COSC410 Lecture 9
BDI agents and their Logic

Willem Labuschagne
University of Otago

2016

Introduction

What is a BDI agent? There may be differences in detail, but all versions and implementations of BDI agents should have three things in common: ways to represent Beliefs, Desires, and Intentions. We further assume that a BDI agent is rational, i.e. able to reason and to communicate its reasoning. We do not want a BDI agent to be a black box achieving results in mysterious ways. Our paradigmatic example of a BDI agent is an agent (human or robot) capable of solving murder mysteries and giving coherent evidence in court to explain the solution.

In this lecture we explore the general features of such an agent and the sort of logic language that might be used to model its reasoning. It's important to say clearly at the very start that no such agent has yet been implemented in full – only bits and pieces that seem to show in principle that if we work hard we may be able to design and implement a real BDI agent, capable of reasoning, at some future instant (hopefully before the heat death of the universe).

To start with, we should say a bit more about 'reasoning'. All reasoning is thinking but not all thinking is reasoning. (To sound profound, replace "thinking" by "cognition".)

Consider a robot that uses machine-learning algorithms to balance itself while walking on ice. Balancing requires fast responses to correct the growing tilt from the vertical. In humans such processes are always unconscious, because conscious processes are much slower — too slow for smooth motion. Partly, the speed results because unconscious processes can occur in parallel, whereas conscious processes are sequential. Speed and parallelism are the key to co-ordinating complex motion. There are many things we humans do that we consider easy, such as grasping a teacup or doorknob, but which are actually very subtle problems that seem easy to us only because we don't have to consciously decide "This finger should go here, that one over there". Other processes may be unconscious because they need to be maintained constantly, not just when we're thinking about them — breathing, digesting food, registering the size and shape of obstacles to our locomotion (think of stepping up onto a kerb that you perceived unconsciously, through your peripheral vision, without saying to yourself "Behold, I see before me a kerb, of height six inches.").

Our BDI agent must, if it is to be interesting, be equipped with various unconscious abilities related to perception and performing actions, e.g. grasping objects with its effectors, balancing, walking, recognising the sound of an approaching bus (hopefully before feeling its impact). These unconscious processes involve the processing of information and so we may consider them to be

forms of cognition. But the information processing involved in these unconscious processes is not what we have in mind when we use the word 'reasoning'.

We take reasoning to be a conscious process of inferring conclusions from premisses. There are two separate components to this assertion.

Firstly, reasoning involves language because the agent must be able to say what the premiss is and what the conclusion might be. Lots of cognition occurs without language, but equally, in humans a lot of thinking does occur in language and we may often have the feeling of talking silently to ourselves. So, some thinking uses language, and this is what we have in mind when we talk of reasoning — unless the starting point and endpoint of the cognition are tied down in language we don't consider that cognition to involve reasoning.

Furthermore, reasoning is a 'conscious' process in the sense that the inferences are sequential and can be tracked by the agent herself. This is an important safety mechanism. Someone once said: "Writing is nature's way of letting you know how sloppy your thinking is." Well, it's not just the writing, it's being able to read through what you've written. The fact that reasoning involves language allows an agent, if gifted with the ability to monitor its own use of language, to step through the inference sequentially and discover mistakes, which helps to ensure that the agent's "thinking goes right".

We shall not explore the ability of an agent to monitor its own reasoning, i.e. the concept of consciousness. It's a big topic, and neuroscience is making great strides in solving its mysteries – see for example Antonio Damasio's book "*Self Comes to Mind*". Instead we shall focus on the language used by BDI agents for their reasoning.

That a reasoning agent must be equipped with a capacity for language has wider implications. Agents are usually not alone in their environments. If premisses and conclusions are expressed in language then the agent's reasoning can be shared with and checked by other agents. Such sharing allows second-hand learning to occur and ultimately provides the precondition for multi-agent societies to develop a shared culture. This is how it works for humans, at any rate, and humans are a particular kind of BDI agent, the kind we know best but didn't design ourselves, which is why we're so interested in designing other types of BDI agents — to test our understanding of how humans work.

The class of BDI agents therefore includes not only humans but also a vast range of robots we might design. Keeping in mind the robot detective example, BDI agents will need to be embodied, i.e. not limited to input in the form of language like a pure software agent but equipped with various sensors (perhaps a video camera instead of the human eye, or sonar, or radar) and effectors (perhaps metal pincers instead of human hands, or magnetic pads, or suction devices), as well as being able to use language.

BDI agents may use language in three main ways:

- to build and store an internal model of their environment in the form of sentences of some language (the set of **beliefs**)
- to perform acts of reasoning
- to communicate with other agents (e.g. to request information or to inform other agents).

Beliefs are important but on their own are of little use. Every BDI agent will have a number of **desires**, ranging from short-term goals (to get something to eat, to park the car) to long-term

goals (to become famous, to be a good person). So a desire is some sort of goal towards which the agent is vaguely disposed to work.

Goals in the form of desires are important but may be mutually incompatible. One of the great truths Aunt Maud is ever willing to share with you is that you can't have everything you want. So a BDI agent must have the capacity to decide what goals to pursue in the immediate future and to commit to achieving those particular goals. These are then the agent's **intentions**. Intentions are desires but some desires are not intentions.

Intentions are important but we don't want BDI agents to be ineffectual dreamers, full of good intentions while realising none of them. A BDI agent must be able to use its beliefs about the world in order to formulate **plans** to achieve its intentions, and then carry out the actions that will bring about the desired state of the world. Plans are recipes for achieving intentions.

Summarising, a BDI agent has beliefs and desires expressed as sentences of a language, has a way to select some desires and commit to achieving them, thus forming intentions (also expressed in language), and is capable of planning to use actions to change the state of the system/environment until its intentions are realised. The plans may involve acting individually or recruiting the help of other agents via communication (since many goals cannot be achieved alone, at least not without impractical investments of time and effort – think of building a house or writing a really big piece of software).

The BDI agent control loop

In the best-known textbook on BDI agents, Michael Wooldridge's *Reasoning about Rational Agents*, the behaviour of a BDI agent is summarised by the following loop:

1. observe the world
2. update internal world model
3. deliberate about what intention to achieve next
4. use means-end reasoning to get a plan for the intention
5. execute the plan
6. begin the loop again.

How does this match up with what we've done in logic? Let's take a closer look.

Step 1:
In step 1 the agent forms beliefs about the world by observation using sensors or senses such as vision. In other words, step 1 is about acquiring information by first-hand experience. Step 1 is part of AI rather than logic, and involves a difficult problem that I'll try to explain.

Imagine a human being using an eye, or a robot equipped with a video camera, so that visual information can be gathered from the environment. Initially, an analog image is formed on the retina of the eye or on the film in a camera.

Problem: *How does this analog image give rise to beliefs which are sentences?*

Somehow, information given in analog form must be converted into discrete sentences.

We more or less have an idea of how this works in the case of the human eye and brain, and presumably can use this understanding as inspiration to find a way to do it in the case of a BDI agent. Stevan Harnad is a prominent researcher in cognitive science who in many publications has espoused a tripartite system of representation in the human brain. Here is his theory of how an analog image on the retina eventually gives rise to sentences expressing beliefs.

Consider a distal stimulus object in the environment, for example suppose the agent is looking at Aunt Maud's new car. The stimulus object casts a proximal projection on the transducer surface of the agent's eye, i.e. light reflected from the car produces an image on the retina of the eye. This sensory input is registered by changing the state of the body at the relevant site (the retina), and these changes are propagated to the brain by nerve fibres.

It is important to realise that at every stage in this propagation information is lost. The eye has around 100 million light-sensing cells, and there are only about 1 million fibres leading to the brain, so each incoming image is simplified (reduced in complexity, i.e. information is thrown away) by a factor of 100. At later stages, filtering occurs (as we shall discuss), which again throws some information away. Basically, our heads (or the corresponding parts of an artificial agent) are much smaller than the universe, and so we can't hold the full complexity of all the information available out there inside our little skulls.

In the early stages of perception, the pattern of activation of the nerve cells (i.e. the mental image) is an *iconic* representation of the stimulus object, by which we mean that the internal representation is topographically organised, with parts of the retinal image that are close together corresponding to areas of excitation that are close together in the brain. (Think of it this way — the left headlight of Aunt Maud's car is closer to the left front wheel than to the right front wheel, and in a topographically organised representation, the neural pattern representing the left headlight will be closer to the neural pattern representing the left front wheel than to the neural pattern representing the right front wheel.) Call this (topographically organised) iconic image the first level of representation in the brain.

The iconic representation in the brain is an analog of the image cast by the car on the retina. This holds, by the way, not just for visual inputs but for the other senses as well. For example, in the case of sound the hair cells in the cochlea are ordered in such a way that a tonotopic map is produced, i.e. the representation of sound in the brain is organised spatially in a manner ordered by the frequencies of the tones.

These iconic representations suffice to allow discrimination, i.e. to permit the agent to make judgements about whether two stimuli are the same or different. But it is not enough to say that this object is different from that object — we need to recognise the object as a car. To allow identification of the distal stimulus object we need *categorisation*. In other words, suppose we have a photo of Aunt Maud's car and a photo of our own car. We can see they are different, since Aunt Maud has splashed out on a Jaguar XK while we chug along in our 1993 Toyota Corona. But unless we have some sort of generic concept of what a car is, we won't be able to say "Ah, that thing is a car!" The generic concept is what we mean by the *category*. How does one learn the generic concept?

To build the generic concept (the category), or to recognise the category into which a new iconic representation fits, the iconic representation is filtered, i.e. undergoes changes that exaggerate some features by discarding others, in much the same way as an accurate photo of a person might be caricatured by a cartoonist's line drawing. Or, as Harnad puts it, the iconic representation is "selectively reduced to those invariant features of the sensory projection that will reliably distinguish a member of the category from any nonmembers with which it could be confused"

(Harnad (1990): The symbol-grounding problem, *Physica D* 42:335–346). This reduction is like a hardening of boundaries that changes what was originally continuous variation into discrete parts that can more readily be distinguished. The omission of detail makes the result more generic, so that it can stand for a whole class of items rather than one specific item, in other words the output of the process is a category. Call this categorical representation the second level of representation.

So now we have two sorts of mental images: iconic representations that are "analog copies of the sensory projection, preserving 'shape' faithfully" as well as categorical representations which are "icons that have been selectively filtered to preserve only some of the features of the shape of the sensory projection" (Harnad 1990).

Symbolic representations are the third level of representation. Some sort of symbolic system assigns labels to categories, i.e. associates names with the categorical representations, such as "car", or even "Toyota". These symbols (names) are grounded in the sense that they are ultimately linked with distal objects in the environment by a chain composed of iconic representations and categorical representations. The symbolic system in modern humans has syntax, allowing the grounded names of categories to be strung together into propositions such as "The Toyota is blue". By the use of symbols with syntax, i.e. language, an agent acquires a new way to manufacture concepts (i.e. categories) by what Harnad calls 'theft instead of honest toil'. The idea is that categories can now be acquired not just by experience of the environment but through language. Suppose the agent already has the categories "horse" and "stripe". Then the agent can create a new category, which might be labelled "zebra" and which includes all horselike objects having stripes. This is theft instead of honest toil because the agent can construct this new category even if the agent has never seen a zebra, just by combining symbols that have been grounded in categories of which the agent does have direct experience.

The three levels of representation constitute Harnad's tripartite model of representation in the brain, and we would need something of this kind in order to design an agent that can convert the sensory input from a video camera (or other sensor) into sentences of a language. It would be fair to say that no single approach to designing such systems has triumphed as yet, but machine learning algorithms are achieving great success in changing iconic to categorical representations, for example in Google's driverless car. Finding a suitable way to implement Harnad's tripartite model (or some other model that addresses the same issues) is an important problem of AI rather than logic. We shall therefore say no more about the conversion of sensory input to sentences (beliefs), and merely assume that some way has been found to achieve this conversion, thus implementing step 1 of the agent control loop.

Step 2:
Much that you have learnt about logic in COSC410 is really aimed at step 2 of the control loop. A belief set is formed by taking whatever sentences represent the agent's initial information and closing under classical entailment, and the set of beliefs is changed in the light of new information. Our familiarity with classical entailment, nonmonotonic logic, and AGM belief change theory gives us the tools to cope with step 2.

Step 3:
Logic will also influence the mechanism by which the deliberation process is implemented (step 3 in the control loop). Deliberation is about choosing intentions and committing to them. This is a complex process because there are various constraints that must be satisfied. For example, suppose we interpret "committed to achieving the intention φ" as saying that the agent is committed to undertaking a plan of action that will change the state of the system so

that φ becomes true. Then this commitment needs to be compatible with the agent's beliefs — it would be irrational for the agent to commit to achieving φ if the agent in fact believed it was impossible for φ to be true. It would not be rational for you to form the intention of becoming a concert pianist if you believed that you would be unable to achieve the finger dexterity required to play Chopin etudes. It would not be rational of a student to commit to becoming a surgeon if he knows that he faints at the sight of blood and would be unable to slice boldly through the skin of his patients. So logic plays a role in the deliberation process, both in the form of taking the future into account and in terms of reasoning about the agent's own beliefs (possibly formalised as knowledge using epistemic logic, for example).

Step 4:
Step 4 is about planning, and although logic would obviously be relevant, we shall not go into the details. The kind of reasoning most often used in planning is called *means-ends analysis* and was invented by Herbert Simon and Allen Newell, whom you may remember as the authors of the Logic Theorist reasoning program exhibited at the 1956 Dartmouth Conference. We won't examine means-end analysis in detail. (Basically, it treats the construction of a plan as a problem-solving exercise in which one searches for the solution to a problem through a space of system states, choosing an action at every step that will reduce the distance between the current state and the goal state.)

Step 5 is about executing the plan, which we may consider a part of AI.

The language of BDI logic

To begin with, the agent is situated in an environment (the system of interest) and so the language must be capable of expressing the elementary facts of relevance to that environment. Let L_A be a propositional language with a suitable set A of atomic sentences. For instance, if the agent is the driver of a car then A would include sentences such as "The traffic light is red". Or if the agent is a robot detective, then the atomic sentences would express facts such as "Colonel Mustard has an alibi". The language L_A forms the basis of our BDI logic, but needs to be enriched with a variety of modal operators of both the epistemic and the temporal varieties.

Talking about beliefs, desires, and intentions

There are three 'modalities' or agent-attitudes we want to associate with BDI agents, namely believing, desiring, and intending.

First we need a way to say "Agent i believes that φ". Typically a modal box operator is used for this. If we are satisfied to equate belief and knowledge, then we could use the epistemic language having an operator $[i]$ for each agent i, expressing "agent i knows that". If we want to distinguish between knowledge and belief then the possibilities vary, but the conservative old-fashioned approach would be to use a modal box operator often written as B_i, as in our brief discussion of doxastic logic in lecture 7 (page 7). One still has an accessibility relation for agent i and defines satisfaction in terms of that accessibility relation in the usual way, but one doesn't require the accessibility relation to be an equivalence relation. As mentioned in lecture 7, this treatment of belief doesn't really match up with nonmonotonic logic and AGM belief change theory, but that's something it may be possible to change. Let's just agree for now that the

language of BDI logic will have a modal box operator for each agent to represent that the agent knows or believes something, and for compatibility with Wooldridge's textbook we'll write the box operator as Bel_i.

More specifically, we add to the language three modal box operators for each agent, writing

- $Bel_i\varphi$ to express that agent i believes that φ
- $Des_i\varphi$ to express that agent i desires φ to be the case
- $Int_i\varphi$ to express that agent i intends to bring about φ.

Note that since we are using modal box operators, we may nest the operators and write, say,

- $Bel_iBel_m\varphi$ to express "Agent i believes that agent m believes that φ"
- $Int_iBel_m\varphi$ to express "Agent i intends that agent m should believe φ"
- $Des_iInt_m\varphi$ to express"Agent i desires that agent m should intend to make φ true".

We are already able to express a number of important properties that BDI agents may have or should not have, depending on what we consider to be rational:

- The sentence $Bel_i\varphi \to \neg Bel_i\neg\varphi$ expresses that if agent i believes φ true then agent i must not believe that φ is false. We would want our agents to satisfy this property, otherwise they would be insane, and who wants a six-ton rocket-firing robot that is totally insane? Well, maybe the American Department of Defence.
- The sentence $Int_i\varphi \to Des_i\varphi$ expresses that if agent i has the intention of making φ true then agent i must desire that φ should be true, which makes sense given that intentions are desires the agent selects in order to commit to achieving them. But what if the agent is acting under coercion like a bank manager opening the safe because the robbers are waving guns around? The bank manager intends to open the safe but really hopes something will happen to prevent it, such as the police arriving in time. So perhaps the way intentions and desires are related in BDI agent design needs some re-thinking!
- $Bel_i\varphi \to Des_i\varphi$ expresses a property known as *realism*, saying that if agent i believes φ to be the case, then agent i desires φ to be the case, i.e. agent i accepts the inevitable. Those who think it is a property that rational agents ought to have would argue that if I believe the sun will rise tomorrow then it makes no sense for me to desire that the sun will not rise. Those who believe the property is too strong to be rational would argue that in the first place I may have no desire one way or the other, and in the second place I may be dreading tomorrow and therefore could indeed desire that the sun not rise.
- The sentence $Int_i\varphi \to Bel_i\varphi$ expresses that if an agent intends something then the agent believes it. At first glance this seems reasonable, but if you think about it, it becomes less convincing. Suppose I have the intention of becoming an All Black. Does that mean I believe I am an All Black? Surely not. Even if we think of φ becoming true in future, does having the intention mean I definitely believe I will become an All Black? Surely not, because many things can go wrong (e.g. an injury). What I would need to believe is that it is *possible* to realise my ambition, not that it is inevitable. Intentions typically involve the agent acting according to some plan, and the plan needs to take into account the various ways in which the state of the system can be altered over time, so we'll come back to this after the next section.

Talking about time

In order to do their planning, BDI agents need to be able to reason about the way the system state may change over time, and so we introduce a number of temporal connectives into our language. However, it is usual to adopt a rather complex branching model of time instead of the simple linear discrete time model we used in lecture 8. Take a deep breath and hold onto your head so it doesn't fall off.

Suppose we are in a particular state of the system, which we take to be *now*.

Towards the past, our model of time remains simple — we should imagine a line of discrete time instants beginning at some time t_0, the starting time of the universe as far as our agent is concerned, and extending all the way to whenever 'now' may be. In other words, the system has only one past history up to the present moment, which we can visualise as a line of time points (instants) extending from t_0, the starting moment, to t_n, the present instant of 'now'.

The future is more complicated because it has not yet been determined. An agent who is planning to achieve an intention will need to consider the effects of various actions, each of which may change the state of the system in a different way. Standing at now and gazing into the future, we therefore visualise time as a tree branching into the future from the root now, with each branch being a sequence of time instants that represents a possible future. So branches in time represent the choices available to agents with respect to the actions they may perform. As time passes and the instant 'now' moves forward, the branches not taken are pruned away so that the past remains linear.

Technically, we adopt what is called a forward-branching (or left-linear) time structure, and we choose the time structure to be discrete instead of dense or continuous (so that we can talk of one instant being immediately before or immediately after another). Moreover, the time structure is bounded towards the past but unbounded towards the future. None of these decisions is forced and we could have chosen the time structure to be different, but have stuck close to the model of time adopted in Wooldridge's *Reasoning About Rational Agents*.

Now let's make things more precise. The set of time instants is $T = \{t_0, t_1, t_2, ...\}$ and as the temporal accessibility relation we take a total left-linear binary relation on T.

What do we mean by this? Well, a binary relation on T is a relation R consisting of ordered pairs (t, s) of time instants, the idea being that if the pair (t, s) is in R then t comes before s.

- We require R to be *total* in the sense that for any time instant t there is a time instant s that is 'later' according to R, i.e. for every t in T there is at least one s in T such that (t, s) is in R. This ensures that time continues infinitely towards the future.
- And we require R to be *left-linear* in the sense that if t and s are two different instants occurring before now, then either (t, s) must be in R or (s, t) must be in R. More formally, since every instant is a possible 'now', we define R to be left-linear iff for all t, s, n in T, if (t, n) is in R and (s, n) is in R, then either (t, s) must be in R or (s, t) must be in R. (So if t and s occur before now $(= n)$, then one of them must occur before the other, i.e. they cannot lie on different time lines but have to lie on the same line, forcing there to be just one past.)

Note that the usual ordering which makes t_i come before t_j if $i < j$ is one possible choice of the relation R, in which there happens to be only a single branch into the future (because the whole timeline is now just a single straight line).

More generally, however, we may vary from this simple choice of R in order to allow as many branches into the future as it takes to accommodate the possible actions available to the agent at each point. Since the relation R is then not linear towards the future, you will be able to find two instants t and s in the future which are not comparable using R, i.e. neither (t, s) nor (s, t) is in R, because t and s lie on different branches. Suppose Aunt Maud sent me a hideous tie for Christmas. One branch, on which t lies, might be the branch that resulted because I neglected to write a thank you letter to Aunt Maud, so she left me out of her will, and the other branch, on which s lies, might be the branch that resulted because I politely wrote to thank her for her thoughtful gift, so that she left me her shares in Apple.

This makes the semantics rather more complicated than in lecture 8, and we will not go into it in detail. There are two main changes.

The first is that a state is no longer just a naked time instant. A state (or 'world', as Wooldridge prefers to call it) is a pair $w = (T_w, R_w)$ where w is a miniature time structure, with T_w being a subset or part of T and R_w being a part of R that includes only instants from T_w and that has the same properties as R, namely is total (thus unbounded towards the future) and left-linear.

It is easy to see that a world w has a time structure that begins at some time instant, which may or may not be the same as the initial time instant t_0 of the whole T. The beginning time in world w may be considered the instant 'now' in that world. There is a unique sequence of time instants from the global starting time t_0 to 'now', which is easily recovered from the global time structure R, so we need not include it in the local time structure of w. We can limit the time structure of w to what may happen from the instant 'now' in w. And the idea is that the time structure of w represents all the possible ways in which, from 'now', the system could evolve as a result of actions performed by the agent.

The second change is that we can no longer define satisfaction of a temporal sentence with respect to a single time instant. Instead we consider satisfaction relative to a *path*.

Intuitively a path is one possible timeline from 'now' onwards, where 'now' is determined by the world. Suppose w is a world. Then a path through w is a sequence of time instants, i.e. a function $p : \mathbb{N} \to T_w$, such that $p(0)$ is the instant 'now' of world w and each adjacent pair $(p(i), p(i + 1))$ corresponds to some pair (t, s) in R_w. Informally, a path is a sequence of time instants that start at 'now' and form a single line — the path doesn't have any branches because it represents a particular choice being exercised at every possible branching point.

All right, bearing these small complications in mind, what sort of temporal connectives do we wish to include in our language?

As our temporal connectives we take a mixture of the familiar and the new, but we shall adopt a bias towards the future (because of the emphasis on planning), and this bias changes some of the familiar connectives by ignoring their past dimensions:

- $\bigcirc$ means "at the next instant", and the idea is that $\bigcirc\varphi$ will be satisfied by a path through a world w if, on that path, φ is satisfied at the instant immediately after 'now'.
- $\Diamond$ means "either now or at some future instant", and we should really use $\langle F \rangle$ with the accessibility relation $\leq$ on T, but then you would be unprepared for Wooldridge's use of the unadorned diamond for this purpose. In any case, the idea is that $\Diamond\varphi$ will be satisfied by a path through a world w if, on that path, φ is satisfied at some instant, including possibly the instant 'now'.
- $\Box$ means "now and at all future instants", so again we would have been more consistent

with our notation if we had used $[F]$ with the accessibility relation $\leq$, but ... you know why ... Wooldridge! The idea is again that $\Box\varphi$ will be satisfied by a path through a world w if, on that path, φ is satisfied at every instant, including the instant 'now'.

- $\mathcal{U}$, the Until operator. The idea is that $\varphi \,\mathcal{U}\, \psi$ will be satisfied by a path through a world w if, on that path, there is some future instant at which ψ is true and from now until just before then φ is true.
- **A** means "it is inevitable that" or, if you like, "on all paths from now to the future". Here we really have a different kind of animal, which is specifically designed for our branching time model. All our previous temporal connectives are satisfied by a path, which means we can restrict attention to the instants on that path, so we are dealing in effect with linear time and the temporal connectives work pretty much as we are used to them working. But now $\mathbf{A}\varphi$ is satisfied at a world w iff φ is satisfied by all paths through w.
- **E** means "it is possible that" or, if you like, "on at least one path from now to the future". This is again an operator built for the branching time structure and $\mathbf{E}\varphi$ is satisfied at a world w iff φ is satisfied by some path through w.

Talking about actions

The language of BDI logic described thus far is incomplete — we should include a way to represent actions. Since this is something new and fairly complicated, we won't. Just keep in mind that in order to do planning, an agent must be able to reason about actions and their effect.

Talking about groups of agents

The language of BDI logic should include ways to represent plans involving groups of agents working together. For instance, instead of merely being able to express that agent i believes φ, we should be able to express that all the agents in some group share the common knowledge or common belief that φ. We won't go into the details. Just keep in mind that this is an important feature of BDI logic because solitary BDI agents are the exception, not the rule. (Take humans, for example.)

Properties of BDI agents

We are now able to express a number of important properties that BDI agents or their environments may have or should not have. Two important kinds of property are *safety* and *liveness* properties.

Safety properties have the general form of saying that some bad thing won't ever happen, or equivalently that something you like will always be true. So a typical shape for a safety property is $\Box\varphi$, where φ is the 'invariant' saying that the closet is monster-free, or that robot drivers stop at red lights, or that witnesses don't lie in court.

Liveness properties have the general form of saying that some particular good thing will eventually happen, and so a typical shape for a liveness property is $\Diamond\varphi$, where φ is the good thing.

Safety and liveness properties involve temporal operators only. Most of the important properties involve agent attitudes, e.g. belief.

Should every BDI agent have the property given below?

- $Bel_i\varphi \rightarrow \neg Bel_i\neg\varphi$?
  We've seen this property already. It says that if agent i believes that φ is true, then agent i does not believe that φ is false. Clearly we do want every BDI agent to have this property.
- $Des_i\varphi \rightarrow \neg Des_i\neg\varphi$?
  This property says that if agent i desires that φ should be true, then agent i does not desire that φ should be false. Perhaps we do want every BDI agent to have this property, but that depends on the mental complexity of our agents. If we consider humans as examples of BDI agents, then humans certainly have occasional mixed feelings about something, mingled anticipation and dread, for example. A rugby player before an important match wants to play but fears losing. So this is not quite as simple a property as it may seem.
- $Int_i\varphi \rightarrow \neg Int_i\neg\varphi$?
  This property says that if agent i intends that φ should become true, then agent i does not intend that φ should become false. Clearly we do want every BDI agent to have this property.

The most interesting properties are those that involve mixed modalities.

Should every BDI agent have the property given below?

- $Bel_i\varphi \rightarrow Int_i\varphi$?
  This says that if agent i believes φ is true then agent i intends to make φ true. This seems pretty stupid. The fact that the agent believes φ to be true ought to free the agent to turn attention to other matters, not expend time and effort trying to achieve something that is already the case. However, if we introduce a temporal dimension then we get something more interesting.
- $Bel_i\varphi \rightarrow Int_i\ \mathbf{A}\varphi$?
  This says that if agent i believes φ to be true then agent i intends to make sure that φ remains true, come what may. One would certainly not want an agent to have this property for all possible φ, but one can imagine that some properties may be important enough to justify such a commitment of an agent's efforts. For example, φ might express that the agent had found a really satisfying job (and therefore should fight to keep it), or bought a very comfortable and well situated home (and therefore should resist all temptation to sell it), or finally managed to aim the pesky rocket ship at Pluto. Sometimes, good things must be maintained.
- $Int_i\varphi \rightarrow Bel_i\ \mathbf{E}\varphi$?
  Earlier we saw that the property $Int_i\varphi \rightarrow B_i\varphi$ was too crude to be realistic, amongst others because it ignored the temporal aspect of 'planning to achieve an intention'. But now we have a property that includes this temporal aspect: if agent i intends to make φ true, then agent i really should believe it is possible to make φ true, i.e. that there is some path into the future, determined by some selection of actions the agent can perform, on which φ becomes true at some point. This fits the thinking of the agent who sanely intends to become an All Black, for example.
- $Des_i\ \mathbf{A}\varphi \rightarrow Int_i\ \mathbf{A}\varphi$?
  This says that if agent i desires φ to be true always then agent i intends to make sure that

φ remains true, come what may. This is not irrational – suppose that you desire to spend the rest of your life happily married, then it makes sense to commit yourself to achieving this state and keeping it that way. Indeed, in human society this commitment is precisely what marriage ceremonies are suppose to signify. However, there is the danger of the agent having mutually incompatible desires, in which case the agent should select only some as intentions. So this is not a property on which one would wish to insist for rational agents.

- $Int_i\ \mathbf{E}\varphi \rightarrow Des_i\ \mathbf{E}\varphi$?
  Suppose agent i intends to make $\mathbf{E}\varphi$ true. What does this mean? Well, think of it as agent i keeping the option φ open, i.e. ensuring that there is a possible future timeline on which φ is true. This often makes sense — a student may want to choose papers that keep the option open of becoming a database administrator. Now the property says you should do this only if you desire to keep that option open, i.e. only if you think you might possibly want to be a database administrator.

Exercises

Quiz: The quiz for lecture 10 will be based on the first exercise below.

1. What does the following say? Do you think every BDI agent should have this property?
   - $Des_i\varphi \rightarrow Int_i\varphi$?
   - $Bel_i\mathbf{A}\varphi \rightarrow Bel_i\mathbf{E}\varphi$?
   - $Int_i\mathbf{E}\ \Box\varphi \rightarrow Int_i\mathbf{A}\varphi$?
   - $Int_i\mathbf{A}\ \Box\varphi \rightarrow Des_i\mathbf{A}\varphi$?
   - $Int_i\varphi \rightarrow \neg(Des_i\neg\varphi)$?
   - $Bel_i\varphi \rightarrow \neg(Des_i\neg\varphi)$?
   - $Des_i\varphi \rightarrow \neg(Bel_i\neg\varphi)$?
   - $Int_i\varphi \rightarrow \neg(Bel_i\neg\varphi)$?
2. How would you express each of the following properties in BDI logic? Does it seem like a reasonable property of rational agents?
   - If an agent intends to bring about the truth of φ then that agent should not desire that φ is false in all possible futures.
   - If an agent intends to bring about the truth of φ then that agent should not believe that φ is false in all possible futures.
   - If an agent believes that φ is inevitable, then that agent should not desire that φ be false.
   - If an agent believes that φ is inevitable, then that agent should not adopt the intention to make φ false.