Agent Theory

Agent Theory

1. Why agent theory?
2. Agents as intentional systems
3. Foundations of formal logic
4. Introduction to modal logic
5. Logic of knowledge
6. Examples of agent theory and models (BDI-architecture)

References - Curriculum

- Wooldridge: ”Introduction to MAS”,
  - Chapter 17

Why Theory?

- Formal theory has (arguably) had restricted impact on the general practice of software development. Why should they be relevant in agent-based systems?

➢ Answer: we need to be able to give a semantics (literally a meaning) to the architecture, languages and tools that we use

➢ Without a semantics, it is never clear exactly what is happening and why it works.

- We need a theory to reach any kind of profound understanding of the tools.

Use of Formalisms

- Formalization of agents has been used for 2 distinct purposes:
  1. As **internal specification language** to be used by the **agent** in its reasoning or action.
  2. As **external metalanguage** to be used by the **designer** to specify, design and verify certain behavioral properties of agents situated in a dynamic environment.

Ref: Singh et. al., 1999

Agent Theory

- Agent theory gives:
  - An overview of the ways in which an agent is conceptualised.
  - Semantics to the architecture, language and tools.
- An agent model is needed to develop a theory of agents.

Agents as Intentional Systems

Where do theorists start from?

- The notion of an **agent as an intentional system**.
- So, agent theorists start with the view of agents as intentional systems: where agent's behavior is explained in terms of attitudes.

Attitudes

- Attitude represents a summary evaluation of a psychological object (e.g., oneself, other people, issue, plan, behavior)
- Attitudes are formed throughout the interaction with the surrounding environment.
- Attitudes help to manage individual's cognitive resources to deal with uncertainties of complex dynamic domains

Adopted from: Jaesuk Ahn

Intentional systems

What is an attitude?

*Julia took her umbrella because she **believed** it was going to rain*

*John worked hard because he **wanted** to possess a Master degree*

Use of a folk psychology, by which human behaviour is predicted and explained through the attribution of attitudes.

The attitudes employed in such folk psychological descriptions are called the intentional notions.

An approach is to describe agent's behaviour in terms of intentional systems, "whose behaviour can be predicted by the method of attributing belief, desire and rational acumen"

Intentional systems

Is it useful to consider agents (similar to humans) as intentional system?

(McCarthy)

- it is legitimate when it express the same information about the machine that it expresses about a person
- it is useful when it helps us understand the structure of the artifact
- it perhaps never logically required even for humans
- ascription of mental qualities is most straightforward for machines of known structure but most useful for entities whose structure is incompletely known

Intentional systems

Do we need describe light switch as an intentional system?

"Light switch is perfectly coherent to treat a light as a (very cooperative) agent with the capability of transmitting current at will, who invariably transmits current when it believes that we want it transmitted and not otherwise; flicking the switch is simply our way of communicating our desires"

The more we know about a system the less we need to rely on intentional explanations of its behaviour

An autonomous agent is a system that is most conveniently described by the intentional stance.

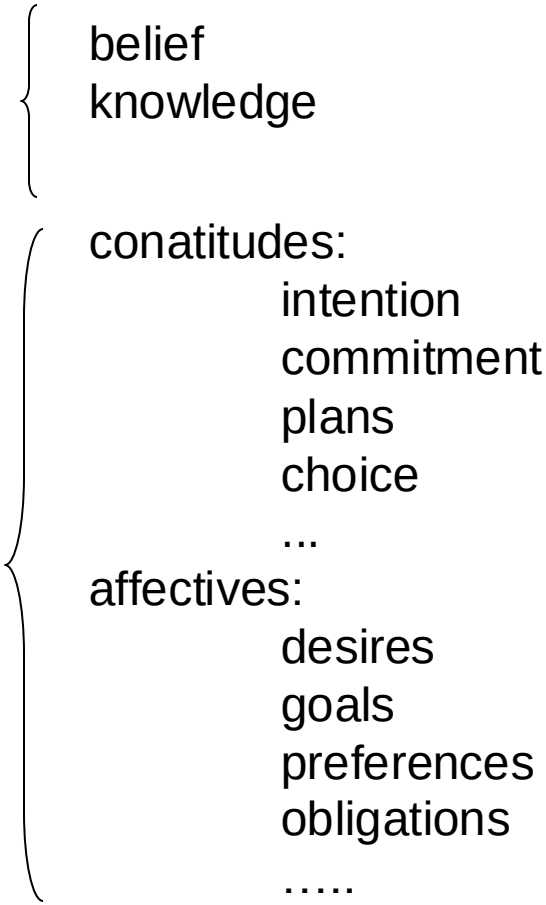
Theories of Attitudes 1

- We want to be able to design and build computer systems in terms of **mentalistic** notions.
- Before we can do this, we need to identify a manageable subset of these attitudes and a model of how they interact to generate system behaviour.

  So first, **which attitudes**?

Theories of Attitudes 2

- Two categories:
  - information
  - attitudes
  - pro-attitudes



Study of Knowledge 1

1. What do we know?
2. What can we know?
3. What does it mean to say someone knows something?
4. What does an agent need to know in order to perform an action?
5. How does an agent know whether it knows enough to perform an action?
6. At what point does an agent know enough to stop gathering information and make a decision?

Study of Knowledge 2

- Individual Perspective
- Group Perspective
  - Knowledge of other agents in the group
  - Everyone knows that everyone knows that everyone knows ... (common knowledge)
  - Distributed Knowledge
    - Alice knows that John is in love with Carol or Suzanne
    - Charlie knows that John is not in love with with Carol

The Wise Men puzzle

There are three wise men.

It's common knowledge -- known by everyone, and known to be known by everyone, etc. -- that there are three red hats and two white hats. The king puts a hat on each of the wise men, and ask them sequentially if they know the color of the hat on their head. Suppose the first man says he does not know; then the second say he does not know either.

It follows that the third man must be able to say that he knows the color of his hat.

Why is this, and what color has the third man's hat?

The Muddy Children puzzle

This is one of the many variations on the wise men puzzle; a difference is that the questions are asked in parallel rather than sequentially.

There is a large group of children playing in the garden. A certain number of children (say k) get mud on their foreheads. Each child can see the mud on others but not on his own forehead. If $k>1$ then each child can see another with mud on its forehead, so each one knows that at least one in the group is muddy.

The Muddy Children puzzle (cntd.)

Consider these two scenarios:

[Scenario 1.] The father repeatedly asks the question 'Does any of you know whether you have mud on your forehead?' The first time they all answer 'no'. But unlike in the wise men example, they don't learn anything by hearing the others answer 'no', so they go on answering 'no' to the father's repeated questions.

[Scenario 2.] The father first announces that at least one of them is muddy (which is something they know already); and then, as before, he repeatedly asks them 'Does any of you know whether you have mud on your own forehead?' The first time they all answer answer 'no'. Indeed they all go on answering 'no' to the first *k-1* repetitions of the same question; but at the *k*-th those with muddy foreheads are able to answer 'yes'.

Foundations of Formal Logic 1

- A formal logic is a game for producing symbolic objects according to given rules.
- Syntax: Alphabet *A*:
  - Variables (X, Y, ..)
  - Constants (a, abc, 15, ...)
  - Functors (f/n)
  - Predicate symbols (p, q, ..)
  - Logical Connectivities ($\neg$, $\vee$, $\wedge$, $\rightarrow$, $\leftrightarrow$)
  - Quantifiers ($\forall$, $\exists$)
  - Auxiliary symbols (,)

Foundations of Formal Logic 2

- Terms (*T*):
  - any constant in *A* is in *T*
  - any variable in *A* is in *T*
  - if *f* is an n-ary functor in *A* and $t_1..t_n \in T$,

    then $f(t_1..t_n) \in T$
  - Examples:
    - john, marry,...
    - X, Y, ...
    - person(john, 1978, male), car(mnb405, person(john, 1978, male))

Foundations of Formal Logic 3

Let T be the set of terms over alphabet A and F is the set of formulae.

- if p is an n-ary predicate symbol and $t_1,...t_n \in$ T then $p(t_1,...t_n) \in$ F
- if H and G $\in$ F so are ($\neg$ H), (H $\vee$ G), (H $\wedge$G), (H $\rightarrow$G) and (H $\leftrightarrow$ G)
- if H $\in$ F and X is a variable then ($\forall$X H) and ($\exists$X H) $\in$ F

Formulae of the form $p(t_1,...t_n)$ are called atomic formulae

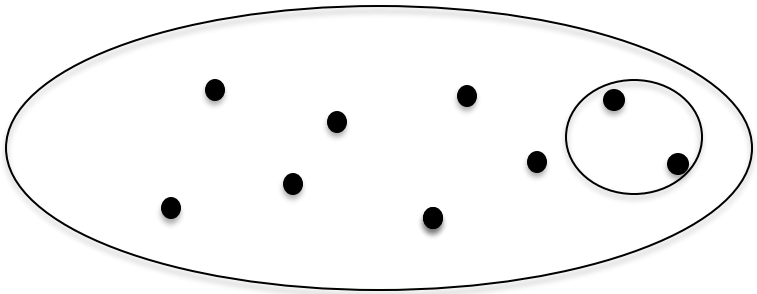
- Examples:
  - owns(john, car(xxx,audi))
  - loves(john,carol) $\vee$ loves(john,suzanne)
  - is_a(student, person) $\wedge$ is_a(employee,person)

Foundations of Formal Logic 4

Interpretation

- An interpretation I of alphabet A is a non-empty domain D and a mapping that associates:

- c $\in$ Const with an element $c_T \in$ D [Const $\subseteq$ A]
- f/n $\in$ Func with an element f_T: Dn $\rightarrow$ D [Func $\subseteq$ A]
- p/m $\in$ Pred with an element $p_T \subseteq$ Dm (= D $\times$... $\times$D) [Pred$\subseteq$ A]



Under an interpretation, each term has a value and each atomic formula is either true or false

Foundations of Formal Logic 3

- Semantics of formulae:
  - Negation: $\neg$A is true if A is false
  - Conjunction: A $\wedge$B is true if both A and B are true
  - Disjunction: A$\vee$B is true if either A or B is true
  - Implication: A$\rightarrow$B is true if either $\neg$A or B are true
  - Universal quantifier $\forall$X: A(x) is true if A is true for every X
  - Existential quantifier $\exists$X: A(x) is true if A is true for some X

Foundations of Formal Logic 4

- Semantics of formulae continued:
  - **Propositional logic** is the logic of connectives, $\neg$, $\vee$, $\wedge$, $\rightarrow$
  - Adding quantifiers give **First-Order Logic**, sometimes called Predicate Calculus
  - Adding quantifiers over formula variables give Higher Order Logic

Foundations of Formal Logic 3

Models and Logical Consequence

An interpretation T is a model of a set of formulae P iff every formula of P is true in T.

- every P has infinitely many interpretations
- not every P has a model (e.g. $F \wedge \neg F$)
- a set of formulae is unsatisfiable if it has no model
- a set of formulae is satisfiable if it has at least one model

Logical consequence

Let P be a set of formulae. F is a logical consequence of P ($P \models F$) iff F is true in every model of P

Foundations of Formal Logic 4

Logical equivalences

F and G are logically equivalent ($F \equiv G$) iff F and G have the same truth value for every interpretation

Some equivalences

- $F \rightarrow G \equiv \neg F \vee G$
- $F \rightarrow G \equiv \neg G \rightarrow \neg F$
- $\neg(F \vee G) \equiv \neg F \wedge \neg G$
- $\neg(F \wedge G) \equiv \neg F \vee \neg G$
- $\neg \forall X\ H(X) \equiv \exists X\ \neg H(X)$
- $\neg \exists X\ H(X) \equiv \forall\ X\ \neg H(X)$
- …

Foundations of Formal Logic 5

Logical inference

Reasoning can be seen as a process of manipulation of formulae, which from a given set of formulae, called premises, produces a new formula, called the conclusion using inference rules

Inference rules

$$\frac{F_1, \ldots, F}{G}$$

$$\frac{F,\ F \rightarrow G}{G}$$

Modus Ponens

$P \vdash F$ means F is derivable from P

If the inference rules are sound then $P \vdash F$ implies $P \models F$

If the inference rules are complete then $P \models F$ implies $P \vdash F$

Formalising Attitudes 1

- How do we formalise attitudes?
- Consider…

  Julie believes Cronos is father of Zeus

- Naive translation into first-order logic:

  believes(julie, father(zeus,cronos))

  – father(zeus, cronos) is a formula of first-order logic and not a term

  ➢ **Need to be able to apply ”believes” to formulae**

Formalising Attitudes 2

- Classical logic allows us to substitute terms with the same denomination:
  - Consider that zeus = jupiter
    believes(janine, father(jupiter,cronos)
  - but believing that father of Zeus is Cronos is not the same as believing that father of Jupiter is Cronos.
  - ➢ **Intentional systems are referentially opaque:**
    Substitution of equivalents into opaque contexts is not going to preserve meaning
  - Standard substitution rules of first-order logic do not apply.
    - (intentional notions are not truth functional)
    - $A \wedge B$ but believes(Janine, p)

Formalising Attitudes 3

- There are 2 sorts of problems to be addressed in developing a logical formalism for intentional notions:

1. **Syntactic**
2. **Semantic**

Formalising Attitudes 4

- Two fundamental approaches to the **syntactic problem**:
  1. Use a modal language, which contains modal operators, which are applied to formulae;
  2. Use a meta-language: a first-order language containing terms that denote formulae of some other object language.
- Two basic approaches to the **semantic problem**:
  1. Possible worlds semantics
  2. Interpreted symbol structures

➢ We will focus on the possible world semantics and modal logic.

Possible Worlds 1

- Intuitive Idea:
  - Besides the true states of affairs, there are a number of other states of affairs, or ”worlds”.
- Each world represents one state of affairs.
- Given his current information, an agent may not be able to tell which of a number of possible worlds describes the correct state of affairs.

➢ An agent’s beliefs can be characterized as a set of possible worlds.

- Can be represented using modal logic.

Possible Worlds 2

- Consider an agent playing a card game (e.g. poker), who possessed the ace of spades.
  - How could the agent deduce what cards were possessed by the opponents?
- First, calculate all the possible ways that the pack of cards could possibly have been distributed among the players.
- Then, systematically eliminate all those configurations which are not possible, given what the agent knows. (e.g. any configuration in which the agent did not possess the ace of spades could be rejected.)

Possible Worlds 3

- Each configuration remaining after this is a **world**;
  - A state of affairs considered possible according to what the agent knows.
- Something that is true in all our agent's possibilities is known by the agent.

How can possible worlds be incorporated into the semantic framework of logic?

Modal Logic 1

- Modal logic was used by philosophers to investigate different **modes of truth**,
  - e.g. *possibly* true, *necessarily* true
- In the study of agents, it is used to give meaning to concepts such as belief and knowledge.

Modal Logic 2

- Modal logic can be considered as the logical theory of necessity and possibility
- It is essentially classical propositional logic extended by two operators
  - $\Box$ necessity
  - $\Diamond$ possibility
- Examples:
  - ”it is necessary that the sun rises in the east” – $\Box$sun-rises-in-the-east
  - ”it is possible that it rains” - $\Diamond$rain

Modal Logic 3

Syntax:

Let S = {p, q, ... } be a set of atomic propositions

- If $p \in S$ then p is a formula
- If A, B are formulae, then so are $\neg A$ and $A \wedge B$
- If A is a formulae, then so are $\Box A$ and $\Diamond A$

Other connectives can be expressed by abbreviations, e.g.

- $F \rightarrow G \equiv \neg F \vee G$
- $\neg(F \wedge G) \equiv \neg F \vee \neg G$

Modal Logic 4

- Duality of operators
  - $\Box A \equiv \neg \Diamond \neg A$
  - $\Diamond A \equiv \neg \Box \neg A$

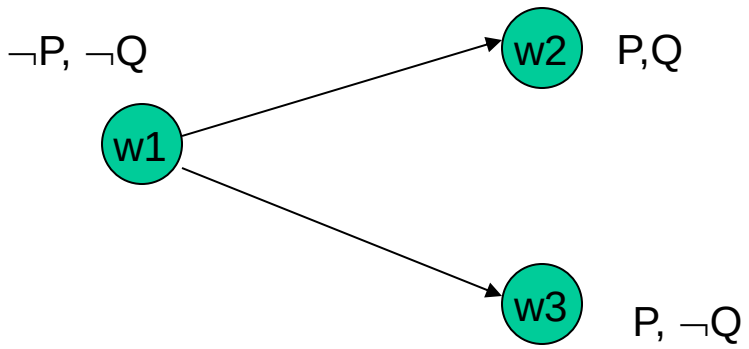
- Two Basic Properties
  1. **K** axiom schema: $\Box(A \rightarrow B) \rightarrow (\Box A \rightarrow \Box B)$ *(K in honour of Kripke)*
  2. Necessitation Rule: if A is valid, then $\Box A$ is valid

Modal Logic 5

- The semantics of modal logic is traditionally given in terms of **possible worlds**.
  - The formula □**A** is true if A is true in **every** world accessible from the current world
  - The formula ◊**A** is true if A is true in **at least one** world accessible from the current world
- With sets of worlds as primitive, the structure of the model is captured by relating the different worlds via a binary **accessibility relation**.

Modal Logic 6

- Formalizing possible worlds (Kripke structure):
  - $(S, \pi, K_1 \ldots K_n)$
    - S – set of possible worlds
    - π – set of formulae true at a world
    - K_i – a binary accessibility relation on S (a set of pairs of elements of S)



Worlds *w2* and *w3* are accessible from world *w1*.

w1: $\Box P$

$\Diamond Q$

Modal Logic 7

- Possible properties of accessibility relations:
  - **Reflexive**, for all $s \in S$, we have $(s,s) \in K$
  - **Symmetric**, for all $s,t \in S$, we have $(s,t) \in K$ iff $(t,s) \in K$
  - **Transitive**, for all $s,t,u \in S$, we have that if

    $(s,t) \in K$ and $(t,u) \in K$, then $(s,u) \in K$
  - **Serial**, for all $s \in S$, there is some t such that $(s,t) \in K$
  - **Euclidian**, for all $s,t,u \in S$, whenever

    $(s,t) \in K$ and $(s,u) \in K$, then $(t,u) \in K$

Modal Logic

If K is reflexive and Euclidean, then K is symmetric and transitive

If K is symmetric and transitive, then K is Euclidian

The following are equivalent

- K is reflexive, symmetric and transitive
- K is symmetric, transitive and serial
- K is reflexive and Euclidean

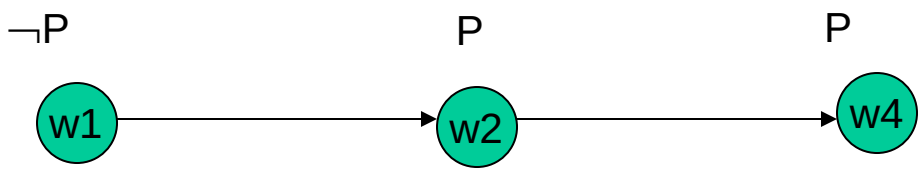
Modal Logic 7

- Properties of accessibility relation are represented by axiom schemas:
  - **T** axiom : corresponds to reflexive accessibility relation
    - $\Box A \rightarrow A$
  - **D** axiom : corresponds to serial accessibility relation
    - $\Box A \rightarrow \Diamond A$
  - **4** axiom : corresponds to transitive accessibility relation
    - $\Box A \rightarrow \Box\Box A$
  - **5** axiom : corresponds to Euclidean accessibility relation
    - $\Diamond A \rightarrow \Box\Diamond A$

S5(KT5) S4(KT4) T(KT) weak-S5(KD45)

Modal Logic 8

Transitive (but not reflexive):

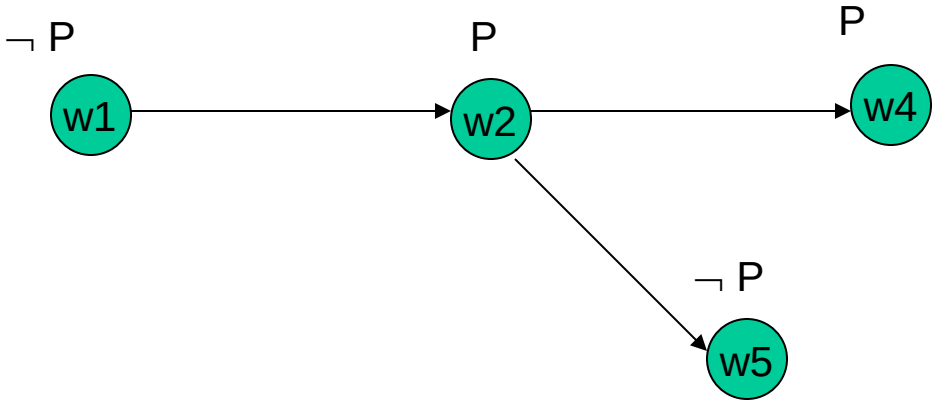


w1: □P

w2: □P

w1: □□P

Another relation



w1: □P

w2: ◊P

w1: □◊P

Logic of knowledge 1

- The formula □**A** is read as ”it is known that A” or ”agent knows A” and denoted as K
- For group knowledge, we have an indexed set of modal operators
- K_1, .., K_n for □
- **K_1 A** is read as ”agent1 knows A”

Logic of knowledge 2

- Some examples:

  $K_1K_2p \wedge \neg K_2K_1K_2p$

  – *Agent1* knows that *Agent2* knows p, but *Agent2* doesn't know that *Agent1* knows that *Agent2* knows p

  $\neg K_1 \neg (K_2 K_1K_2p) \wedge \neg K_1 \neg(\neg K_2 K_1K_2p)$

  – *"Dean* doesn't know whether *Nixon* knows that *Dean* knows that *Nixon* knows that *McCord burgled O'Brien's office at Watergate"*

Modal Logic and Knowledge and Belief 1

- T axiom (Knowledge axiom) $K_iA \rightarrow A$ ($\Box A \rightarrow A$)
  - what is known is true
- D axiom $K_iA \rightarrow \neg K_i \neg A$ ($\Box A \rightarrow \Diamond A$)
  - if *i* knows *A* then *i* doesn't know $\neg A$
- 4 axiom (positive introspection) $K_iA \rightarrow K_iK_iA$ ($\Box A \rightarrow \Box\Box A$)
  - if *i* knows *A* then *i* knows that it knows *A*
- 5 axiom (negative introspection) $\neg K_i \neg A \rightarrow K_i \neg K_i \neg A$ ($\Diamond A \rightarrow \Box\Diamond A$)
  - *i* is aware of what it doesn't know

Modal Logic and Knowledge and Belief 2

- Knowledge is often defined as true belief:
  - agent knows A if agent believes A and A is true.
- Axioms KTD45 are often chosen as a logic of knowledge
- Axioms KD45 are often chosen as a logic of belief

Modal Logic and Knowledge and Belief 3

How well does normal modal logic serves as a logic of knowledge and belief?

- Consider the K axiom and the necessitation rule:

  K axiom schema: $\Box(A \rightarrow B) \rightarrow (\Box A \rightarrow \Box B)$

  Necessitation Rule: if A is valid, then $\Box A$ is valid

  - Necessitation rule: an agent knows all valid formulae, (an agent will have an infinite number of items of knowledge).
  - K axiom: agent's knowledge is closed under implication.

Modal Logic and Knowledge and Belief 4

- **Logical omniscience** – knowing all truth of logic
- **Logical omniscience problem** – constituted by that of knowing all valid formulae and that of knowledge/belief being closed under consequence (must know all logical consequences of one's knowledge or belief).
- Disadvantages of using possible world semantics for agents are:
  - agents believe/know all valid formulae
  - agents' beliefs/knowledge are closed under logical consequence

Other approaches

(Levesque)

In order to avoid logical omniscience problem a distinction is made between explicit and implicit belief

(Konolige)

Deduction model of belief

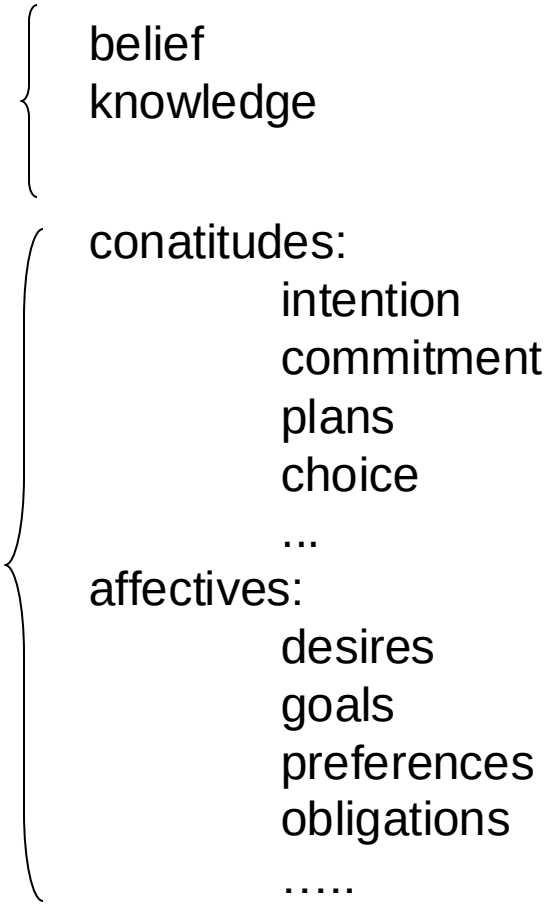
The deduction model defines a belief system as a tuple $d=(\Delta, \rho)$

containing a "base set" of formulae in some internal, logical language of belief, and a set of deduction rules (may be incomplete) for deriving new beliefs

Agent with such a belief system believes A if A can be derived from its base set using its deduction rules

Theories of Attitudes 2

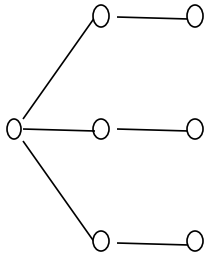
- Two categories:
  - information
  - attitudes
  - pro-attitudes



BDI Architecture 1-
belief, desire, intentions (Rao, Georgeff)

- Systems and formalisms that give primary importance to intentions are often referred to as **BDI-architecture**.
- Formalization of intentions based on branching-time possible worlds future and single past model.
- Crucial elements are:
  - Intentions are treated on a pair with beliefs and goals.
  - Distinguishes between choices and the possibilities of the different outcomes of actions.
  - Interrelationship between beliefs, goals and intentions are specified.
    - (Goals are consistent desires of an agent.)

BDI Architecture 2

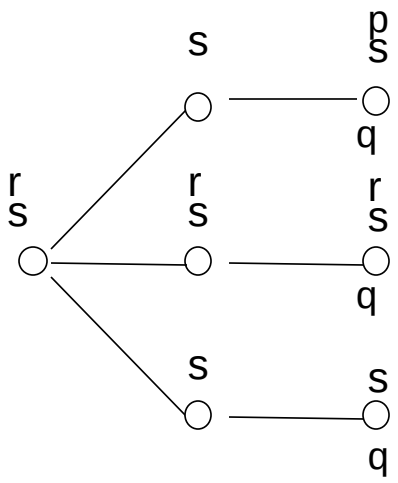


- Informal semantics:
  - The world is modeled by using a temporal structure with a branching time future and a single past – this is called a **time tree**.
  - A particular time point in a particular world is a **situation**.
  - Event types transform one time point to another.
  - Primitive events are those events directly performable by the agent and uniquely determine the next time point.
  - The branches in a time tree represent the **choices** available to an agent.

BDI Architecture 3

- Uses 2 **modal operators**:
  - Optional: a path formula is said to be *optional* if, at a particular time point in a time tree, it is true of at least one path emanating from that point.
  - Inevitable: a path formula is said to be *inevitable* if it is true of all paths emanating from that point.
- **Temporal operators**: *next*, *eventually*, *always* and *until*.
- A combination of these modalities can be used to describe the options available to an agent.

BDI Architecture 4



optionally eventually p

optionally always r

inevitable eventually q

inevitably always s

p: it is optional that John will eventually visit London

r: it is optional that Mary will always live in Australia

q: it is inevitable that the world will eventually come to an end

s: it is inevitable that one plus one will always be two

BDI Architecture

BELIEFS

In each situation a set of belief-accessible worlds is associated – those worlds that the agent believes to be possible.

- Each belief-accessible world is a time-tree

GOALS/DESIRES

For each situation a set of goal-accessible worlds is associated - those represent the goals of the agent

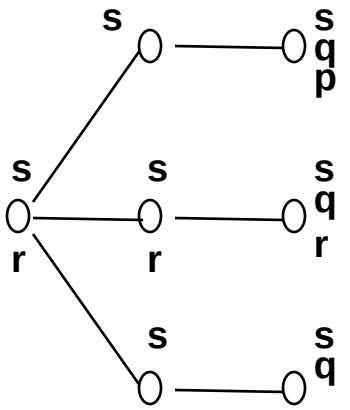
- goals are chosen desires of the agent that are consistent and agent should believe that the goal is achievable; goals must be compatible with beliefs
- for each belief-accessible world w in time t, there must be a goal-accessible sub-world of w at time t

BDI Architecture 6

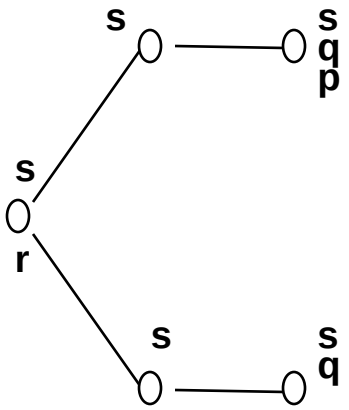
INTENIONS:

intentions are represented by a set of intention-accessible worlds.

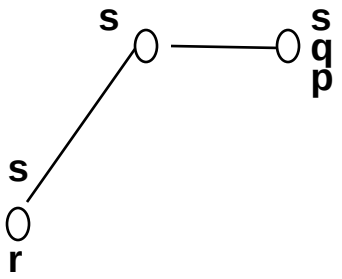
- These worlds are ones that an agent has committed to attempt to realize.
- The intention-accessible worlds of an agent must be compatible with its goal-accessible worlds.
- For each goal-accessible world *w* in time *t*, there must be an intention-accessible sub-world of *w* at time *t.*



belief-accessible world



goal-accessible world



intentions-accessible world

Summary

1. Agents as intentional systems
2. Modal logics are often used
3. Belief-Desire-Intetions (BDI) is most used set of attitudes

Next Lecture: Agent architectures

Uses: Wooldridge: ”Introduction to MultiAgent Systems”

Chapters 2,3,4,5