Time for Action:
On the relation between time, knowledge and action

Yoav Shoham*
Computer Science Department
Stanford University

Abstract

We consider the role played by the concept of action in AI. We first briefly summarize the advantages and limitations of past approaches to taking the concept as primitive, as embodied in the situation calculus and dynamic logic. We also briefly summarize the alternative, namely adopting a temporal framework, and point out its complementary advantages and limitations. We then propose a framework that retains the advantages of both viewpoints, and that ties the notion of action closely to that of knowledge. Specifically, we propose starting with the notion of time lines, and defining the notion of action as the ability to make certain choices among sets of time lines. Our definitions shed new light on the connection between time, action, knowledge and ignorance, choice-making, feasibility, and simultaneous reasoning about the same events at different levels of detail.

1 Introduction

One has strong intuitions about the concept of action in AI, and about its importance in common sense reasoning. Indeed, some formal systems take the concept of action as primitive. Most notable among them in the situation calculus [McCarthy and Hayes, 1981]. As is well known, the situation calculus views the world as consisting of states which are not only connected, but in fact defined, by actions. Specifically, all states are defined by the sequence of actions that led to them from the initial state. For example, given an initial state in which the car engine is off, the action of starting the car defines a new state in which the motor is running.

This view of the world is not an AI idiosyncracy. For example, it is exactly the view of the world within dynamic logic [Pratt, 1976], although there actions are called programs. Indeed, this approach has much intuitive appeal. It captures oui intuition that there are agents who can *choose* to act in one way or another, and the way we reason about the world is by making, hypothesizing or observing these decisions, and computing their ramifications. For example, it is natural to view turn-camera-head as an action, since we can make the fact that the camera head turned true by simply deciding so.

Along with its advantages, however, this action-based approach has severe limitations. In [Shoham and Goyal, 1988] we discussed the general limitations of what were termed *change-bused systems,* those that adopt a 'change indicator' (action, program) as primitive, and have no notion of time other than that implicit in the sequencing of the change indicators. Briefly, the limitations include the inability to represent duration and dates, to represent the effect of concurrent actions (which may bear no interesting relation to the effect of any one action in isolation), to represent more complex temporal relations between actions such as overlapping, and to represent effects that do not follow the action immediately or that last only a bounded amount of time. These limitations are the primary reason dynamic logic has not had a lasting effect on computer science, even though it is a most elegant formalism and attracted much attention initially. From the AI point of view, perhaps the most striking limitation is the inability to integrate the notion of action with that of naturally-occurring processes, such as chemical reactions or the behavior of physical devices, a capability clearly needed in, e.g., planning applications.

The alternative is to explicitly introduce the concept of time. One starts with a temporal structure, and only then states what is the case at different times within that structure. There is sometimes in AI the misconception that this necessarily makes temporal logics inherently complex. However, the structure of time may be very simple (for example time can be assumed to be isomorphic to the integers), in which case the logic is no more complex than, e.g, the situation calculus. The important point is that one *first* lays out the structure of time, and only then describes the world within that structure.

As is discussed in [Shoham and Goyal, 1988], the transition to such temporal logics solves the expressiveness problems associated with the change-based approach. Since one can speak about time, it is easy to represent overlapping events, deadlines, delayed effects, and so on. This is the main reason that temporal logics, which were introduced to computer science one year after dynamic logic was [Pnueli, 1977], have spawned a tremendous industry since the initial publication. It is arguably also

'The author was supported by NSF grant IRI-8721701, IBM grant 1220042, and & gift from DEC Corporation.

the reason that we, people, have the mythical notion of time ingrained so deeply into our conceptualization of the world, even though it is *change* that our perceptual apparatus is sensitive to (witness for example our ability to detect motion in the periphery of our visual field where we cannot see motionless objects, or our ability to detect a change in pitch even though we may be very bad at identifying any single given pitch).

However, while the move to a temporal framework solves the representational problems with the action-based view, it also loses the very advantage of the latter. We no longer capture the intuition that there are agents in the world that may influence the course of events; all we have is a lackluster flow of history.

Our loss is best illustrated when we try to justify keeping the notion of action in our vocabulary. In the change-based framework its role is clear: as was discussed, actions define future states. In a temporal logic action is not needed for that purpose, since we start with a temporal structure in which the future is already defined, even before have said anything about what is true and false in it.

It is tempting to say that the role of actions is to determine what is true in that future: if I perform the action of turning the ignition key at time t, then at time $t + c$ the engine is running. Indeed, several logics in AI (such asMcDermott's [McDermott, 1982], Allen's [Allen, 1984], Georgeff and Lansky's [Georgeff and Lansky, 1985], and Kowalski and Sergot's [Kowalski and Sergot, 1986]) have notions of action along these lines.

The point to realize is that this view of action, as something that merely takes place in time and that implies something about the future, makes the very concept of action unnecessary. One can replace an action by a proposition representing the same fact, and replace the notion of an action having effects by standard implication. For example, to capture the fact that turning the ignition key has the effect of the motor being on, it is sufficient to write the implication TRUE(t,turn-key) $\supset$ TRUE($t+\epsilon$,engine-running) in a suitable temporal logic. Since such implications are needed in the language anyway, why add a seperate notion of action?[1]

Clearly, if the notion of action buys us nothing, parsimony dictates that we drop it. Yet one has strong intuitions that the concept of action is important in AI, even if so far that importance has not been formally captured. In the following I propose a formal role for action in a temporal setting, which, surprisingly, ties it closely to the notions of knowledge and ignorance. To explain this connection I start in the next section with some general thoughts on action, and then go on to fill in the details.

2 Conceptual preliminaries: action

It was mentioned in the previous section that an attraction of the action-based approach has to do with the connection between action and choice making. In particular, one has the intuition that an agent's taking an action is associated with his making a certain choice among several courses of events. Of course, one can easily get into controversial cases in which agents might be said to take an action even if they really cannot help it (see, e.g., [Goldman, 1970]), but we need not do that in this preliminary discussion. What is important is that, although its precise nature is still imperfectly understood, in prototypical cases there does seem to be a connection between action and choice making. Furthermore, this choice seems inherently asymmetric in time - whether or not an action is possible depends only on the past, and it can affect only the future.

There is another property of action that is related to its "free will" aspect, which is perhaps less obvious. Consider an animal performing a simple action such as raising its leg. Suppose now that modern neuroscience has advanced to a point where we can completely map the neuromuscular activity leading the contraction of the leg muscles. In fact, suppose it is *us* who supplied the stimulus that resulted in the raising of the leg in the first place. In such a situation we would no longer be inclined to say that the animal performed an action.

The reader may recognize here Dennett's argument in [Dennett, 1984] that "free will is in the mind of the beholder": We do not ascribe free will to an agent if we can invariably predict his behavior. Carrying the idea slightly further, I claim that action too is in the mind of the beholder - what one observer might call an action, another more knowledgeable observer might describe as an agentless process.

This subjective view of action may seem at first a bit radical, but really there is much evidence for it. Consider, for example, Piaget's theory how children's explanations of the physical world evolve over time [Piagct, 1951]. Without arguing about whether it makes sense to identify precisely seventeen stages of development, it is clear from his data that at first children make much use of the notion of action and actors when explaining the world, until the age of about nine when they know enough to use standard physical explanations. Thus before they get to the age at which they say that it is the wind that causes the clouds to move, they go through an animistic stage at which they report that clouds are agents that follow people around.

This is an example of a single phenomenon (movement of clouds) explained differently by observers with varying degrees of knowledge. It is also easy to come up with examples in which a single observer reasons about phenomena about which she has different degrees of knowledge. Consider our reasoning about a light switch and our reasoning about our dog. For all intents and purposes, we understand the workings of the light switch completely. We therefore view it as governed by a fixed transition function, "causal rules" if you will, rather than as a agent who receives our request to transfer current and decides to grant it.

In the case of the dog, on the other hand, we are in a much more ignorant situation. We know only a fraction of the rules governing the dog's behavior: his transition function is just too complex. Some are tempted to say that indeed no such transition function exists, since the

[1]This is not an argument against the merit of the above-mentioned logics, but about the contribution of the notion of action to that merit.

Figure 1: A future-branching structure

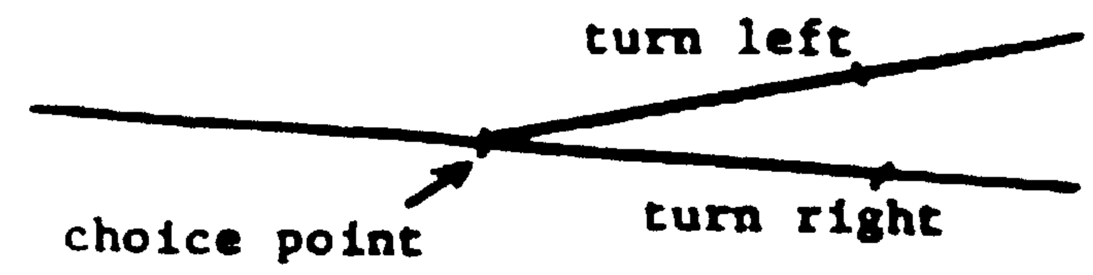

Figure 2: A simple choice

dog has a certain amount of "free will." As Dennett observes, the distinction between a transition function that is too immense to fathom, and the lack of a deterministic function in the first place, is not an illuminating one.

Our robots fall somewhere in between dogs and light switches. On the one hand we understand well the workings of their individual components; after all, we built them. On the other hand, the entire robot's behavior is a complex composition of the function of the individual components, a global behavior that is extremely hard to describe succintly or to control.

One therefore finds both views on robots. Some, typically those closer to the hardware side and working on very circumscribed applications, have no use for the concept of action, but instead use the machinery that describes aspects of the transition function such as electronic circuits and control theory. Others, typically those working in more abstract settings and on very ambitions applications such as general planning, use the concept of action quite heavily.

It is clearly important to reconcile these two views of our robots. For that we must explicate the information hidden in the notion of action. I have already given the intuition behind the two main properties of of action: a choice among possibilities, and the perspective of observers. The question now is how these properties can be captured naturally in our notation.

3 Notational preliminaries: choice-making and knowledge

Let us start by considering choice making. The need for representing choice immediately suggests using a branching-time framework.

3.1 Branching time

One of the more influential temporal logics in AI, due to McDermott, is indeed based on a future-branching structure [McDermott, 1982]. Time lines (or, as they are called there, *chronicles)* may diverge into the future. Once they do split, two chronicles never meet again. This induces a forest-like structure of the form shown in Figure 1.

However, McDermott's motivation for adopting future branching is different from ours. For him the branching represents true indeterminacy in the world, having nothing to do with actions of agents; Actions of agents are associated with intervals within a time line, and not with the branching aspect of the structure. As I have already said, the resulting definition of action does not capture our intuition, and makes the concept unnecessary from the technical point of view.

We get closer to our goal if we retain the branching structure, but associate actions with the branching points. Specifically, we may assume that at every point in the structure, each agent may may select certain subsets among the set of all futures.

For example, consider the simple structure in Figure 2. The two future branches differ on whether the robot turned left or right at the branching point. We can view the robot as having control over his movements, and thus being able to select among these possibilities. The action of "turning left" is defined to be a selection of the branch in which the robot indeed turned left. Notice, however, that the different branches are "synchronized" - they do not describe time differently, only what is true and false in it. I will therefore make the distinction between *dates,* which are common to all paths, and *-points* or *situations,* which make up the different paths.

Here we described the selection as resulting in a single branch, but in general it may result in some subset of the available futures. Thus there exist actions that can only be carried out jointly by several agents. Consider, for example, the action of playing tennis. Each of the two agents can restrict the set of futures to those in which he runs around and hits tennis balls, if any balls come his way. It is only the intersection of those two subsets of futures that yields a tennis game.

This view of action does not prohibit backward branching. Indeed, in many cases it is useful to allow different courses of events to lead to identical situations. I will therefore allow this possibility. I will not, however, allow paths that diverge into the future to meet later on (although for some applications even this would be a sensible thing to allow). The resulting topology is illustrated in Figure 3.

It remains to be explained what the difference is between the future branching and past branching. At first it might seem that an agent can only reduce the future branching. However, in a certain sense, he can also reduce the backward branching. This is exactly what an experiment is about: before the experiment there are several possible pasts, for example one in which the tested substance is (and has been) an acid and another in which it is a base, and after the experiment only one past remains. Of course, the action did not change the

At first glance, this view of action might appear troublesome. After all, if every action is taken in a particular time line, the future is already determined, and the action cannot really influence it. However, a small shift in perspective takes care of this worry. It relies on the observation that when all is said and done, there will indeed have been a single time line. The actions of agents "reveal" to us the time line we will have been in. What in our view is the "free will" of agents, is simply our belief that there could be no one able to predict the time line in advance. A small step for the philosopher, no doubt, but a giant leap for the notational engineer: we are now able to represent both action taking and knowledge in the same framework.

4 Formal treatment

The previous section can be viewed as documentation for the formal construction in this section. The following definitions will explain the distinction between action and mere augmentation of knowledge, the time-asymmetry of action, and the different viewpoints of observers.

Definition 1 Time *is a pair* $\langle T, < \rangle$, *w* T *is a set of time points, and < is a total order on* T *which is unbounded in both directions.*

For our purposes other properties of time, such as density or discreteness, are not important. In the following we assume *Time, T* and <. We also assume a propositional language of discourse £ with primitive propositions $\$$ (the discussion extends trivially to the first order case).

Definition 2 *A* time-line structure 5 *(over T, <,* L, *M and* Φ*) is a pair* $\langle L, M \rangle$ *where*

L is a set of time lines, and

M is a meaning function $M : L \times T \rightarrow 2^{\Phi}$.

We represent the fact that $p \in M(l,t)$ **by** $M, l \models TRUE(t,p)$, **that** $p \in M(l,t)$ **and** $p' \in M(l,t')$ **by** $M, l \models TRUE(t,p) \wedge TRUE(t',p')$, **and so on. If** $M, l' \models \varphi$ **for all** l' **then we say that** φ **is known in** M, **and denote it by** $M \models K\varphi$.

As has been discussed, a time-line structure defines the observer's knowledge of the world in a given situation. Next we define the evolution of this knowledge.

Definition 3 *A time structure S =* (L, *M) is* no more ignorant *than another time structure S' =* $\langle L'$, *M') if* $L \subseteq L'$, *and M is the restriction of M' to* L. *If* $L \subset L'$ *we say that S is* less ignorant *that S'.*

In the following, let *TS* be the set of time structures.

Definition 4 *An* evolving time-line structure (ETS) *(over* T, <, L, Af, Φ *and TS) is a pair* $\langle l_0, F \rangle$ *where* $l_0 \in L$ *is the "real time line", and F is a function FT* $\rightarrow$ *TS, such that a.* *if* (L, M) $\in$ *F(t) then* $l_0 \in L$ *and b.* *if* $t < t'$, *then F(t') is no more ignorant than F(t).*

An ETS describes the evolution of the observer's knowledge about the real world.

It will be useful to define two operations on ETS's.

Definition 5 *Let A and B be two ETS's as above that differ only on the F function. Let the two functions be* F_A *and* F_B. *The* **intersection** *of A and B is a structure* $C = \langle l_0, F_C \rangle$, *where* $F_C : T \rightarrow TS$ *is such that* $F_C(t) = F_A(t) \cap F_B(t)$. *The* **union** *of A and B is a structure C* $= \langle l_0, F_C \rangle$, *where* $F_C : T \rightarrow TS$ *is such that* $F_C(t) = F_A(t) \cup F_B(t)$.

The reader may verify the following:

Proposition 1 *The intersection of two ETS's is a third ETS, which is no more ignorant of either of the first two. The union of two ETS's is a third ETS, of which the first two are no more ignorant.*

Actions of agents may add to an observer's knowledge of the real world. This is defined as follows.

Definition 6 *An* action system *(over T, <, L, M,* Φ *and TS) is a tuple* $\langle l_0, CA, AA \rangle$, *where*

CA is a **capability function** $CF = A_1, \ldots, A_n$, *where each* A_i *defines the actions that are available to an agent i under different conditions. Specifically, it is a collection of pairs* $\langle \varphi_1, \varphi_2 \rangle$, *where* $\varphi_i \in \Phi$. *(If* $\langle \varphi_1, \varphi_2 \rangle \in A_i$ *then we say that "Agent i can do* φ_1 *under condition* φ_2.") *We require that that* $\langle true, true \rangle \in A_i$ *for all i (the null action).*

AA is a list $a_1, \ldots, a_n$ *of actual actions taken by the agents. Specifically,* a_i *is a function* $a_i : T \rightarrow A_i$. *We require that if* $a_i(t) = \langle \varphi_1, \varphi_2 \rangle$, *then* $M, l_0 \models TRUE(t, \varphi_1)$ *and* $M, l_0 \models TRUE(t, \varphi_2)$.

In this definition we assumed that actions depend on what is true, but not on the agent's knowledge. It is possible to capture that too, but since the essence of our proposal does not hinge on that, we ignore the issue here for the sake of understandability.

The last definition relates action to knowledge:

Definition 7 *Let AS be an action structure as above. The* **induced structure** *of AS is the pair* $\langle l_0, F \rangle$, *where* $F : T \rightarrow TS$ *is such if* $F(t) = \langle L', M' \rangle$, *then for all l,* $l \notin L'$ *iff there exist* i, t', φ, φ' *such that a.* $t' < t$, *b.* $M, l \models TRUE(t', \varphi)$, *and c.* $a_i(t') = \langle \neg\varphi, \varphi' \rangle$.

It is not hard to verify the following:

Proposition 2 *An induced structure of an action structure is an ETS.*

From Propositions 1 and 2 we get trivially.

Corollary 3 *The intersection of an ETS with the structure induced by an action structure is an ETS, no more ignorant than either.*

We can thus view the observer's evolution of knowledge about the world as defined by his *a priori* knowledge, some ETS, combined with the knowledge about the actions of agents, knowledge which is gained at the time of action.

5 From single observer to multiple observers

The discussion so far accounts nicely for the choice-making aspect of action. However, our definitions extend naturally to representing different perspectives of observers with varying detail of knowledge.

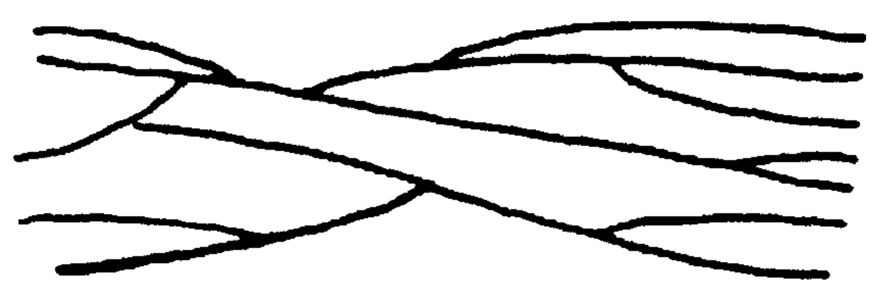

Figure 3: Past and future branching

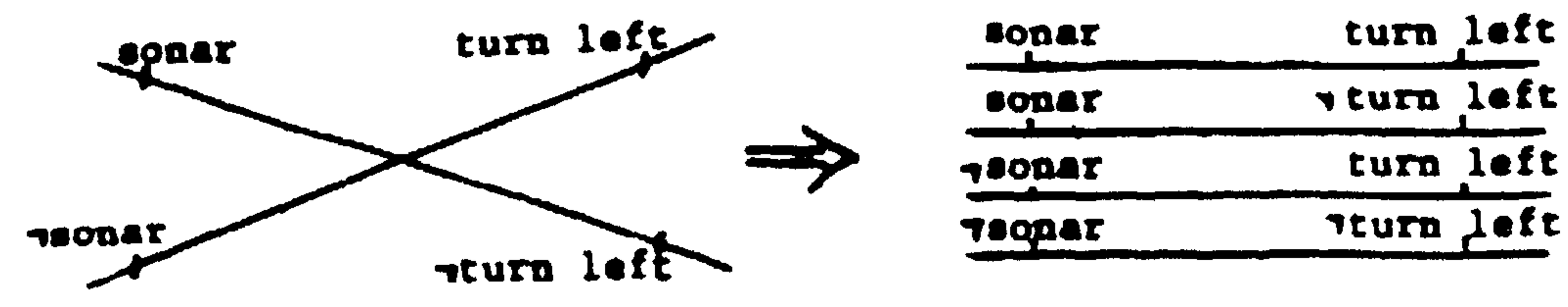

Figure 5: From branching time to parallel time lines

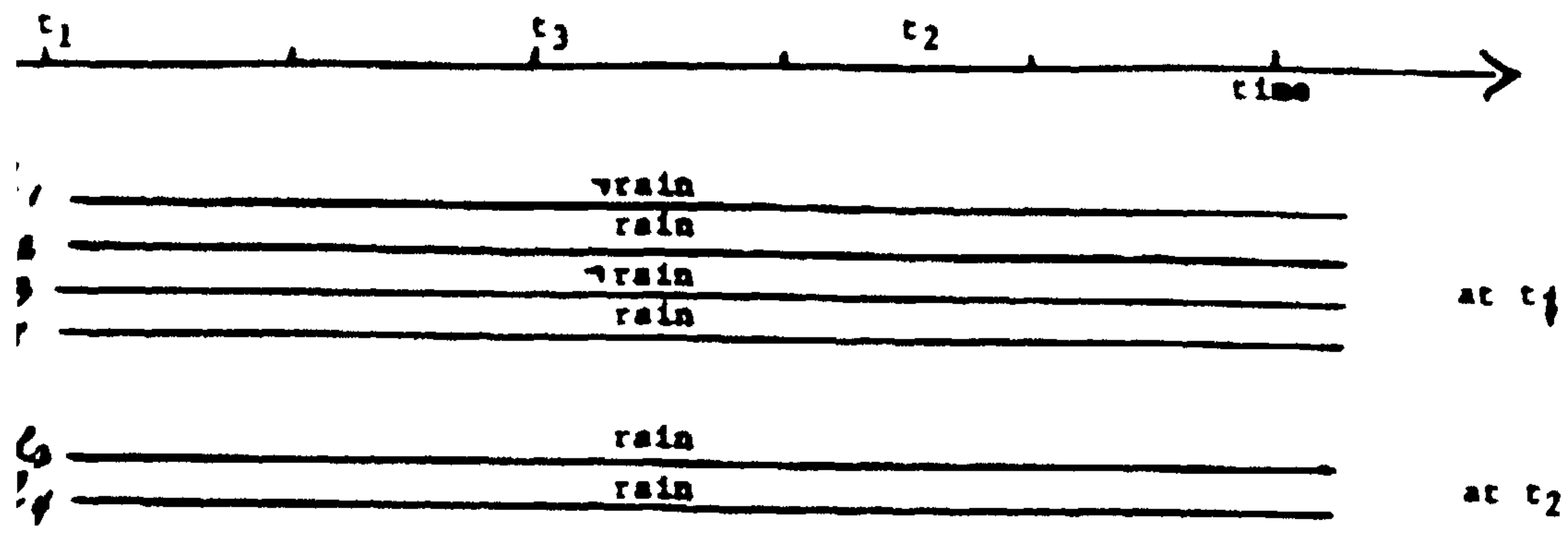

Figure 4: Knowledge changing in time

past, only our knowledge of it. To pursue this issue further, we now turn to the notion of knowledge and its conection to time.

3.2 Knowledge and time

Following the now-standard approach, we will equate knowledge with truth in all possible worlds. For us, however, possible worlds have specific forms: they are each an independent time line. At each point in time, an agent considers some set of time lines possible: the greater his knowledge, the less time lines are possible. This is the standard S5 possible worlds account of knowledge (see, e.g., Halpern and Moses' survey in [Halpern and Moses, 1985]). In our particular case, at each time point an agent considers not only what the possible presents are, but also the possible pasts and futures.

An example is given in Figure 4, where an agent's state of knowledge at two different times, $t_1 < t_2$, is represented. At t_1 he does not know whether it will rain at time t_3 where $t_1 < t_3 < t_2$, since from each world there exist some accessible worlds in which it rains and some in which it does not. At time t_2, presumably having the benefit of hindsight, he knows that it in fact rained at t_3.

In general, the set of possible time lines may change wildly over time, allowing arbitrary learning and forgetting. In pratice we will impose a strong restriction on how the set changes, and assume that no forgetting occurs: What an agent knows at one time, he knows later on too. This means, for example, that if the robot knows at 5 that at that time his batteries are full, he know also at 8 that at 5 his batteries were full; it does not imply that he knows at 8 that at that time (8) his batteries are full.[2] [3]

3.3 Integrating knowledge and choice-making

We now have two mechanisms, a branching structure for choice making, and parallel time lines for knowledge. We note, however, that they are very similar, each describing alternative courses of events. In particular, we note that in the branching structure, every set of branches that meet at a point can be expanded into a set of parallel time lines. In principle, if we have n incoming pasts and m outgoing futures, we simply generate $n \times m$ parallel time lines. Figure 5 decribes a simple case in which $n = m = 2$.

We now have a uniform structure, a set of parallel time lines. The observer, who might be in any of these time lines, at every point considers some set of time lines possible. We assume that this set includes the actual time line of the observer, and, as before, that as time progresses the set of time lines considered by the observer can only shrink, never expand. Some of this shrinking is "natural" acquisition of knowledge, such as the exclusion at time t of all time lines in which there is no house near the robot at time t, when at that time the observer happens to see a house near the robot. Other change in knowledge is more deliberate, and reflects the ability of agents to prune certain time lines at will. For example, in the view of some observer, the robot may "decide" to prune all time lines in which at time *t* the robot did not turn left.

[2] The treatment here depends on agents having some form of memory. Our requirement of perfect memory, however, may be unnecessarily strong.

[3] This construction borrows from a more detailed and general construction in [Lin *et* ai., 1987], where we do not necessarily assume no forgetting, and where we consider belief rather than knowledge. There we provide a logic to accompany this semantic construction, along with a complete axiom system. The basis is a standard interval logic, with atomic propositions such as TRUE(t_1, t_2,p) We then use n modal operators indexed by time, so that KjTRUE(t_1, t_2, p) states that agent i knows (knew, will know) at time t that the proposition p is (was, will be) true at the interval $\langle t_1, t_2 \rangle$. Since these syntactic considerations are not the focus of this article, 1 will not pursue the issue further here.

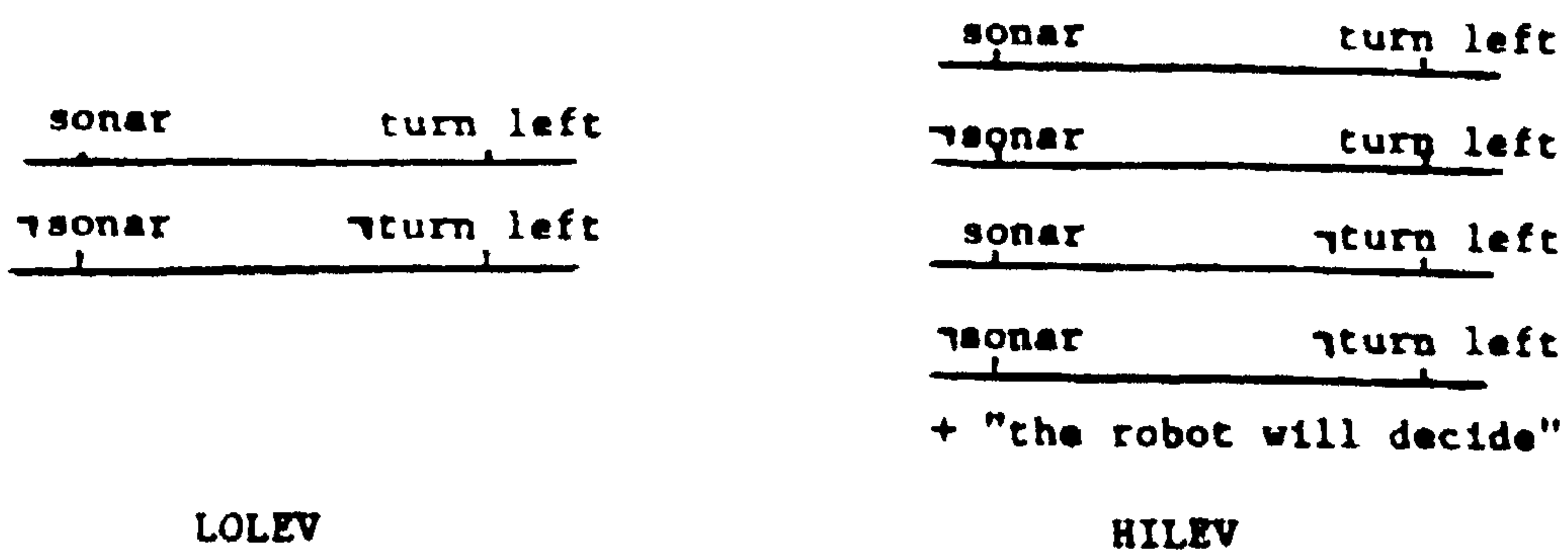

Figure 6: The LoLev and HiLev views of the robot's behavior

We start with intuition. Consider two observers, LoLev and HiLev,[4] such that LoLev is more knowledgeable than HiLev. Specifically, assume that at every point at which LoLev considers some set of time lines L, HiLev considers a larger set of possible time lines $L' \supset L$, and in addition postulates a possible action of an agent (or a set of such) which eliminates from L' those futures in $L' - L$. If nothing is added to the calculus, LoLev clearly has greater predictive power, since he can predict ahead of time the set L. HiLev can only predict the set L', and say that the fate of the futures in $L' - L$ depends on the action of the agent. Figure 6 describes HiLev and LoLev view of the behavior of the same robot.

It is wrong, however, to view HiLev as necessarily inferior to LoLev:

- Sometimes actions are indeed a poor man's physics, as can be seen in the above robot example, or the Piagetian example from Section 2.
- On the other hand, it is often advantageous to deliberately lose some information and reason in terms of action, since the complete information is too rich to contend with. If for example the robot's turning left is the result of 232 factors rather than two, treating it as an action might be the only computationally-viable option. We do not even necessarily lose predictive power that way.
- One reason that we may not lose knowledge is because we ourselves are the agents. If we are the ones "making the decisions," we can always compensate for indeterminacy in our *a priori* knowledge by deciding.
- Even when others are the ones "making the decisions," we are not left clueless as to their future actions. We can often predict their "decisions" quite reliably, and for that purpose have invented a whole slew of related terms, such as "knowledge," "belief," "goals," "plans," "rationality," and other intensional terms. For example, we might predict that a "rational" robot will "choose" to turn left when he "believes" that an object is about to collide with it from the right. Indeed, there is considerable research on the calculus of intensional terms such as belief, knowledge, goals, intentions, desires, and rationality (see, e.g., [Cohen and Levesque, 1987]), although much more remains to be done.

[4]Thc terms 'LoLev' and HiLev' were already used in the Stanford/SRI/Rockwell ICA project for the same general purpose, though not in the specific sense used here.

Thus, important information resides at different levels of abstraction, and it is useful to be able to tap on all of it. For example, it is useful to integrate our knowledge of rational behavior with our understanding of the robot's sonar system. This view is compatible with the *situated automata* approach, in which internal states of the robot are correlated with external conditions, and intensional terms such as knowledge describe this correlation.[Rosenschein and Kaelbling, 1986]

As was said, it is very easy to extend the treatment to accommodate multiple perspectives in our framework. Again, the formal details will appear in the full version of the paper (and some are given in the Appendix). The highlight of is the definition of an ETS-system as a collection of ETS's, and noting that those form a natural lattice with the ETS intersection and union operations.

6 Formal treatment (cont.)

It is very easy to extend the formal treatment to accommodate multiple perspectives in our framework. In fact, we already have a special case of multiple perspectives: the ETS of the observer, and the induced structure of his postulated action structure. As we saw, the two intersected to yield a new ETS. We now simply generalize the picture.

Definition 8 *An* ETS system *(over* $T, <, L, M, \Phi$ *and TS) is a pair* $\langle l_0, F \rangle$, *where* $l_0 \in L$, *and* $F = F_1, \ldots, F_n$ *is a set of functions suck that* $\langle l_0, F_i \rangle$ *are all ETS's (over* $T, <, L, M, \Phi$ *and* TS*).*

The following is immediate from the earlier propositions:

Corollary 4 *An ETS system defines a lattice, with the nodes being the ETS's, and the join and meet operations being intersection and union of ETS's.*

7 Summary and Discussion

I started by pointing out that taking actions as primitive and ignoring time is problematic, and that taking time as primitive and ignoring action is problematic in a complementary way. I then developed an approach to retaining the best of both worlds. I first identified two notions that seem closely related to action: choice making, and relativity to observer.

(Continued as an appendix to the article "Belief as Defeasible Knowledge" in these proceedings.)