcal{K}$ are defined exactly as was done in the proof of Theorem 3.1. The interesting problem is to define $\mu(w, \mathbf{n})$: what agents belong to name $\mathbf{n}$ at world w?

Recall that the basic idea in the proof of Theorem 3.1 was to define one agent who knows φ, for each $S_{\mathbf{n}}\varphi$ in a world w. There is another, stronger and often more useful, approach we could have taken instead.[14] If $\varphi_1, \varphi_2, \varphi_3, \ldots$ is a sequence of sentences that are increasingly stronger ($\varphi_{i+1} \Rightarrow \varphi_i$), such that $S_{\mathbf{n}}\varphi_i \in w$, we could define an agent in $\mu(w, \mathbf{n})$ who knows *all* of the φ_i. This agent's knowledge is, essentially, the upper bound of the agents who know φ_1, φ_2, etc.

In this proof, we use this idea of sequences (although somewhat modified). There are other necessary changes as well. As discussed in the paper, when agents know their names, each agent (in $\mathbf{n}$ at w) necessarily either knows formula φ, or else knows that $\neg E_{\mathbf{n}}\varphi$. Our definition of agents must respect this. Let us call a set of sentences *definite* if either φ or $\neg E_{\mathbf{n}}\varphi$ is provable from it, for every φ.

Combining these ideas, we say that a (definite) *sequence*, relative to w and $\mathbf{n}$, is a collection of formulas $\sigma = \{S_{\mathbf{n}}\varphi_1, S_{\mathbf{n}}\varphi_2, \ldots\}$ satisfying (1) $S_{\mathbf{n}}\varphi_i \in w$, (2) $S_{\mathbf{n}}\varphi_{i+1} \Rightarrow S_{\mathbf{n}}\varphi_i$ is provable, and (3) $\{S_{\mathbf{n}}\varphi_1, S_{\mathbf{n}}\varphi_2, \ldots\}$ is definite. Any such sequence determines

[14] For instance, in Section 3.3 we mentioned two axioms that, in conjunction with the others, are sound and complete for the case where one name is contained in another. This is much more easily proved using the idea of sequences discussed here.

an agent, by $a_{\sigma,w,\mathbf{n}} = \{w' : \sigma \subseteq w'\}$ (recall that in this model, an agent is just a set of worlds). We say that $\mu(w, \mathbf{n})$ is just the set of these agents (i.e., agents determined by sequences relative to $w, \mathbf{n}$). Note that $\sigma \subseteq w$ automatically, so $w \in a$ $(= W_a)$ as required. As we have suggested, agent $a_{\sigma,w,\mathbf{n}}$ is being determined here by what it should know: just the formulas in σ. This explains why σ is required to be definite. It also justifies looking at $S_{\mathbf{n}}\varphi_i$ rather that just φ_i. Whatever the agent knows, it also knows that someone (itself) knows it. If we had defined sequences simply as $\{\varphi_1, \varphi_2, \ldots\}$ where $S_{\mathbf{n}}\varphi_i \in w$, this property would not be guaranteed.

The remainder of the proof is to show (1) that the model, as just constructed, is indeed a possible-worlds structure where agents know their names, and (2) show that the Truth Lemma holds: $\mathcal{M}, w \models \varphi$ if and only if $\varphi \in w$. We start with the former.

The problem is to check that "agents know their names". So let us look at some $a = a_{\sigma,w,\mathbf{n}} \in \mu(w, \mathbf{n})$. Suppose w' is a world that a considers possible from w; that is, $w' \in a$. We want this same agent to be in $\mu(w', \mathbf{n})$ also, and this will be so exactly if σ is also a sequence at w'. We really only need to show that $S_{\mathbf{n}}\varphi_i \in w'$ (because the second and third properties in our definition of what a sequence is do not depend on the world we are considering). But $S_{\mathbf{n}}\varphi_i \in w'$ is immediate from the construction of $a = a_{\sigma,w,\mathbf{n}}$ (and noting that $w' \in a$).

Next, we turn to the proof of the Truth Lemma. As usual, this proof is by induction on the structure of φ, and the only interesting cases are those involving $E_{\mathbf{n}}$ and $S_{\mathbf{n}}$.

Suppose, $S_{\mathbf{n}}\varphi$ is in w. We construct an agent in $\mu(w, \mathbf{n})$ that knows φ. We can do this, because there is a sequence $\sigma = \{S_{\mathbf{n}}\varphi_1, S_{\mathbf{n}}\varphi_2, \ldots\}$ such that $\varphi_1 \Rightarrow \varphi$. For let $\psi_1, \psi_2, \ldots$ be a complete enumeration of *all* formulas in the language. The sequence we want is as follows. First, $S_{\mathbf{n}}\varphi_1$ is $S_{\mathbf{n}}\varphi$. Next, assume we have already chosen φ_i. By C3, $S_{\mathbf{n}}\varphi_i \Rightarrow S_{\mathbf{n}}(\varphi_i \wedge \psi_i) \vee S_{\mathbf{n}}(\varphi_i \wedge \neg E_{\mathbf{n}}\psi_i)$. Since $S_{\mathbf{n}}\varphi_i \in w$ (by assumption) one or both of $S_{\mathbf{n}}(\varphi_i \wedge \psi_i)$ or $S_{\mathbf{n}}(\varphi_i \wedge \neg E_{\mathbf{n}}\psi_i)$ is also in w. Select one of these: this will be $S_{\mathbf{n}}\varphi_{i+1}$. This construction determines our sequence. It is clearly definite. Moreover, agent $a_{\sigma,w,\mathbf{n}}$ knows φ (since $S_{\mathbf{n}}\varphi$ is in all worlds is considers possible, then so (by S1) is φ; and by induction, φ is actually true at all such worlds).

Suppose $E_{\mathbf{n}}\varphi \in w$. Consider *any* agent $a_{\sigma,w,\mathbf{n}} \in \mu(w, \mathbf{n})$. By definiteness, there is some formula $S_{\mathbf{n}}\varphi_i \in \sigma$ such that $S_{\mathbf{n}}\varphi_i \Rightarrow \varphi$ or $S_{\mathbf{n}}\varphi_i \Rightarrow \neg E_{\mathbf{n}}\varphi$. But the latter is impossible; since then $\neg E_{\mathbf{n}}\varphi \in w$ contrary to consistency. Thus $S_{\mathbf{n}}\varphi_i \Rightarrow \varphi$. Arguing as in the conclusion of the previous paragraph, $a_{\sigma,w,\mathbf{n}}$ knows φ.

Next, suppose $\neg S_{\mathbf{n}}\varphi \in w$. We must show that no agent in $\mu(w, \mathbf{n})$ knows φ. Suppose, to the contrary, that $a = a_{\sigma,w,\mathbf{n}}$ does know φ. Let $\sigma = \{S_{\mathbf{n}}\varphi_1, S_{\mathbf{n}}\varphi_2, \ldots\}$. Then $\{\neg\varphi, S_{\mathbf{n}}\varphi_1, S_{\mathbf{n}}\varphi_2, \ldots\}$ is inconsistent (otherwise, we could extend this to a maximal consistent set, which would be a world a considers possible where φ is false). But then there must be some $S_{\mathbf{n}}\varphi_i$ such that $S_{\mathbf{n}}\varphi_i \Rightarrow \varphi$ is provable (we only need to consider one φ_i because of the nesting property of sequences). Applying E2 and C1, it follows that $S_{\mathbf{n}}S_{\mathbf{n}}\varphi_i \Rightarrow S_{\mathbf{n}}\varphi$ is provable. But $S_{\mathbf{n}}\varphi_i \in w$, so by S2, $S_{\mathbf{n}}S_{\mathbf{n}}\varphi_i \in w$ also. Then it follows that $S_{\mathbf{n}}\varphi \in w$, a contradiction.

Finally, and most difficult, suppose φ is such that $\neg E_{\mathbf{n}}\varphi \in w$. We must demonstrate an agent, in $\mu(w, \mathbf{n})$, which does not know φ.

Consider again our enumeration of all the sentences in the language, $\psi_1, \psi_2, \ldots$, and fix some $l \geq 1$. Let us consider all the 2^l sentences $S_{\mathbf{n}}\chi_i$ of the form $S_{\mathbf{n}}((\neg E_{\mathbf{n}})\psi_1 \wedge (\neg E_{\mathbf{n}})\psi_2 \wedge \ldots \wedge (\neg E_{\mathbf{n}})\psi_l))$. (That is, we construct the 2^l alternatives by including or omitting the $E_{\mathbf{n}}$ operators preceding each of the ψ_i.) We claim that there is at least one χ_j such that $S_{\mathbf{n}}\chi_j \in w$, and also that $S_{\mathbf{n}}\chi_j \Rightarrow \varphi$ is *not* provable. For suppose this is false, and let $A = \{i : S_{\mathbf{n}}\chi_i \not\Rightarrow \varphi\}$. Then our assumption is that $\bigwedge_{i \in A} \neg S_{\mathbf{n}}\chi_i \in w$, and so from C4, $E_{\mathbf{n}}(\bigvee_{i \notin A} S_{\mathbf{n}}\chi_i) \in w$. For such i, we know $S_{\mathbf{n}}\chi_i \Rightarrow \varphi$, and so $\bigvee_{i \notin A} S_{\mathbf{n}}\chi_i \Rightarrow \varphi$. By E1 and E2 it follows that $E_{\mathbf{n}}(\bigvee_{i \notin A} S_{\mathbf{n}}\chi_i) \Rightarrow E_{\mathbf{n}}\varphi$. But then $E_{\mathbf{n}}\varphi \in w$, which is a contradiction, and it follows that a suitable $S_{\mathbf{n}}\chi_j$ must exist.

Our sequence is constructed by letting $S_{\mathbf{n}}\varphi_l$ be the $S_{\mathbf{n}}\chi_j$ constructed above, and doing this for every l. We need to do this so that the sequence satisfies the nesting property $(S_{\mathbf{n}}\varphi_{i+1} \Rightarrow S_{\mathbf{n}}\varphi_i)$. However, by E1, E2, C1, S1 we note that if some $S_{\mathbf{n}}((\neg E_{\mathbf{n}})\psi_1 \wedge \ldots \wedge (\neg E_{\mathbf{n}})\psi_l)$ does not entail φ, then neither will $S_{\mathbf{n}}((\neg E_{\mathbf{n}})\psi_1 \wedge \ldots \wedge (\neg E_{\mathbf{n}})\psi_{l-1})$; in addition, if the former sentence is in w so will be the latter. This observation is sufficient to show that we can construct the sequences to possess the nesting property. (For any l, consider the set of acceptable $S_{\mathbf{n}}\chi_j$. We have seen that this is nonempty. Further, each such sentence implies at least one other in the set of $(l-1)$. We can consider this structure as a tree where the nodes on level l are these sentences, and the arcs are relations of implication. This tree is infinite in size, yet only has a finite number of nodes at any level. Using König's lemma, there must be at least one infinite path: this is our sequence.)

The sequence constructed is clearly definite. Finally, observe that the agent corresponding to this sequence does not know φ. For if it did, there would necessarily be some $S_{\mathbf{n}}\varphi_i$ with $S_{\mathbf{n}}\varphi_i \Rightarrow \varphi$, but the sequence was constructed so that this cannot happen. ∎

D Proof of Theorem 4.1

Theorem 4.1 states that axiom system $A\mathcal{X}_{\mathcal{N}}^{\mathbf{ksi}}$, consisting of A1, R1, E1, E2, C1, C2, S1$'$, K1, K2, K3, and K4 is sound and complete respect the class of all possible-worlds structures with knowledge about self-identity.

Here we prove completeness for the more general semantics, where some names and propositions are designated as being absolute (a discussion of this case appeared after the statement of Theorem 4.1 in the paper). Recall that in this logic, axiom S1$''$ ($S_{\mathbf{n}}\varphi \Rightarrow \varphi$ for objective φ) replaces S1$'$. After the proof, we will relate this more general result back to Theorem 4.1.

We follow the pattern of our earlier proofs. We do not discuss soundness because it is relatively straightforward to verify that the axioms are all true in all possible-worlds structures with knowledge about self-identity. Completeness is shown by constructing a canonical model, $M = (W, A, \alpha, \mathcal{K}, \pi, \mu)$.

Consider the set of all maximal consistent sets of sentences. Here, we do *not* wish to regard these as individual worlds, because they also are dependent upon the choice of viewpoint (consider two sets that differ only on sentences like $K_I\varphi$, for example). Instead, each possible world in W will be an equivalence class of those maximal consistent sets agreeing on all objective formulas. If $w \in W$, then $Obj(w)$ is the set of objective formulas corresponding to w (i.e., contained in all members of w).

The set A of agents is just the collection of all maximal consistent sets of sentences. Each $a \in A$ belongs to one $w \in W$, and we make the obvious definition $\alpha(w)$ $(= A_w) = \{a : a \in w\}$; that is, $A_w = w$. Note that each agent exists in just one world.

$\mathcal{K}$ contains $((w, a), (w', a'))$ exactly if $\{\varphi : K_I\varphi \in a\} \subseteq a'$ (where a, a' belong to the equivalence classes w, w' respectively). Intuitively, a' contains everything that a knows. To be a correctly specified model, this relation should be reflexive, symmetric, and transitive. It is not too difficult to show that these properties follow from the presence in the logic of the formulas $K\varphi \Rightarrow \varphi$, $K_I\varphi \Rightarrow K_IK_I\varphi$, and $\neg K_I\varphi \Rightarrow K_I\neg K_I\varphi$, respectively (these are essentially instances of K4, K2, K3).

To define the truth assignment π, we will say that $\pi(w, a)(p) = \mathtt{true}$ exactly if $p \in a$. Note that this works correctly for absolute propositions: in this case, π is independent of the agent, as we would wish.

Finally, and most complex, is the definition of μ. Given $a \in w \in W$, we say an $(a, \mathbf{n})$ *sequence* is a set $\{K_I\varphi_1, K_I\varphi_2, \ldots\}$ such that (1) $S_{\mathbf{n}}\varphi_i \in a$, (2) $K_I\varphi_{i+1} \Rightarrow K_I\varphi_i$ is provable, and (3) for any ψ, there is some φ_i such that either $K_I\varphi_i \Rightarrow \psi$ or $K_I\varphi_i \Rightarrow \neg K_I\psi$ is provable.

Let σ be any $(a, \mathbf{n})$ sequence in $a \in w \in W$. We use this to select another agent in w as follows. Find any maximal consistent extension of σ which is also in w. (There could be many possibilities. It does not matter which is used, except that if $\mathbf{n}$ is I, we simply choose a itself, and if $\mathbf{n}$ is an absolute name, the same set should be chosen when looking at this sequence from any other $a' \in w$.) The chosen maximal consistent set (i.e., the chosen agent) will be denoted $b_{\sigma,w,a}$. We simply let $(a, b) \in \mu(w, \mathbf{n})$ if and only if $b = b_{\sigma,w,a}$ for some σ. (This works correctly for absolute names also; because then formulas like $S_{\mathbf{n}}\varphi$ are objective and it does not matter which a in w we look at.) Note that in the case where $\mathbf{n} = I$, then we see that $\mu(w, I)$ will be simply the identity relation on agents in w (as is required).

Our insistence that we look at maximal consistent sets in w ensures that $b_{\sigma,w,a} \in w$, which is necessary for this definition of μ to be valid. However, we are not yet finished, for we have not shown that there always is a suitable $b_{\sigma,w,a}$. We must show that $Obj(a) \cup \sigma$ is consistent (for then this has a maximal consistent extension, which could be $b_{\sigma,w,a}$). Suppose this was not consistent. Then there must be φ_i and some objective ψ such that $K_I\varphi_i \Rightarrow \neg\psi$ is provable. (We can assume just one ψ because objective sentences are closed under boolean connectives, and just one φ_i because of the nesting property of sequences.) But then $S_{\mathbf{n}}K_I\varphi_i \Rightarrow S_{\mathbf{n}}\neg\psi$ (by axioms E2 and C1). So, using axiom S1$''$ and K2 (as well as E1, E2, C1), $S_{\mathbf{n}}\varphi_i \Rightarrow \neg\psi$ (we are allowed to use S1$''$ here because $\neg\psi$ is objective). Yet this contradicts the consistency of a itself (which contains $S_{\mathbf{n}}\varphi_i$ and ψ). We conclude that the agent $b_{\sigma,w,\mathbf{n}}$ does exist. Note that the set $b_{\sigma,w,\mathbf{n}}$ contains a formula $K_I\varphi$ exactly if $\varphi = \varphi_i$ for some i (this follows from the third property of sequences, as well as K2, K4). This observation will be useful later.

At this point, we have defined the structure M so that it is indeed a correctly specified possible-worlds structure with knowledge about self-identity. Now, we must prove the Truth Lemma, which in this case means we must show that, for any world w and agent a in that world, $w, a \models \varphi$ if and only if $\varphi \in a$. Showing this demonstrates that any consistent set of sentences is true at some world/agent pair (because such a set has at least one maximal consistent extension), and from this completeness follows.

The proof is by induction on the structure of φ, and, as usual, the boolean connectives are easy so we look at the cases involving the modal operators. In the following, $a \in w \in W$ is some agent/maximal consistent set. In our model a uniquely determines w, so we often omit mentioning the latter.

Suppose $S_{\mathbf{n}}\varphi \in a$. We can show there is an $(a, \mathbf{n})$ sequence, with $\varphi_1 = \varphi$, using the following construction. First, $\varphi_1 = \varphi$. Now let $\psi_1, \psi_2, \ldots$ be an enumeration of all formulas in the language, and suppose that we have determined φ_i somehow. We could try to set $\varphi_{i+1} = \varphi_i \wedge \psi_i$. This could fail, but only if $\neg S_{\mathbf{n}}(\varphi_i \wedge \psi_i) \in a$. But then, by K3, $E_{\mathbf{n}}\neg K_I(\varphi_i \wedge \psi_i) \in a$, and so $E_{\mathbf{n}}(\neg K_I\varphi_i \vee \neg K_I\psi_i) \in a$ (this last step follows from several applications of E1 and E2). But we know that $S_{\mathbf{n}}\varphi_i \in a$, so by K2 we also have that $S_{\mathbf{n}}(K_I\varphi_i \wedge \varphi_i) \in a$. From this, the previous observation, and C1, a contains $S_{\mathbf{n}}(\varphi_i \wedge \neg K_I\psi_i)$. So we can let φ_{i+1} be $\varphi_i \wedge \neg K_I\psi_i$.

Completing this construction will determine a sequence (that is, the three required properties of sequences will hold). The corresponding agent, a' say, satisfies $(a, a') \in \mu(w, \mathbf{n})$ and knows φ (because this sequence contains $K_I\varphi$, and by the definition of $\mathcal{K}$ we know that φ must be true in all world/agent pairs considered possible by a'). So $w, a \models S_{\mathbf{n}}\varphi$, which is what we wished to show.

Next, suppose that $E_{\mathbf{n}}\varphi \in a$, and look at any agent $b = b_{\sigma,w,a}$ so that $(a, b) \in \mu(w, \mathbf{n})$. There is some $K_I\varphi_i \in \sigma$, so that $\mathcal{K}_I\varphi_i \Rightarrow \varphi$ or $K_I\varphi_i \Rightarrow \neg K_I\varphi$. If the former, we are done, because then b knows φ (by E1, E2, K1, K2, we have $K_I\varphi \in b$, so φ is contained in all worlds b considers possible). On the other hand, suppose $K_I\varphi_i \Rightarrow \neg K_I\varphi$. From C2 and E2, $S_{\mathbf{n}}K_I\varphi_i \Rightarrow S_{\mathbf{n}}\neg K_I\varphi$ is provable. We know $S_{\mathbf{n}}\varphi_i \in a$, so (by K2, K4) $S_{\mathbf{n}}K_I\varphi_i \in a$, and it follows that $S_{\mathbf{n}}\neg K_I\varphi \in a$ also. But, with C1 and $E_{\mathbf{n}}\varphi \in a$, we see that a contains $S_{\mathbf{n}}(\neg K_I\varphi \wedge \varphi)$, and so (K2 again) $S_{\mathbf{n}}(\neg K_I\varphi \wedge \varphi \wedge K_I(\neg K_I\varphi \wedge \varphi))$. But this latter sentence is inconsistent (axiom K4 can be used to show that the formula inside $S_{\mathbf{n}}$ is equivalent to *false*, but $S_{\mathbf{n}}$*false* contradicts S1$''$). So we are done.

Next, suppose that $\neg S_{\mathbf{n}}\varphi \in a$. Suppose that there is an agent $b = b_{\sigma,w,a}$, with $(a, b) \in \mu(w, \mathbf{n})$, who knows φ. It must be that case that φ is true in all (w', b') pairs considered possible from (w, b). So the set $\{\neg\varphi, \varphi_1, \varphi_2, \varphi_3, \ldots\}$ is inconsistent (where $K_I\varphi_i \in \sigma$; here we are using the fact that $K_I\varphi \in b$ only if $K_I\varphi \in \sigma$). However, this is impossible, for then there is φ_i so that $\varphi_i \Rightarrow \varphi$ is provable, and so $S_{\mathbf{n}}\varphi_i \Rightarrow S_{\mathbf{n}}\varphi$ is also provable, which contradicts the consistency of a (which contains the antecedent but not $S_{\mathbf{n}}\varphi$).

Finally consider if $\neg E_{\mathbf{n}}\varphi \in a$. We can show that $w, a \models \neg E_{\mathbf{n}}\varphi$, using a similar argument to that in the proof of Theorem 3.3. Here, as there, we consider an enumeration of all formulas in the language, $\psi_1, \psi_2, \ldots$, and some $l \geq 1$. Now for any ψ_i, $K_I\psi_i \vee K_I\neg K_I\psi_i$ is provable (by K3). Considering the conjunction of this for $i \leq l$, rearranging according to propositional logic, and finally using E1 and E2 we conclude that the formula χ

$$\bigvee K_I((\neg K_I)\psi_1 \wedge (\neg K_I)(\psi_2) \wedge \ldots \wedge (\neg K_I)\psi_l)$$

is provable. The disjunction is over all 2^l formulas $K_I\chi_i$ which are obtained by omitting or retaining the $(\neg K_I)$ before the ψ_i.

We can duplicate the rest of the argument from the proof of Theorem 3.3 essentially unchanged, if only we show that there is some χ_i such that $S_{\mathbf{n}}\chi_i \in a$, and that $K_I\chi_i \Rightarrow \varphi$ is *not* provable. But suppose this were not so. Then, by K3, $E_{\mathbf{n}}\neg K_I\chi_i$ is in a for each of the χ_i for which $K_I\chi_i$ fails to imply φ. But from these sentences and $E_{\mathbf{n}}\chi$ (which we know to be provable), we conclude that $E_{\mathbf{n}}(K_I\chi_1 \vee \ldots \vee K_I\chi_k) \in a$ (where $K_I\chi_1, \ldots, K_I\chi_k$ *do* imply φ). Then it is easy to see (by E1, E2, propositional logic) that $E_{\mathbf{n}}\varphi \in a$, which is a contradiction. So the desired $S_{\mathbf{n}}\chi_i$ must exist, and $K_I\chi_i$ can be used as the l'th formula in the constructed sequence. (The argument using König's lemma to show that we can choose the successive $K_I\chi_i$ to satisfy the nesting property is the same as for the proof of Theorem 3.3.) ∎

We conclude with a few observations about this proof.

First, suppose that the language does not contain the symbol I. Then a sound and complete axiomatization is given by the axiomatization in Theorem 4.1 but excluding K1, K2, K3, K4. This is not surprising; we might expect that K1, K2, K3, and K4 are irrelevant in this case because their role seems to be in describing properties of the symbol I. It is not hard to prove this observation. In essence, the proof of Theorem 3.1 can be adapted, by using the idea of clustering maximal consistent sets according to the objective formulas they contain, just as in the proof just seen. The ideas of sequences and definiteness are not required. However, because this observation is of little practical interest, we do not give further details.

Our second observation concerns the assumption, made in the proof above, that some propositions and names are intended to be absolute. If this is not the case, then S1$''$ reduces to just $S_{\mathbf{n}}\mathit{true} \Rightarrow \mathit{true}$ and $S_{\mathbf{n}}\mathit{false} \Rightarrow \mathit{false}$ because *true* and *false* are (up to logical equivalence) the only objective formulas. The former sentence is provable from propositional logic anyway, while the latter is more simply written as $\neg S_{\mathbf{n}}\mathit{false}$. This is Axiom S1$'$ which appeared in the statement of Theorem 4.1. It is interesting that, in this case, there is only *one* possible world in the canonical model structure. Knowledge and ignorance is modeled as an agent's uncertainly about who it might be within this one world.

Finally, suppose $\mathbf{n}$ is a symbol we wish to use as both a (relative) proposition and an absolute name. Naturally, we want $\mathbf{n}$ to "denote" the same agents in each world, no matter which way it is interpreted. As mentioned in the main paper, a sound and complete axiomatization for such structures is obtained by adding axioms $\mathbf{n} \wedge K_I\varphi \Rightarrow S_{\mathbf{n}}\varphi$, and $S_{\mathbf{n}}(\psi \vee \neg\mathbf{n}) \Rightarrow \psi$ (for objective ψ). To show this, we must carry out the above proof so that $(b, a) \in \mu(w, \mathbf{n})$ if and only if $\pi(w, a)(\mathbf{n}) = \mathtt{true}$ (for any $b \in A_w$, since $\mathbf{n}$ is absolute).

Very briefly, the proof runs as follows. First, look at $a \in w$, with $\mathbf{n} \in a$ (so $\pi(w, a)(\mathbf{n}) = \mathtt{true}$). We are able to write all the $K_I\varphi$ formulas in a as a sequence $\{K_I\varphi_1, K_I\varphi_2, \ldots\}$. This is possible because the set $\{\varphi : K_I\varphi \in a\}$ is closed under conjunction (essentially, this enables us to cover all the $K_I\varphi$ somehow, as part of a sequence), and because our first new axiom allows us to conclude $S_{\mathbf{n}}\varphi_i \in a$. By our definition of μ, this sequence must select an agent in $\mu(w, \mathbf{n})$; however, the sequence is such that a itself will be an acceptable choice. So we can perform the construction so that any a with $\pi(w, a)(\mathbf{n}) = \mathtt{true}$ also has name $\mathbf{n}$.

Conversely, suppose $\sigma = \{K_I\varphi_1, K_I\varphi_2, \ldots\}$ is a $(b, \mathbf{n})$ sequence in w (it does not actually matter what agent $b \in w$ we consider since all formulas $S_{\mathbf{n}}\varphi$ are objective). This sequence is used to define an agent, i.e., $a \in w$ with $\sigma \subseteq a$, and we wish to choose a so that $\mathbf{n} \in a$ also. This will be impossible only if $Obj(w) \cup \sigma \cup \{\mathbf{n}\}$ is inconsistent, and *this* is so if there is some i and some objective ψ with $K_I\varphi_i \Rightarrow (\neg\psi \vee \neg\mathbf{n})$ provable. But from this we can show that $S_{\mathbf{n}}\varphi_i \Rightarrow S_{\mathbf{n}}(\neg\psi \vee \neg\mathbf{n})$ is also provable, and since $S_{\mathbf{n}}\varphi_i \in b$ (because σ is a sequence) so is the consequent. But then our second axiom ensures that $\neg\psi \in a$, which gives a contradiction.

E Proof of Theorem 4.2

Theorem 4.2 stated that the validity problem for the class of structures with knowledge about self-identity is *PSPACE*-complete. As with Theorem 3.2, we do not give full details of the proof of this result. We have already discussed, in Appendix B, how to show that the validity problem is *PSPACE*-hard. Then it remains for us to show that there is a *PSPACE* algorithm for satisfiability. As we did in Appendix B, we refer the reader to [HM85] for details of the *tableau* algorithms that can be modified to decide satisfiability in our logic.

The observation that allows these results to be used is that, when trying to decide satisfiability for a formula φ of length m, it is sufficient to look for models where $\{b : (a, b) \in \mu(w, \mathbf{n})\}$ is either empty or of size exactly m, for all w and a and all names $\mathbf{n}$ other than I. The argument for this parallels that given in Appendix B. As a result of this, we find it useful to consider an extended language with new modal operators $K_{\{\mathbf{n},i\}}$ for $i \leq m$ and $\mathbf{n} \neq I$.

Consider the formula φ', constructed from φ by replacing every occurrence of $E_{\mathbf{n}}\psi$ by $(K_{\{\mathbf{n},1\}}K_I\psi \wedge \ldots \wedge K_{\{\mathbf{n},m\}}K_I\psi)$, and occurrences of $S_{\mathbf{n}}\psi$ by $(K_{\{\mathbf{n},1\}}K_I\psi \vee \ldots \vee K_{\{\mathbf{n},m\}}K_I\psi)$. It is relatively straightforward to verify that φ has a model (with knowledge about self-identity) just if φ' has a conventional possible-worlds model. (More precisely, in this new model $\mathcal{K}_I$ must be an equivalence relation, and each $\mathcal{K}_{\{\mathbf{n},i\}}$ must be a functional relation.)

Finally, we can verify that algorithms in [HM85] can be used to check the consistency of formulas like φ', with space complexity which is polynomial in the number of distinct subformulas in φ' (which is important, because although φ' might be exponentially longer than φ, the number of distinct subformulas grows by a factor m at most). The result follows. ∎

References

[ABLP91] M. Abadi, M. Burrows, B. Lampson, and G. Plotkin. A calculus for access control in distributed systems. Technical report, SRI, 1991.

[Cas68] H.-N. Castañeda. On the logic of attributions of self knowledge to others. *Journal of Philosophy*, LXV(15):439–456, 1968.

[Che80] B. F. Chellas. *Modal Logic*. Cambridge University Press, 1980.

[CM86] K. M. Chandy and J. Misra. How processes learn. *Distributed Computing*, 1(1):40–52, 1986.

[DM90] C. Dwork and Y. Moses. Knowledge and common knowledge in a Byzantine environment: crash failures. *Information and Computation*, 88(2):156–186, 1990.

[FH88] R. Fagin and J. Y. Halpern. Belief, awareness, and limited reasoning. *Artificial Intelligence*, 34:39–76, 1988.

[Gro] A. J. Grove. Naming and identity in epistemic logics. Part II: A first-order logic for naming. To appear in *Artificial Intelligence*. A preliminary paper reporting this work appeared in the Proceedings of the Second International Conference on Principles of Knowledge Representation and Reasoning (KR), 1991.

[Gro92] A. J. Grove. *Topics in Multi-agent Logics*. PhD thesis, Stanford University, 1992.

[Hal87] J. Y. Halpern. Using reasoning about knowledge to analyze distributed systems. In J. Traub et al., editors, *Annual Review of Computer Science, Vol. 2*, pages 37–68. Annual Reviews Inc., 1987.

[HC78] G. E. Hughes and M. J. Cresswell. *An Introduction to Modal Logic*. Methuen, 1978.

[HC84] G. E. Hughes and M. J. Cresswell. *A Companion to Modal Logic.* Methuen, 1984.

[HF85] J. Y. Halpern and R. Fagin. A formal model of knowledge, action, and communication in distributed systems: preliminary report. In *Proc. 4th ACM Symp. on Principles of Distributed Computing*, pages 224–236, 1985.

[Hin62] J. Hintikka. *Knowledge and Belief.* Cornell University Press, 1962.

[HM85] J. Y. Halpern and Y. Moses. A guide to the modal logics of knowledge and belief. In *Ninth International Joint Conference on Artificial Intelligence (IJCAI-85)*, pages 480–490, 1985.

[HM90] J. Y. Halpern and Y. Moses. Knowledge and common knowledge in a distributed environment. *Journal of the ACM*, 37(3):549–587, 1990. An early version appeared in *Proceedings of the 3rd ACM Symposium on Principles of Distributed Computing*, 1984.

[HZ87] J. Y. Halpern and L. D. Zuck. A little knowledge goes a long way: Simple knowledge-based derivations and correctness proofs for a family of protocols. In *Proc. 6th ACM Symp. on Principles of Distributed Computing*, pages 269–280, 1987. A revised and expanded version appears as IBM Research Report RJ 5857, 1987 and will appear in *Journal of the ACM*.

[Kap89] D. Kaplan. Demonstratives. In J. Almog, J. Perry, and H. K. Wettstein, editors, *Themes from Kaplan.* Oxford University Press, New York, N.Y., 1989.

[Lad77] R. E. Ladner. The computational complexity of provability in systems of modal propositional logic. *SIAM Journal on Computing*, 6(3):467–480, 1977.

[Les89] Y. Lespérance. A formal account of self-knowledge and action. In *Eleventh International Joint Conference on Artificial Intelligence (IJCAI-89)*, pages 868–874, 1989.

[Les91] Y. Lespérance. *A Formal Theory of Indexical Knowledge and Action.* PhD thesis, University of Toronto, 1991.

[Lew79] D. Lewis. Attitudes *de dicto* and *de se*. *Philosophical review*, 88(4):513–543, 1979.

[Maz88] M. S. Mazer. A knowledge theoretic account of recovery in distributed systems: the case of negotiated commitment. In M. Y. Vardi, editor, *Proceedings of the Second Conference on Theoretical Aspects of Reasoning about Knowledge*, pages 309–324. Morgan Kaufmann, 1988.

[Moo85] R. C. Moore. A formal theory of knowledge and action. In J. Hobbs and R. C. Moore, editors, *Formal Theories of the Commonsense World*, pages 319–358. Ablex Publishing Corp., 1985.

[MR89] Y. Moses and G. Roth. On reliable message diffusion. In *Proc. 8th ACM Symp. on Principles of Distributed Computing*, 1989.

[MT88] Y. Moses and M. R. Tuttle. Programming simultaneous actions using common knowledge. *Algorithmica*, 3:121–169, 1988.

[NT87] G. Neiger and S. Toueg. Substituting for real time and common knowledge in asynchronous distributed systems. In *Proc. 6th ACM Symp. on Principles of Distributed Computing*, pages 281–293, 1987. To appear, *Journal of the ACM.*

[Per79] J. Perry. The problem of the essential indexical. *Noûs*, 13:3–21, 1979.

[RK86] S. J. Rosenschein and L. P. Kaelbling. The synthesis of digital machines with provable epistemic properties. In J. Y. Halpern, editor, *Theoretical Aspects of Reasoning about Knowledge: Proceedings of the 1986 Conference*, pages 83–97. Morgan Kaufmann, 1986.

[Rot89] G. Roth. Message diffusion in anonymous distributed systems. Master's thesis, Weizmann Institute of Science, 1989.

[Sea90] W. Seager. The logic of lost Lingens. *Journal of Philosophical Logic*, 19:407–428, 1990.