Ascribing beliefs to resource bounded agents

Natasha Alechina
School of Computer Science and IT
University of Nottingham
Nottingham NG8 1BB, UK
nza@cs.nott.ac.uk

Brian Logan
School of Computer Science and IT
University of Nottingham
Nottingham NG8 1BB, UK
bsl@cs.nott.ac.uk

ABSTRACT

Logical approaches to reasoning about agents often rely on idealisations about belief ascription and logical omniscience which make it difficult to apply the results obtained to real agents. In this paper, we show how to ascribe beliefs and an ability to reason in an arbitrary decidable logic to an agent in a computationally grounded way. We characterise those cases in which the assumption that an agent is logically omniscient in a given logic is 'harmless' in the sense that it does not lead to making incorrect predictions about the agent, and show that such an assumption is not harmless when our predictions have a temporal dimension: 'now the agent believes p', and the agent requires time to derive the consequences of its beliefs. We present a family of logics for reasoning about the beliefs of an agent which is a perfect reasoner in an arbitrary decidable logic L but only derives the consequences of its beliefs after some delay Δ. We investigate two members of this family in detail, L_Δ in which all the consequences are derived at the next tick of the clock, and L^*_Δ in which the agent adds at most one new belief to its set of beliefs at every tick of the clock, and show that these are sound, complete and decidable.

Categories and Subject Descriptors

I.2 [**Artificial Intelligence**]; F.3 [**Logics and Meanings of Programs**]

General Terms

Theory

Keywords

Formalisms and logics

1. INTRODUCTION

A major goal in intelligent agent research is the formal modelling of agent systems. Such an account is key both in deepening our understanding of the notion of agency, e.g., the relationships between agent architectures, environments and behaviour, and for the principled design of agent systems. A common approach is to model the agent in some logic and prove theorems about the agent's behaviour in that logic. It is perhaps most natural to reason about the behaviour of the agent in an epistemic logic, and there has been a considerable amount of work in this area, for example, [15, 8, 14, 18, 21, 9, 17, 28, 25, 30]. Epistemic notions such as knowledge and belief provide a compact and powerful way of reasoning about the structure and behaviour of agents [16]. Such an approach is useful as an abstraction tool even when we have perfect knowledge of the design of the system, but can also be applied when the system in question may not or is known not to employ intentional notions.

Approaches based on epistemic logic must address two main problems: the problem of correctly ascribing beliefs to the agent, and the problem of logical omniscience.

The problem of belief ascription is concerned with the difficulty of actually knowing what the agent's beliefs are. Many agent designs do not make use of an explicit representation of beliefs within the agent. For example, the behaviour of an agent may be controlled by a collection of decision rules or reactive behaviours which simply respond to the agent's current environment. However, when modelling such agents, it can still be useful to view them as having beliefs. For example, when modelling a behaviour-based agent we may say that "the agent believes there is an obstacle to the left" and "if the agent believes there is an obstacle to the left, it will turn to the right". However, for this to be possible, we need some principled way of deciding what the agent believes.

In such cases one approach is to view the agent as an *intentional system*, that is, we ascribe to it the beliefs and goals it *ought* to have, given what we know of its environment, sensors and (putative) desires. This approach, which Dennett [4, 5] calls "adopting the intentional stance", allows us to ascribe propositional attitudes to agent systems which do not explicitly represent beliefs, without having to know anything about the agent's state or architecture. In many cases this works tolerably well; for example, the predictions we can make by ascribing a belief that there is an obstacle to the left to a behaviour-based agent with an 'avoid obstacles' behaviour will be similar to the behaviour exhibited by the system. In other cases it is more problematic, largely due to the arbitrary nature of intentional attribution to such minimal intentional systems. Given only the agent's desires and its environment, we must assume some sort of design for the agent and work backwards to what sorts of events in the environment are significant, and hence the sorts of percepts and beliefs it 'ought' to have. The more we know about the design of an agent, e.g., what sorts of sensors it has, the easier it is to choose between alternative competing designs, and the sorts of beliefs the agent 'ought' to have.

The second problem is that of logical omniscience. The concept of logical omniscience was introduced by Hintikka in [11],

AAMAS'02, July 15-19, 2002, Bologna, Italy.

and is usually defined as the agent knowing all logical tautologies and all the consequences of its knowledge. Logical omniscience is problematic when attempting to build realistic models of agent behaviour, as closure under logical consequence implies that deliberation takes no time.

Most logical approaches to reasoning about agents rely on idealisations about belief ascription and logical omniscience which make it difficult to apply the results obtained to real agents. In many cases, belief ascription is not grounded in the state of the agent; rather beliefs are simply posited of the agent. Moreover, agents are typically modelled as logically omniscient, with the result that it is impossible to say when a real, resource bounded agent will hold a particular belief or even whether it ever does. For example, the influential Belief, Desire, Intention (BDI) framework of Georgeff and Rao [21] models agents as logically omniscient.

In this paper we present a new approach to modelling agents which addresses these problems. We distinguish between beliefs and reasoning abilities which we ascribe to the agent ('the agent's logic') and the logic we use to reason *about* the agent. In this we follow, e.g., [15, 12, 10]. Our approach grounds the ascription of belief in the state of the agent and allows us to explicitly model the computational delay involved in updating the agent's state. In the spirit of [29], we would like to design a logic to reason about the agent's beliefs which is grounded in a concrete computational model. However, unlike [29, 23] we choose not to interpret the agent's beliefs as propositions corresponding to sets of possible states or runs of the agent's program, but syntactically, as formulas 'translating' some particular configuration of variables in the agent's internal state. One of the consequences of this choice is that we avoid modelling the agent as logically omniscient. In representing beliefs syntactically and explicitly modelling computational delay in deriving consequences our approach has some similarities with the bounded-resources approach of [6].

In section 2 we motivate our approach. We develop a precise characterisation of those agents for whom the assumption of logical omniscience is harmless, in the sense that assuming the agent is logically omniscient does not lead to incorrect belief ascription. Then we show that for some agents (those performing some computations on their internal state) the logical omniscience assumption may lead to incorrect predictions concerning their beliefs. In section 3 we introduce a logic L_Δ which remedies the problem by introducing an explicit computational delay into the language of epistemic logic. The agent is still a perfect reasoner in the sense that it can derive all consequences of its beliefs (from a finite set of potential consequences) but this happens after a fixed delay. In section 4 we relax this assumption by introducing a logic L^*_Δ where each consequence (again from a fixed finite set) will be derived after a finite number of delays. Finally we define a family of logics 'between' L_Δ and L^*_Δ each with a different upper bound on the number of delays before all consequences are derived, sketch how those logics can be used to model agents at different levels of abstraction, and outline directions for further work.

2. BELIEF ASCRIPTION AND LOGICAL OMNISCIENCE

In this section we characterise situations when the assumption that an agent is logically omniscient is harmless in the sense that it does not lead to making incorrect predictions about the agent. We call an agent *logically omniscient* (in some logic) if the agent's beliefs are closed under logical consequence in that logic. For example, if the agent's beliefs are modelled in $S4$, the agent is logically omniscient since its beliefs are closed under the $S4$ consequence relation.

In this section we investigate the consequences of assuming that the agent is logically omniscient when reasoning about the agent's beliefs in an epistemic logic E. Identifying the problems resulting from this assumption for a certain kind of agent, we proceed in the next section to introduce an explicit temporal dimension in our epistemic logic and to make a distinction between the agent's 'internal logic' and the external epistemic logic logic in which we reason about the agent's beliefs .

We begin with a simple agent which we call $agent_1$, which we model as two functions and some internal state (see Figure 1).

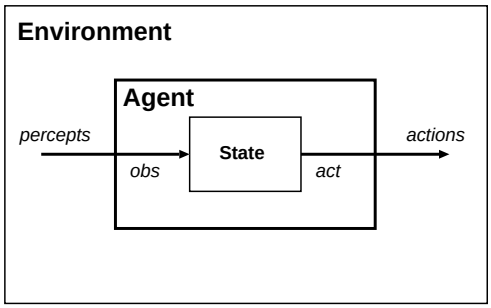

Figure 1: $\mathbf{Agent}_1$

The agent's state contains all the internal representations which determine the behaviour of the agent. Typically, some parts of the state can be interpreted as referring to (real or hypothetical) objects or events in the environment, e.g., that there is an obstacle dead ahead, or to properties of the agent itself, e.g., the level of the agent's battery. In such cases, changes in the environment or the agent will result in changes in the internal state of the agent in some more or less predictable way (depending on how noisy the agent's sensors are etc.)

The functions are the perception function, obs, which takes the percept(s) provided by the agent's sensors and the current state and returns a new state updated by the percepts. The second function is the action selection function, act, which takes the current state and returns an (external) action to perform. The perception function maps from events in the environment to representational states in the agent. The action selection function maps from states in the agent to events in the environment.

For the moment, we make two further assumptions: (I) that the perception function obs simply updates the relevant variables in the agent's state, e.g., it doesn't do any computation on the state such as belief revision or problem solving; and (II) that the actions don't modify the state of the agent.[1]

At each cycle the agent performs the computation: $act(obs(o, s))$, i.e., it updates its state s on the basis of its percepts o and chooses which action to perform on the basis of this new state. To correctly predict the evolution of the agent's beliefs and hence its actions, we need to be able to ascribe beliefs to the agent after its state has been updated by its percepts. In what follows we show how to ascribe beliefs to the agent in a computationally grounded way.

We begin by assuming that the agent's state does not change in the interval between perception and action selection. Suppose the state s is defined by the values of the variables $x_1, ..., x_k$, for example: the value of the temperature sensor variable x_t is 20, the value of the left collision detector x_l is 0 and the value of the right collision detector x_r is 1, etc. Based on those values, we can ascribe

[1]Below we relax some of the restrictions on $agent_1$.

beliefs about the external world to the agent: for example, based on $x_t = 20$ we ascribe to the agent a belief that the outside temperature is 20° C; based on $x_l = 0$ and $x_r = 1$ we ascribe the agent a belief that there is an obstacle to the right and so on. Note that this 'translation' is fixed and does not depend on the truth or falsity of the propositions in the real world.

Assuming that each variable can take only a finite number of values, we fix the set of atomic propositions which correspond to ascribable beliefs to be $\mathcal{P} = \{p_1, ..., p_m\}$. There is a mapping from the state s of the agent to set of propositions it 'believes' in: $P^+(s)$ assigns some propositions from $\mathcal{P}$ as the agent's positive beliefs, $P^+(s) \subseteq \mathcal{P}$. The mapping $P^-(s)$ assigns some propositions from $\mathcal{P}$ as the agent's negative beliefs (the agent believes the negation of the propositions), $P^-(s) \subseteq \mathcal{P}$. We assume $P^+(s) \cup P^-(s) \subseteq \mathcal{P}$ and $P^+(s) \cap P^-(s) = \emptyset$. This model of belief ascription (an agent only has beliefs in atomic formulas or their negations) is very basic. We consider it here not for its intrinsic interest but because it allows us to make a point concerning the logical omniscience assumption. In the remainder of this section we show that, under this model of belief ascription, the logical omniscience assumption is harmless, while if we extend it in a natural way to arbitrary formulas (e.g., by adding implications) we get potentially paradoxical results.

We reason about our model of the agent in a logic E containing a belief operator B. We denote the positive beliefs we ascribe to the agent as $Bel^+(s) = \{Bp : p \in P^+(s)\}$ and the agent's negative beliefs as $Bel^-(s) = \{B\neg p : p \in P^-(s)\}$. The set of ascribed beliefs is then $Bel(s) = Bel^+(s) \cup Bel^-(s)$. If we reason in say, S4, we can derive infinitely many more formulas about beliefs the agent holds: for example, $Bp_1 \vee Bp_2$, $Bp_1 \in Bel^+(s)$ (which many people would consider harmless), $Bp_1 \vee \neg Bp_1$ (tautology: harmless as well), BBp_1 (perhaps less intuitive), and if we don't restrict the language to just $p_1, ..., p_m$, we can derive some clearly irrelevant statements such as $B(p_1 \vee \phi)$ where ϕ says that the moon is made out of green cheese. Apart from some of those consequences being counterintuitive, can we say that they are really harmful in any precise sense?

One criterion would be: if we translate from the set of the derived consequences back into the statements about the agent's state, can we derive anything which does not agree with the actual state of the agent?

Let $P^+(s)$ include $p_1, \ldots, p_n$, and $P^-(s)$ include $\neg p_{n+1}, \ldots, \neg p_{n+k}$. Consider the set of formulas $Cons_E(Bel(s))$ ($Cons_E$ for short) which are all the consequences we can derive in E concerning the agent's beliefs at s. $Cons_E$ is closed under logical consequence in E and is consistent provided that E is *a reasonable modal logic of belief*.

Definition 1. E is *a reasonable modal logic of belief* if it is consistent and the modality B in E has the following properties:

1. $B\phi$ is interpreted in a way which depends only on the interpretation of propositional variables in ϕ and not any other propositional variables;
2. if ϕ is an atomic formula or a negation of an atomic formula, $B\phi$ can be true as well as false; and
3. Bp does not logically entail $B\neg p$ and vice versa.

Many epistemic logics including S4 and S5 fall under this definition. One can come up with definitions of modal logics which violate any of the conditions but the interpretation of B would be very far from the usual understanding of 'belief'. For example, the last condition would be violated if B were interpreted as 'has probability 1/2'.

Now we can formulate precisely what we mean by saying that the assumption that an agent is logically omniscient in E is harmless.

Definition 2. The assumption that an agent is logically omniscient in E is *harmless* if $Bp \in Cons_E(Bel(s))$ implies $Bp \in Bel^+(s)$ and $B\neg p \in Cons_E(Bel(s))$ implies $B\neg p \in Bel^-(s)$.

THEOREM 1. *For agents which only have beliefs in atomic formulas or their negations, the assumption of being logically omniscient in any reasonable modal logic of belief is harmless.*

PROOF. We need to prove that neither of the following consequences holds:

$$Bp_1, ..., Bp_n, B\neg p_{n+1}, ..., B\neg p_{n+k} \models_E Bp_i$$

unless $i \in \{1, \ldots, n\}$ and

$$Bp_1, ..., Bp_n, B\neg p_{n+1}, ..., B\neg p_{n+k} \models_E B\neg p_i$$

unless $i \in \{n+1, \ldots, n+k\}$. The first consequence would hold if for all interpretations of $p_1, \ldots, p_{n+k}$, if $Bp_1, \ldots, Bp_n$, $B\neg p_{n+1}, \ldots, B\neg p_{n+k}$ then Bp_i. If $i \notin \{n+1, \ldots, n+k\}$ the truth value of Bp_i is completely independent from the truth values of the premises by the first condition on reasonable modal logics. By the second condition there is an interpretation of $p_1, \ldots, p_{n+k}$ such that $Bp_1, \ldots, Bp_n, B\neg p_{n+1}, \ldots, B\neg p_{n+k}$ are true and Bp_i is false. If $i \in \{n+1, \ldots, n+k\}$ we still know (by the third condition) that there is an interpretation under which $B\neg p_i$ is true and Bp_i false. Hence the first consequence can't hold. A similar argument works for the second consequence. □

In the model of belief ascription described above, any agent which satisfies assumptions (I) and (II) can be viewed as logically omniscient in any reasonable logic of belief. For example, the simple reactive agents defined in [22] can be modelled as ideal reasoners without ascribing incorrect beliefs.

$Agent_1$ is still very simple. It might be termed a purely reactive agent in the sense that it will always do the same thing in the same situation (assuming obs is perfect and act is deterministic). Many agents do additional processing, for example, an agent may decide whether to 'believe' its percepts on the basis of its current state or it may derive consequences of its beliefs when combining inputs from several sensors. Similarly, when selecting an action to perform, the agent may deliberate about the merits of various possible actions or sequences or actions. In the simplest case, the agent may just record that it has performed an action a_i, so that it can try some other action if a_i doesn't have the desired effect. Such processing can be viewed as the result of *internal actions* which modify the agent's state. Agents with internal actions violate assumptions (I) and (II).

We therefore define a new agent, $agent_2$, which incorporates an additional inference step modelled by a belief update function inf which takes an agent state as argument and returns a new state (see Figure 2). At each cycle, $agent_2$ performs the computation: $act(inf(obs(o, s)))$, i.e., it updates its state s on the basis of its percepts o, derives any additional consequences of its new beliefs about its percepts and then selects an action to perform on the basis of this new state.

Note that $agent_2$ may believe different things at different times, since $Bel(inf(s)) \neq Bel(s)$. For example, after inf is evaluated the agent may start 'believing' an extra formula p. We need to model this ability in some way in the epistemic logic E. However, if we continue to only ascribe beliefs in atomic formulas or their

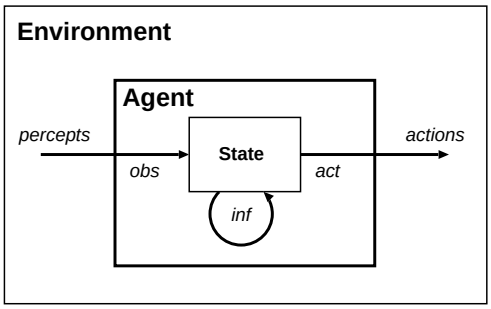

Figure 2: $\mathbf{Agent}_2$

negations to the agent, we cannot make any predictions concerning such new beliefs. If we expand $Bel(s)$ with some rules which allow us to derive new beliefs, e.g. $Bq \rightarrow Bp$, then we can derive Bp from Bq and Bp would be added to $Cons_E(Bel(s))$. However this would result in incorrect belief ascription at s, since p only appears in $Bel(inf(s))$. While beliefs in rules such as $Bq \rightarrow Bp$ cannot be grounded in the agent's state in the same way as beliefs in atomic propositions, in other work [1] we show how to ground beliefs in arbitrary formulas in the state of an agent.

The assumption that $agent_2$ is logically omniscient (or that $Bel(s)$ is closed under logical consequence in E) is not harmless. The evaluation of the belief update function inf will take some time, say Δ. If belief update starts at time t_0, then for times $t_0 < t < t_{0+\Delta}$, the assumption that the agent is logically omniscient will result in the ascription of incorrect beliefs to the agent, specifically those beliefs it has yet to derive.

To solve this problem, in the next section we make explicit a distinction between the agent's internal logic and the epistemic logic we use to reason about the agent's beliefs, and introduce an explicit temporal dimension to the latter logic.

3. A LOGIC FOR A PERFECT REASONER WITH A FIXED COMPUTATIONAL DELAY

In this section we consider a logic L_Δ for delayed belief. Strictly speaking, L_Δ is a family of logics parameterised by a decidable logic L which is the agent's internal logic, for example, classical propositional logic. (The material in this section and the next extends that in [2].) We assume there is a mapping from the belief update function inf to a logic L (set of inference rules) which mimics inf in the following way: if $inf(s) = s'$ then $Bel(s) \vdash_L Bel(s')$. The relation $\Phi \vdash_L \psi$ ('Φ implies ψ in L') between Φ and ψ is decidable if inf terminates after a fixed number of steps and there are fintely many agent states. The assumption that L exists and can be modelled this way is non-problematic if the agent really implements some (restricted) reasoning in a logic, as for example the agents considered in [12, 17]. If the agent is just an arbitrary program resetting some variables according to a set of instructions, this is more problematic, but still possible.

In L_Δ, after a fixed delay Δ the agent believes in all L-consequences of its beliefs at the previous step. The intuition underlying the notion of delayed belief is that an agent is able to draw all inferences from its beliefs, but needs time to derive them, so it does not believe the consequences *instantaneously*. The delay operator is not very expressive but it allows us to state some simple properties of a reasoner which requires some time to derive consequences, and nestings of delay operators can express the fact that some conclusions will be reached later than others: for example, $B\phi \wedge \Delta B(\phi \rightarrow \psi) \rightarrow \Delta\Delta B\psi$ is true if L contains modus ponens. In what follows we restrict the set of L-consequences to be a subset of some large but finite set $\mathcal{C}$ of 'potentially interesting' formulas. For example, if L is classical propositional logic and the language contains finitely many propositional variables, $\mathcal{C}$ could be all formulas in disjunctive normal form. We assume that the set of beliefs grows monotonically since we model just one application of the inf function.

The language L_Δ consists of a finite set $\mathcal{P}$ of propositional variables, $p_1, \ldots, p_m$, the usual boolean connectives $\neg, \wedge, \rightarrow, \ldots$ and two unary modalities: B which should be read as 'Believes' and Δ, standing for 'After a delay'.[2] A well formed formula is defined as usual: $p|\neg\phi|\phi \wedge \psi|\Delta\phi|B\phi$ with the proviso that Δ does not occur in ϕ in the clause for $B\phi$. We make this requirement since statements concerning the agent's computational delay do not correspond to anything in the agent's internal state and belong to the external language describing the agent's reasoning. From the technical point of view this restriction is not important and all the proofs below will go through if the syntax is not restricted. We denote the set of all well formed formulas as $Form$ and the set of all potentially interesting consequences the agent may derive as $\mathcal{C}$. We require $\mathcal{C}$ to be finite which may entail for example restricting the number of nestings of belief operators.

Definition 3. The models of L_Δ are structures of the form $M = \langle W, V, R, \delta\rangle$ where W is a non-empty set of possible worlds, $V : \mathcal{P} \longrightarrow 2^W$ assigns subsets of W to propositional variables, $R \subseteq W \times \mathcal{C}$ is a relation used to interpret B and $\delta : W \longrightarrow W$ is a function which is used to describe the next state of the world (after a delay) and interpret Δ. The satisfaction relation of a formula being true in a world in a model ($M, w \in W \models \phi$) is as follows:

$M, w \models p \Longleftrightarrow w \in V(p)$;

$M, w \models \neg\phi \Longleftrightarrow M, w \not\models \phi$;

$M, w \models \phi \wedge \psi \Longleftrightarrow M, w \models \phi$ and $M, w \models \psi$;

$M, w \models B\phi \Longleftrightarrow R(w, \phi)$;

$M, w \models \Delta\phi \Longleftrightarrow M, \delta(w) \models \phi$;

There are two conditions on δ:

Inclusion $\{\phi : R((w), \phi)\} \subseteq \{\phi : R(\delta(w), \phi)\}$.

Consequences For every $\psi \in \mathcal{C}$, if $\{\phi : R(w, \phi)\} \vdash_L \psi$, then $R(\delta(w), \psi)$.

If for any formula ϕ in L , $\phi \vdash_L \phi$, then the *Inclusion* condition follows from the *Consequences* condition. Although *Consequences* is a semantic condition on L_Δ models, it is formulated in terms of a syntactic relationship $\vdash_L$ which is the derivability relation in logic L. This is not problematic so long as derivability in L can be effectively established and as a result the property of being an L_Δ model is decidable.

The notions of L_Δ-valid and satisfiable formulas are standard: a formula ϕ is L_Δ-satisfiable if there exists an L_Δ-model M and a world w such that $M, w \models \phi$. A formula ϕ is L_Δ-valid ($\models \phi$) if all worlds in all models satisfy ϕ.

Consider the following axiom system (which we also refer to as L_Δ in the light of the completeness theorem which follows):

[2] Δ is in fact the same modality as the 'next' operator X of linear time temporal logic LTL [19].

A1 Classical propositional logic;

A2 $\Delta\phi \vee \Delta\neg\phi$

A3 $\neg\Delta(\phi \wedge \neg\phi)$

A4 $\Delta(\phi \rightarrow \psi) \rightarrow (\Delta\phi \rightarrow \Delta\psi)$

A5 $B\phi \rightarrow \Delta B\phi$

MP $\vdash \phi,\ \phi \rightarrow \psi \implies \vdash \psi$

R1 For all $\psi \in \mathcal{C}$, $\bigwedge_i \phi_i \vdash_L \psi \implies \vdash \bigwedge_i B\phi_i \rightarrow \Delta B\psi$

R2 $\vdash \phi \implies \vdash \Delta\phi$

We say that ϕ is derivable in L_Δ if there is a sequence of formulas $\phi_1, \ldots, \phi_n$, each of which is either an instance of an axiom schema from L_Δ or is obtained from the previous formulas using the inference rules of L_Δ, and $\phi_n = \phi$. A formula ϕ is L_Δ-consistent if $\phi \nvdash_{L_\Delta} \bot$.

It is natural to assume that $\phi \vdash_L \phi$ in which case **A5** becomes derivable by **R1**. Note that **A5** states that the agent never gives up its beliefs (its belief set grows monotonically). This is a strong assumption in itself, and in conjunction with additional belief axioms may lead to paradoxical results. In particular, if the belief operator satisfies the axioms of KD45 (5 being negative introspection, $\neg B\phi \rightarrow B\neg B\phi$) then also $\neg B\phi \rightarrow \Delta\neg B\phi$ is derivable.[3]

THEOREM 2. *L_Δ is complete and sound, namely* $\vdash_{L_\Delta} \phi \iff \models_{L_\Delta} \phi$

PROOF. First we give a proof of soundness: $\vdash_{L_\Delta} \phi \implies \models_{L_\Delta} \phi$. All instances of the axiom schemas are obviously valid. **A2-A4** state that after a delay the world is still a classical boolean universe. **A5** is valid because of the *Inclusion* condition.

Next we need to show that if the premises of the rules are valid, then the conclusions are. Rule **R1** expresses the main point of L_Δ: if an agent believes in $\bigwedge_i \phi_i$ and $\bigwedge_i \phi_i$ implies ψ in the agent's internal logic L, then after a delay the agent believes ψ. This follows from the condition on δ. **R2** states that after a delay all tautologies are still true.

Next we prove completeness: $\models_{L_\Delta} \phi \implies \vdash_{L_\Delta} \phi$. We show that for every ϕ if $\nvdash_{L_\Delta} \neg\phi$ then ϕ is satisfiable, that is $\not\models_{L_\Delta} \neg\phi$.

Assume that ϕ is an L_Δ-consistent formula. In a standard way, we can show that ϕ can be extended to a maximally consistent set of formulas w_ϕ, which is a consistent set closed under L_Δ derivability and containing either ψ or $\neg\psi$ for each $\psi \in Form$. We construct a canonical model M^c satisfying ϕ as follows:

W^c is the set of all maximally consistent sets;

$w \in V^c(p) \iff p \in w$;

$R^c(w, \psi) \iff B\psi \in w$;

$\delta^c(w) = \{\psi | \Delta\psi \in w\}$. In other words, $\forall\psi \in Form(\Delta\psi \in w \iff \psi \in \delta(w))$.

In order to complete the proof, we need to show:

Truth Lemma: for every $\psi \in Form$ and every $w \in W^c$, $M^c, w \models \psi \iff \psi \in w$.

Correctness of δ^c: for every $w \in W^c$, $\delta^c(w)$ is unique and is a maximally consistent set.

Inclusion For every $\psi \in Form$, if $R^c((w), \phi)$ then $R^c(\delta^c(w), \phi)$.

Consequences: For every $\psi \in Form$, if $\{\phi : R^c(w, \phi)\} \models_L \psi$, then $R^c(\delta(w), \phi)$.

From the Truth Lemma, it follows that ϕ is true in w_ϕ, hence ϕ is satisfiable.

The proofs of these statements are given below.

Truth lemma. The proof is standard and goes by induction on subformulas of ψ.

Correctness of δ^c. Consistency of $\delta^c(w)$ follows from **A3**. Maximality follows from **A2, A4** and **R2**. Uniqueness follows from the fact that each $w' \in W^c$ is unique.

Inclusion Suppose $R^c((w), \phi)$, then $B\phi \in w$. Hence by **A5** $\Delta B\phi \in w$ and by the definition of δ^c, $B\phi \in \delta^c(w)$. This implies $R^c(\delta^c(w), \phi)$.

Consequences Suppose $\psi \in \mathcal{C}$ and $\{\phi : R^c(w, \phi)\} \vdash_L \psi$. Therefore in L there exist $\phi_1, \ldots, \phi_n$ such that $\bigwedge_i \phi_i \models_L \psi$. By assumption $B\phi_i \in w$. By **R2**, $\Delta B\psi \in w$. By definition of δ^c, $B\psi \in \delta^c(w)$. Hence $R^c(\delta^c(w), \psi)$.

□

The logic L_Δ is decidable and has the bounded model property. Before proving this, we need a simple lemma. Below, $Subf(\phi)$ denotes the set of all subformulas of ϕ, and $ModSubf(\phi) = \{\psi \in Subf(\phi) : B\psi \in Subf(\phi)\}$ are modal subformulas of ϕ.

LEMMA 1. *For every $\phi \in Form$, and every two L_Δ models $M_1 = \langle W_1, V_1, R_1, \delta^1\rangle$ and $M_2 = \langle W_2, V_2, R_2, \delta^2\rangle$, if $W_1 = W_2$, $\delta_1 = \delta_2$, V_1 and V_2 agree on $p \in \mathcal{P} \cap Subf(\phi)$ and R_1 and R_2 agree on $\psi \in ModSubf(\phi)$, then for every w,*

$$M_1, w \models \phi \iff M_2, w \models \phi$$

PROOF. The proof is just a simple induction on subformulas of ϕ. □

Let us call the number of nestings of Δ operator in ϕ *Δ-depth of ϕ*, $d(\phi)$. More precisely,

$d(p) = 0$ for $p \in Prop$;

$d(\neg\psi) = d(\psi)$;

$d(B\psi) = d(\psi)$;

$d(\psi_1 \wedge \psi_2) = max(d(\psi_1), d(\psi_2))$;

$d(\Delta\psi) = d(\psi) + 1$.

Clearly $d(\phi) \leq |\phi|$ where $|\phi|$ is the size (number of subformulas) of ϕ. So the result below is better than usual results for modal logics obtained by filtrations which produce models of size less or equal to $2^{|\phi|}$.

THEOREM 3. *L_Δ has the bounded model property, that is, if a formula ϕ is satisfiable then it has a model where the cardinality of the set of worlds is less than or equal to $d(\phi)$ (hence less than or equal to $|\phi|$).*

[3] We are grateful to one of the anonymous referees for pointing this out.

PROOF. We can show that if a formula ϕ of Δ-depth $d(\phi) = k$ is satisfied in a world w of a model M then it is satisfied in a model M' where the set of worlds contains only w and the worlds reachable from w in k δ-steps, i.e., $W' = \{w, \delta(w), \delta(\delta(w)), \ldots, \delta^k(w)\}$. Obviously W' is of size at most $d(\phi)$ even if W is infinite ($|W'|$ could be less than k if for some $m < k, \delta^m(w) = \delta^{m+1}(w)$).

The proof that $M, w \models \phi \iff M', w \models \phi$ is standard (see for example [27], Lemma 2.8) and is omitted here. □

THEOREM 4. *The satisfiability problem for L_Δ is decidable.*

PROOF. Suppose we would like to check whether a formula ϕ is satisfiable in L_Δ. By the previous theorem, it suffices to check whether ϕ is satisfiable in any L_Δ model of size less or equal to $|\phi|$. The set of models of size less or equal to $|\phi|$ is strictly speaking infinite since R is defined on the set of all formulas which is infinite, so there are infinitely many models of a fixed finite size which differ in R. However, by lemma 1 the only part of R in every model which really matters for checking whether ϕ is satisfied or not is the part dealing with all subformulas of ϕ of the form $B\psi$. There are only finitely many different relations R with respect to the set $ModSubf(\phi)$, so we need to check only finitely many cases. Being an L_Δ model is a decidable property; in particular checking whether the *Consequences* condition holds is decidable given that L is decidable. □

An important property for an epistemic logic is also the complexity of model checking: is there an efficient procedure to establish whether a certain formula holds in a given model. The following theorem shows that L_Δ is extremely efficient in this respect:

THEOREM 5. *Given a formula ϕ and a pair M, w (a model and a world), there is is an algorithm which checks whether $M, w \models_{L_\Delta} \phi$ which is in O($|\phi|$).*

PROOF. Assume that $d(\phi) = k$; recall that $|\phi| \geq k$. Each subformula of ϕ is at some depth i where $0 \leq i \leq k$ and should be evaluated relative to $\delta^i(w)$. We start at the subformulas at depth k and replace propositional variables which are true in $\delta^k(w)$ by $\top$, the ones which are false by $\bot$, formulas of the form $B\psi$ by $\top$ if $R(\delta^k(w), \psi)$ and by $\bot$ otherwise. Evaluate the resulting formula using propositional logic and replace all subformulas at depth k by $\top$ or $\bot$. Remove the innermost Δ operator. Now we have a new formula of depth $k-1$. repeat until there are no occurences of Δ and evaluate the resulting formula. The process is obviously linear in the length of ϕ. □

4. A LOGIC FOR A MORE REALISTIC REASONER

In L_Δ, the agent is a perfect reasoner in L: after a fixed delay Δ, it derives all the L-consequences of its beliefs. Here we introduce a slightly more realistic logic which describes an agent deriving L-consequences one at a time, although it is still guaranteed to derive each consequence after some finite sequence of delays. We call this logic L^*_Δ. The language of L^*_Δ is expanded by an extra modality Δ^* which stands for a finite (possibly empty) sequence of delays. It is interpreted as a reflexive transitive closure of the delay relation $\{(w, v) : v = \delta(w)\}$. The rest of the language is the same as the language of L_Δ; if ϕ is a formula, then $\Delta^*\phi$ is also a formula.

Definition 4. The models of L^*_Δ are the same as the models of L_Δ, with an additional clause defining the truth conditions of the Δ^* operator:

$M, w \models \Delta^*\phi \iff M, \delta(\delta(....\delta(w)..) \models \phi$, that is, after finitely many (possibly 0) applications of δ we reach a state where ϕ is true.

The condition on δ called *Consequences* which held in L_Δ models does not hold for L^*_Δ models. Instead we have

Finiteness $\{\phi : R((w), \phi)\}$ is always finite.

Inclusion $\{\phi : R((w), \phi)\} \subseteq \{\phi : R(\delta(w), \phi)\}$.

Uniqueness $\{\phi : R((w), \phi)\}$ and $\{\phi : R(\delta(w), \phi)\}$ differ in at most one formula.

Eventually-Consequences If $\psi \in \mathcal{C}$ and $\{\phi : R((w), \phi)\} \vdash_L \psi$ then there exists an $n \geq 0$ such that $\psi \in \{\phi : R(\delta^n(w), \phi)\}$.

Note that $\Delta^*\phi$ means $(\Delta^n\phi)$ for some non-negative n, so it is essentially an existential modality. Its dual $[\Delta^*]$ which means $\forall n \geq 0\, (\Delta^n\phi)$ can be defined as $\neg\Delta^*\neg\phi$. To give an axiomatisation, it is helpful to think of Δ and Δ^* as PDL (Propositional Dynamic Logic, see [20]) modalities $\langle\Delta\rangle$ and $\langle\Delta^*\rangle$, with the additional property that $\langle\Delta\rangle$ is the same as $[\Delta]$ because it is interpreted by a function (δ) rather than a relation.

THEOREM 6. *The following axiom system is weakly sound and complete for L^*_Δ:*

A1 Classical propositional logic;

A2 $\Delta\phi \vee \Delta\neg\phi$

A3 $\neg\Delta(\phi \wedge \neg\phi)$

A4a $\Delta(\phi \to \psi) \to (\Delta\phi \to \Delta\psi)$

A4b $[\Delta^*](\phi \to \psi) \to ([\Delta^*]\phi \to [\Delta^*]\psi)$

A5' $B\phi \to [\Delta^*]B\phi$

A6 $\Delta B\phi \wedge \Delta B\psi \to B\phi \vee B\psi \;\; (\phi \neq \psi)$

A7 $\Delta^*\phi \leftrightarrow \phi \vee \Delta\Delta^*\phi$

A8 $[\Delta^*](\phi \to \Delta\phi) \to (\phi \to [\Delta^*]\phi)$

MP $\vdash \phi,\ \phi \to \psi \implies \vdash \psi$

R1' For all $\psi \in \mathcal{C}$, $\bigwedge_i \phi_i \vdash_L \psi \implies \vdash \bigwedge_i B\phi_i \to \Delta^*B\psi$

R2' $\vdash \phi \implies \vdash [\Delta^*]\phi$

PROOF. The proof of soundness is straightforward and is omitted here. Observe that **A5'** replaces **A5** of L_Δ and **R1', R2'** replace **R1, R2** which become derivable since $[\Delta^*]\phi \to \Delta\phi$ is derivable. Axiom **A6** corresponds to the *Uniqueness* condition. The axioms **A7** and **A8** axiomatise the transitive closure relation, see [24, 13].

The completeness proof is based on the completeness proof for PDL given in [3] which in turn is based on [26]. We are going to show that any L^*_Δ-consistent formula ϕ has a model.

First we define a *closure* of a set of formulas Σ, $Cl(\Sigma)$ as the smallest set containing Σ and closed under subformulas and the following conditions:

if $\Delta^*\phi \in Cl(\Sigma)$ then $\Delta\Delta^*\phi \in Cl(\Sigma)$

if $\phi \in Cl(\Sigma)$ then $\sim\phi \in Cl(\Sigma)$ where $\sim\phi$ is ψ if $\phi = \neg\psi$ and $\neg\phi$ otherwise.

Note that $Cl(\Sigma)$ is finite if Σ is finite.

Then we define a set of atoms over Σ, $At(\Sigma)$, as the set of all maximally consistent subsets of $Cl(\Sigma)$. It can be shown that if ϕ is consistent then there is an atom $A \in At(Cl(\{\phi\} \cup B\mathcal{C}))$ such that $\phi \in A$, where $B\mathcal{C}$ is the set of all possible consequences prefixed by a belief operator. Finally we build a model $M = \langle W, V, R, \delta \rangle$ where

$W = At(Cl(\{\phi\} \cup B\mathcal{C}))$;

For every $p \in \mathcal{P}$ and $A \in At(Cl(\{\phi\} \cup B\mathcal{C}))$, $A \in V(p)$ iff $p \in A$;

For every ϕ and $A \in At(Cl(\{\phi\} \cup B\mathcal{C}))$, $R(A, \phi)$ iff $B\phi \in A$;

For every $A, B \in At(Cl(\{\phi\} \cup B\mathcal{C}))$, $B = \delta(A)$ if $\hat{A} \wedge \Delta\hat{B}$ is consistent, where $\hat{A} = \bigwedge_{\phi \in A} \phi$.

Then we need to show that δ so defined is indeed a function (if $\hat{A} \wedge \Delta\hat{B}$ and $\hat{A} \wedge \Delta\hat{C}$ are consistent, then $B = C$). This is easy since in L^*_Δ (and L_Δ) $\Delta\phi \wedge \Delta\neg\phi$ is inconsistent.

We also need to show that other conditions on R and δ hold and that the *Truth lemma* holds. The only really difficult part of the proof of the *Truth lemma* is showing that

$\Delta^*\phi \in A$ iff there exists an atom B such that $\phi \in B$ and there is a sequence of atoms C_0,...,C_n such that $C_0 = A$, $C_n = B$ and either $n = 0$ or $C_{i+1} = \delta(C_i)$.

The proof of this is identical to the proof in [3] for arbitrary PDL modalities $\langle \pi^* \rangle$.

Of the remaining conditions, *Finiteness* follows from construction (the atoms are finite hence the number of beliefs associated with each atom by construction of R is finite as well). *Uniqueness* follows easily from **A6** and the definition of δ. *Inclusion* follows from **A5'** and the definition of δ. *Eventually-Consequences* is the only slightly non-trivial property. Assume that $\{\phi : R(A, \phi)\} \vdash_L \psi$. We want to show that there is an atom B reachable by a set of δ-steps such that $R(B, \psi)$. In the completeness proof for L_Δ it sufficed to show that there is a δ-reachable possible world consistent with $B\psi$ to conclude that it contains $B\psi$. However, in our model the atoms are finite and do not contain all formulas they are consistent with. However we constructed the model using the atoms not just over $Cl(\{\phi\})$, but over $Cl(\{\phi\} \cup B\mathcal{C})$, so $B\psi$ is guaranteed to belong to an atom it is consistent with. We show that there is a δ-path to such an atom from A using **R1'**. □

THEOREM 7. *Satisfiability in L^*_Δ is decidable.*

PROOF. We have shown in the previous theorem that for every consistent formula ϕ we can build a finite model the size of which depends on ϕ and C. □

THEOREM 8. *Given a formula ϕ and a model, state pair M, w there is an O($|M| \times |\phi|$) algorithm for checking whether $M, w \models_{L^*_\Delta} \phi$.*

PROOF. Since L^*_Δ is a variant of PDL, this follows from the result on complexity for model checking for PDL ([7]). □

There are infinitely many logics between L^*_Δ and L_Δ depending on how many or which beliefs are added at each δ step. Intuitively, L^*_Δ corresponds to the most low-level view of the agent (although we don't say in which order the formulas are derived, in principle we can specify this). If we prefer to think of the agent on a higher level of abstraction we can specify, for example, in which order inference rules are going to be applied and, at each timestep, add all the formulas derivable by one application of some particular rule. Alternatively, we can specify (in L^*_Δ) an upper bound n on the number of delays after which each consequence will be derived:

$$\bigwedge_i \phi_i \vdash_L \psi \implies \vdash \bigwedge_i B\phi_i \rightarrow \Delta^n B\psi$$

Given the above, we have an upper bound on the number of delays to omniscience. We can use this to define a notion of an *effectively omniscient* agent. If we assume that the agent must derive all consequences of its beliefs to be sure of choosing the correct action to perform, then the agent is effectively omniscient if the number of delays required for omniscience is less than the rate at which the environment changes. If the agent requires more applications of Δ, then either it must abandon those inferences it has managed to draw and start over from its new beliefs (conservative strategy) or risk that some of its derived beliefs are based on out of date information (optimistic strategy).

The modularity of L^*_Δ (the fact that we can substitute any decidable logic for L) means that we can use different logics to model different phases of the agent's processing. Many agent designs organise processing into layers or phases, for example, layering of behaviours in a subsumption architecture or the grouping of rules into sets concerned with perceptual processing, planning and so on in a rule-based architecture. We can model this organisation as a series of logics, $L_1, \ldots, L_k$, to allow finer-grained ascription of beliefs to the agent. For each logic L_i, we can impose an upper bound on the number of delays required to derive all consequences in this phase. Assuming the processing of the L_i is sequential, we can simply sum the delays for each phase to give the overall upper bound on deriving all consequences.

5. CONCLUSIONS

In this paper, we have investigated ascribing beliefs to an agent based on the values of the variables constituting the agent's internal state. We looked at the consequences of assuming that the agent not just has the given beliefs but is also logically omniscient in an epistemic logic E. We have shown that although this assumption may result in counterintuitive consequences (such as the agent believing all tautologies etc.) in a certain precise sense this assumption may be harmless: namely, when translated back into the agent's 'internal language' (involving only the values of the agent's state variables) the extra derived beliefs don't amount to anything more than what is already in the agent's state. We characterise the kinds of agents where this is the case (the logical omniscience assumption is harmless). We also showed that if the agent is able to update its internal state (e.g., by revising its beliefs), then the logical omniscience assumption may not be harmless as we ascribe to the agent beliefs which it may not have derived. To remedy this, we proposed a family of logics which explicitly model computational delay. The logic L_Δ describes an agent which is a perfect reasoner in an arbitrary logic L (for example, S4, or classical propositional logic) which explicitly models the computational delay in reasoning about the agent's beliefs. For any decidable logic L, L_Δ (parameterised by L) has a complete and sound axiomatisation and is decidable. The logic L^*_Δ models a more realistic agent which derives at most one consequence at each computational step—each consequence from a specified finite set $\mathcal{C}$ is guaranteed to be derivable after a finite sequence of delays. L^*_Δ also has a complete and sound axiomatisation and is decidable.

The approach we have presented can be used to model the agent at different levels of abstraction. Propositions can be supervenient on complex patterns of state variables and the rules used to model

the belief update function can be at any level of abstraction. For example, the proposition "colliding with obstacle", p_b, could be true if any of the agent's bump switches are closed (represented by one of the variables $x_i, \ldots, x_k$ having the value 1, say). In future work, we hope to extend our approach to explicitly allow modelling at multiple levels of abstraction, from fine grained operations on the variables comprising the agent's state to more coarse grained models based on intentional notions such as beliefs, desires and intentions, with each level of abstraction grounded in the one below and ultimately in the agent's state.

6. REFERENCES

[1] N. Alechina and B. Logan. Grounding knowledge and action selection in agent-based systems. In *Proceedings of the Workshop on Logics for Agent-Based Systems (LABS)*.

[2] N. Alechina and B. Logan. Logical omniscience and the cost of deliberation. In R. Nieuwenhuis and A. Voronkov, editors, *Proceedings of the 8th International Conference on Logic for Programming, Artificial Intelligence and Reasoning (LPAR'01)*, LNAI No. 2250, pages 100–109. Springer Verlag, 2001.

[3] P. Blackburn, M. de Rijke, and Y. Venema. *Modal Logic*, volume 53 of *Cambridge Tracts in Theoretical Computer Science*. Cambridge University Press, 2001.

[4] D. C. Dennett. *The Intentional Stance*. MIT Press, 1987.

[5] D. C. Dennett. *Kinds of Minds: Towards an understanding of consciousness*. Basic Books, 1996.

[6] J. J. Elgot-Drapkin and D. Perlis. Reasoning situated in time I: Basic concepts. *Journal of Experimental and Theoretical Artificial Intelligence*, 2:75–98, 1990.

[7] E. Emerson and C.-L. Lei. Efficient model checking in fragments of the propositional mu-calculus. In *Proceedings LICS'86*, pages 267–278, 1986.

[8] R. Fagin and J. Y. Halpern. Belief, awareness, and limited reasoning. In *Proceedings of the Ninth International Joint Conference on Artificial Intelligence (IJCAI-85)*, pages 480–490, 1985.

[9] R. Fagin, J. Y. Halpern, Y. Moses, and M. Y. Vardi. *Reasoning about Knowledge*. MIT Press, Cambridge, Mass., 1995.

[10] R. Fagin, J. Y. Halpern, and M. Y. Vardi. A non-standard approach to the logical omniscience problem. *Artificial Intelligence*, 79(2):203–240, 1996.

[11] J. Hintikka. *Knowledge and belief*. Cornell University Press, Ithaca, NY, 1962.

[12] K. Konolige. *A Deduction Model of Belief*. Morgan Kaufmann, San Francisco, Calif., 1986.

[13] D. Kozen and R. Parikh. An elementary proof of the completeness of PDL. *Theoretical Computer Science*, 14:113–118, 1981.

[14] G. Lakemeyer. Steps towards a first-order logic of explicit and implict belief. In J. Y. Halpern, editor, *Theoretical Aspects of Reasoning About Knowledge: Proceedings of the 1986 Conference*, pages 325–340, San Francisco, Calif., 1986. Morgan Kaufmann.

[15] H. J. Levesque. A logic of implicit and explicit belief. In *Proceedings of the Fourth National Conference on Artificial Intelligence, AAAI-84*, pages 198–202. AAAI, 1984.

[16] J. McCarthy. Ascribing mental qualities to machines. Technical report, Stanford AI Lab, 1978.

[17] R. C. Moore. *Logic and Representation*. Number 39 in CSLI Lecture Notes. CSLI Publications, 1995.

[18] R. Parikh. Knowledge and the problem of logical omniscience. In *Methodologies for Intelligent Systems, Proceedings of the Second International Symposium*, pages 432–439. North-Holland, 1987.

[19] A. Pnueli. A temporal logic of concurrent programs. *Theoretical Computer Science*, 13:45–60, 1981.

[20] V. R. Pratt. Semantical considerations on Floyd-Hoare logic. In *Proceedings of the Seventeenth IEEE Symposium on Computer Science*, pages 109–121, 1976.

[21] A. S. Rao and M. P. Georgeff. Modeling rational agents within a BDI-architecture. In *Proceedings of the Second International Conference on Principles of Knowledge Representation and Reasoning (KR'91)*, pages 473–484, 1991.

[22] S. Russell and P. Norvig. *Artificial Intelligence: a modern approach*. Prentice Hall, 1995.

[23] N. Seel. The 'logical omniscience' of reactive systems. In *Proceedings of the Eighth Conference of the Society for the Study of Artificial Intelligence and Simulation of Behaviour (AISB'91)*, pages 62–71, Leeds, England, 1991.

[24] K. Segerberg. A completeness theorem in the modal logic of programs. *Notices of the American Mathematical Society*, 24:A–552, 1977.

[25] M. P. Singh. Know-how. In M. Wooldridge and A. Rao, editors, *Foundations of Rational Agency*, pages 81–104. Kluwer Academic, Dordrecht, 1999.

[26] J. van Bentham and W. Meyer Viol. Logical semantics of programming. Unpublished lecture notes, 1993.

[27] J. van Benthem. *Modal logic and classical logic*. Bibliopolis, 1983.

[28] W. van der Hoek, B. van Linder, and J.-J. C. Meyer. An integrated modal approach to rational agents. In M. Wooldridge and A. Rao, editors, *Foundations of Rational Agency*, pages 133–168. Kluwer Academic, Dordrecht, 1999.

[29] M. Wooldridge. Computaionally grounded theories of agency. In E. Durfee, editor, *Proceedings of the Fourth International Conference on Multi-Agent Systems (ICMAS-2000)*, pages 13–20. IEEE Press, 2000.

[30] M. Wooldridge and A. Lomuscio. A computationally grounded logic of visibility, perception, and knowledge. *Logic Journal of the IGPL*, 9(2):273–288, 2001.