

Lecture 6: Practical Reasoning Agents

Autonomous Agents and Multiagent Systems
DIS, La Sapienza - PhD Course

Sebastian Sardina¹

¹Department of Computer Science and Information Technology
RMIT University
Melbourne, AUSTRALIA



November 22, 2007

Roadmap for Next Lectures

Lecture 6

Practical Reasoning

Bratman, Dennett
IRMA, PRS

Roadmap for Next Lectures

Lecture 6

Practical Reasoning

Bratman, Dennett
IRMA, PRS

Lecture 7

Formal Models of
BDI Programming

CAN

Roadmap for Next Lectures

Lecture 6

Practical Reasoning

Bratman, Dennett
IRMA, PRS

Lecture 7

Formal Models of
BDI Programming

CAN

Lecture 8

Declarative Goals

Planning in BDI

CANPlan

Roadmap for Next Lectures

Lecture 6

Practical Reasoning

Bratman, Dennet
IRMA, PRS

Lecture 7

Formal Models of
BDI Programming

CAN

Lecture 8

Declarative Goals

Planning in BDI

CANPlan

Lecture 9

Practical BDI
Programming I

JACK

Roadmap for Next Lectures

Lecture 6

Practical Reasoning

Bratman, Dennet
IRMA, PRS

Lecture 7

Formal Models of
BDI Programming

CAN

Lecture 8

Declarative Goals

Planning in BDI

CANPlan

Lecture 9

Practical BDI
Programming I

JACK

Lecture 10

Practical BDI
Programming II

JACK & PDT

Roadmap for Next Lectures

Lecture 6

Practical Reasoning

Bratman, Dennet
IRMA, PRS

Lecture 7

Formal Models of
BDI Programming

CAN

Lecture 8

Declarative Goals

Planning in BDI

CANPlan

Lecture 9

Practical BDI
Programming I

JACK

Lecture 10

Practical BDI
Programming II

JACK & PDT

Outline

- 1 Introduction
- 2 The Process of Deliberation
 - Desires
 - Intentions
 - Plans
- 3 Commitment
 - Example
 - Strategies
 - Reconsideration
- 4 Agent Architectures
- 5 Agent Theory
- 6 Conclusions

Next Topic:

Introduction

Intelligent Agents

An intelligent (software) agent is an **autonomous** entity, existing over time in a dynamic environment, that is able to **rationally** balance **pro-active** and **reactive** behavior.

autonomous: does not require continuous external control;

pro-active: pursues goals over time; goal directed behavior;

situated: observe & act in the environment;

reactive: perceives the environment and responds to it.

Other features: flexible, robust, social, etc.

Agents are not Objects!

1 Agents are autonomous:

- ▶ they decide for themselves whether or not to perform an action on request from another agent.

2 Agents are smart:

- ▶ capable of flexible (reactive, pro-active, social) behavior, and
- ▶ standard object model has nothing to say about such types of behavior.

3 Agents are active:

- ▶ a multi-agent system is inherently multi-threaded;
- ▶ each agent is assumed to have at least one thread of active control.

Agents are not Objects!

1 Agents are autonomous:

- ▶ they decide for themselves whether or not to perform an action on request from another agent.

2 Agents are smart:

- ▶ capable of flexible (reactive, pro-active, social) behavior, and
- ▶ standard object model has nothing to say about such types of behavior.

3 Agents are active:

- ▶ a multi-agent system is inherently multi-threaded;
- ▶ each agent is assumed to have at least one thread of active control.

Objects do it for free; agents do it because they want to!

Understanding Behavior

Philosopher **Daniel Dennett** defines three levels of abstraction to explain and predict the behavior of an object:

Understanding Behavior

Philosopher **Daniel Dennett** defines three levels of abstraction to explain and predict the behavior of an object:

- 1 **Physical Stance:** at the level of **physics** and **chemistry**.
 - ▶ concerned with things such as mass, energy, velocity, and chemical comp.;
 - ▶ e.g., **predict a thermometer based on the chemical properties of mercury.**

Understanding Behavior

Philosopher **Daniel Dennett** defines three levels of abstraction to explain and predict the behavior of an object:

- 1 **Physical Stance:** at the level of **physics** and **chemistry**.
 - ▶ concerned with things such as mass, energy, velocity, and chemical comp.;
 - ▶ e.g., predict a thermometer based on the chemical properties of mercury.
- 2 **Design Stance:** at the level of **biology** and **engineering**.
 - ▶ concerned with things such as purpose, function, and design;
 - ▶ e.g., predict that a bird will fly when it flaps its wings, on the basis that wings are made for flying.

Understanding Behavior

Philosopher **Daniel Dennett** defines three levels of abstraction to explain and predict the behavior of an object:

- 1 **Physical Stance:** at the level of **physics** and **chemistry**.
 - ▶ concerned with things such as mass, energy, velocity, and chemical comp.;
 - ▶ e.g., predict a thermometer based on the chemical properties of mercury.
- 2 **Design Stance:** at the level of **biology** and **engineering**.
 - ▶ concerned with things such as purpose, function, and design;
 - ▶ e.g., predict that a bird will fly when it flaps its wings, on the basis that wings are made for flying.
- 3 **Intentional Stance:** at the level of **software** and **minds**.
 - ▶ we are concerned with things such as belief, thinking and intent;
 - ▶ e.g., predict that the bird will fly away because it knows the cat is coming;
 - ▶ e.g., predict that Mary will leave the theater and drive to the restaurant because she sees that the movie is over and is hungry.

Understanding Behavior

Philosopher **Daniel Dennett** defines three levels of abstraction to explain and predict the behavior of an object:

3 Intentional Stance: at the level of **software** and **minds**.

- ▶ we are concerned with things such as belief, thinking and intent;
- ▶ e.g., predict that the bird will fly away because it knows the cat is coming;
- ▶ e.g., predict that Mary will leave the theater and drive to the restaurant because she sees that the movie is over and is hungry.

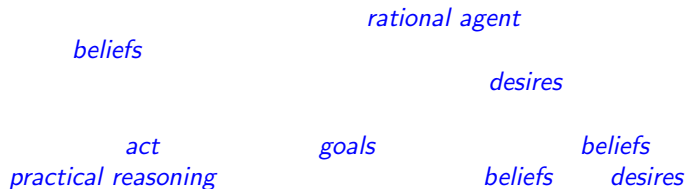
Dennett argues that it is best to understand human's behavior at the level of the intentional stance.

The Intentional Stance

*“Here is how it works: first you decide to treat the object whose behavior is to be predicted as a **rational agent**; then you figure out what **beliefs** that agent ought to have, given its place in the world and its purpose. Then you figure out what **desires** it ought to have, on the same considerations, and finally you predict that this rational agent will **act** to further its **goals** in the light of its **beliefs**. A little **practical reasoning** from the chosen set of **beliefs** and **desires** will in most instances yield a decision about what the agent ought to do; that is what you predict the agent will do.”*

(Daniel Dennett, The Intentional Stance, p. 17)

The Intentional Stance



(Daniel Dennett, The Intentional Stance, p. 17)

The Intentional Stance

Two major ideas:

- 1 Rational behavior can be understood in terms of **mental properties** (i.e., propositional attitude):
 - ▶ beliefs (“he thinks that Peter is wise”);
 - ▶ desires & goals (“she wants that piece of cake”);
 - ▶ fear (“Alex is afraid of spiders”);
 - ▶ hopes (“she hopes that he is on time today”);
 - ▶ ...
- 2 Rational behavior relies on a special kind of “thinking”
 - ▶ **practical reasoning**

So, Dennett coined the term **intentional system** to describe entities *“whose behavior can be predicted by the method of attributing belief, desires and rational acumen.”*

Taking the Intentional Stance: A Student



- ▶ **Rational agent:** a student.
- ▶ **Beliefs:** 5pm now; game at 5:10pm; library closes 5:30pm; no book; ...
- ▶ **Desires:** read book; watch game; ...
- ▶ **Practical reasoning** states to go to the library now!

Intentional Stance for Computer Systems?

John McCarthy argued that there are occasions when the intentional stance is appropriate:

“To ascribe beliefs, free will, intentions, consciousness, abilities, or wants to a machine is legitimate when such an ascription expresses the same information about the machine that it expresses about a person. It is useful when the ascription helps us understand the structure of the machine, its past or future behavior, or how to repair or improve it. [...]”

Taking the Intentional Stance: A Thermometer



- ▶ **Rational agent:** thermometer.
- ▶ **Beliefs:** the temperature of the air around.
- ▶ **Desires:** to tell the current temperature.
- ▶ **Practical reasoning** states that if n is believed to be the current temperature, n is displayed.

Taking the Intentional Stance: A Thermometer



- ▶ **Rational agent:** thermometer.
- ▶ **Beliefs:** the temperature of the air around.
- ▶ **Desires:** to tell the current temperature.
- ▶ **Practical reasoning** states that if n is believed to be the current temperature, n is displayed.

But most adults would find such a description absurd!
Why is this?

When is the Intentional Stance Useful?

- ▶ The answer seems to be that while the intentional stance description is consistent,
. . . it does not buy us anything, since we essentially understand the mechanism sufficiently to have a simpler, mechanistic description of its behavior. (Yoav Shoham)

When is the Intentional Stance Useful?

- ▶ The answer seems to be that while the intentional stance description is consistent,
. . . it does not buy us anything, since we essentially understand the mechanism sufficiently to have a simpler, mechanistic description of its behavior. (Yoav Shoham)
- ▶ So: the more we know about a system, the less we need to rely on animistic, intentional explanations of its behavior.
- ▶ But with very complex systems, a mechanistic, explanation of its behavior may not be practicable.

When is the Intentional Stance Useful?

- ▶ The answer seems to be that while the intentional stance description is consistent,
. . . it does not buy us anything, since we essentially understand the mechanism sufficiently to have a simpler, mechanistic description of its behavior. (Yoav Shoham)
- ▶ So: the more we know about a system, the less we need to rely on animistic, intentional explanations of its behavior.
- ▶ But with very complex systems, a mechanistic, explanation of its behavior may not be practicable.
- ▶ As computer systems become ever more complex, we need more powerful abstractions and metaphors to explain their operation — low level explanations become impractical.

The intentional stance is such an abstraction!

The Intentional Stance

Two major ideas:

- 1 Rational behavior can be understood in terms of **mental properties** (i.e., propositional attitude):
 - ▶ beliefs (“he thinks that Peter is wise”);
 - ▶ desires & goals (“she wants that piece of cake”);
 - ▶ fear (“Alex is afraid of spiders”);
 - ▶ hopes (“she hopes that he is on time today”);
 - ▶ ...
- 2 Rational behavior relies on a special kind of “thinking”
 - ▶ **practical reasoning**

So, Dennett coined the term **intentional system** to describe entities *“whose behavior can be predicted by the method of attributing belief, desires and rational acumen.”*

What is Practical Reasoning?

- ▶ Practical reasoning is reasoning **directed towards actions** — the process of figuring out what to do.
- ▶ Principles of practical reasoning applied to agents largely derive from work of philosopher **Michael Bratman (1990)**:

“Practical reasoning is a matter of weighing conflicting considerations for and against competing options, where the relevant considerations are provided by what the agent desires/values/cares about and what the agent believes.”



- ▶ Distinguish practical reasoning from **theoretical reasoning**.

Theoretical vs Practical Reasoning

*"In theory, there is no difference between theory and practice.
But, in practice, there is." – Jan L. A. van de Snepscheut*

- 1 Theoretical reasoning is reasoning directed towards beliefs — concerned with deciding what to believe.
 - ▶ Tries to assess the way things are.
 - ▶ Process by which you change your beliefs and expectations;.
 - ▶ Example: you believe q if you believe p and you believe that if p then q .
- 2 Practical reasoning is reasoning directed towards actions — concerned with deciding what to do.
 - ▶ Decides how the world should be and what individuals should do.
 - ▶ Process by which you change your choices, plans, and intentions.
 - ▶ Example: you go to class, if you must go to class.

The Components of Practical Reasoning

Human practical reasoning consists of two activities:

- 1 **Deliberation**: deciding **what** state of affairs we want to achieve.
- 2 **Means-ends reasoning**: deciding **how** to achieve these states of affairs:

The Components of Practical Reasoning

Human practical reasoning consists of two activities:

- 1 **Deliberation**: deciding **what** state of affairs we want to achieve.
 - ▶ considering preferences, choosing goals, etc.;
 - ▶ balancing alternatives (decision-theory);
 - ▶ the outputs of deliberation are **intentions**;
 - ▶ interface between deliberation and means-end reasoning.
- 2 **Means-ends reasoning**: deciding **how** to achieve these states of affairs:

The Components of Practical Reasoning

Human practical reasoning consists of two activities:

- 1 **Deliberation**: deciding **what** state of affairs we want to achieve.
 - ▶ considering preferences, choosing goals, etc.;
 - ▶ balancing alternatives (decision-theory);
 - ▶ the outputs of deliberation are **intentions**;
 - ▶ interface between deliberation and means-end reasoning.
- 2 **Means-ends reasoning**: deciding **how** to achieve these states of affairs:
 - ▶ thinking about suitable actions, resources and how to “organize” activity;
 - ▶ building courses of action (planning);
 - ▶ the outputs of means-ends reasoning are **plans**.

The Components of Practical Reasoning

Human practical reasoning consists of two activities:

- 1 **Deliberation**: deciding **what** state of affairs we want to achieve.
 - ▶ considering preferences, choosing goals, etc.;
 - ▶ balancing alternatives (decision-theory);
 - ▶ the outputs of deliberation are **intentions**;
 - ▶ interface between deliberation and means-end reasoning.
- 2 **Means-ends reasoning**: deciding **how** to achieve these states of affairs:
 - ▶ thinking about suitable actions, resources and how to “organize” activity;
 - ▶ building courses of action (planning);
 - ▶ the outputs of means-ends reasoning are **plans**.

Fact: agents are **resource-bounded** & world is **dynamic**!

The key: To combine **deliberation** & **means-ends reasoning** appropriately.

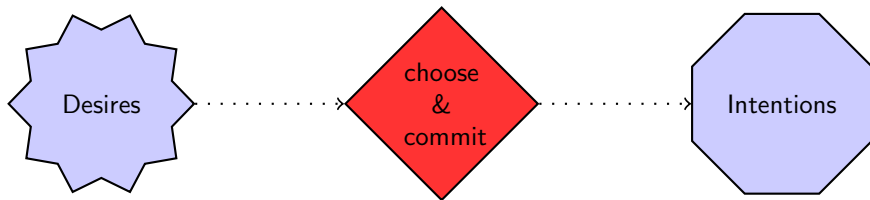
Next Topic:

The Process of Deliberation

Deliberation

How does an agent **deliberate**?

- 1 Begin by trying to understand what the **options** available to you are:
 - ▶ options available are **desires**.
- 2 **Choose** between them, and **commit** to some:
 - ▶ chosen options are then **intentions**.



Desires

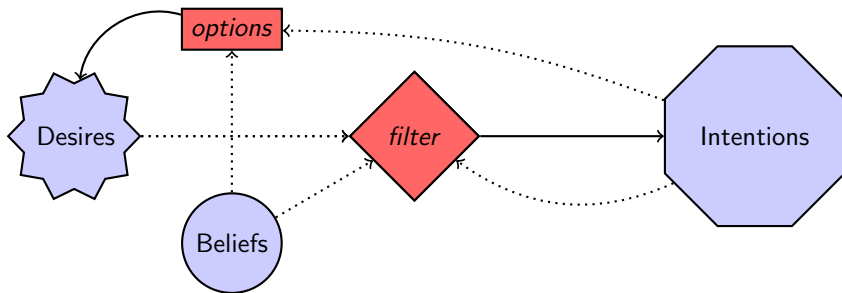
- ▶ Desires describe the states of affairs that are considered for achievement, i.e., **basic preferences** of the agent.
- ▶ Desires are much **weaker** than intentions; not directly related to activity:

*“My **desire** to play basketball this afternoon is **merely a potential influence** of my conduct this afternoon. It must vie with my other relevant desires [...] before it is settled what I will do. In contrast, once I **intend** to play basketball this afternoon, **the matter is settled**: I normally need not continue to weigh the pros and cons. When the afternoon arrives, I will normally just proceed to execute my intentions.”*
(Bratman 1990)

Functional Components of Deliberation

Option Generation agent generates a set of possible alternatives; via a function, *options*, which takes the agent's current beliefs and intentions, and from them determines a set of options/desires.

Filtering in which the agent chooses between competing alternatives, and commits to achieving them. In order to select between competing options, an agent uses a *filter* function.

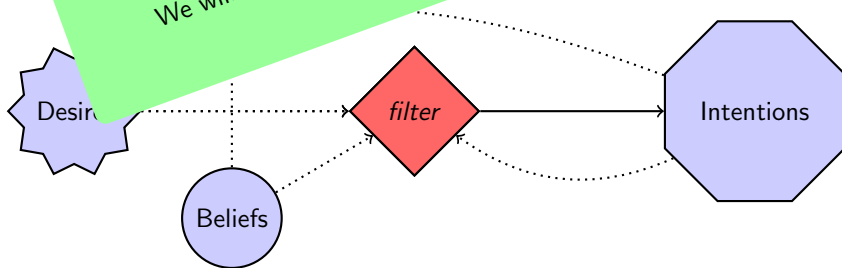


Functional Components of Deliberation

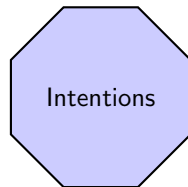
Option Generation agent generates a set of possible alternatives; via a function, *options*, which takes the agent's current beliefs and intentions, and from them produces a set of options/desires.

Filtering in which the agent evaluates competing alternatives, and chooses the one to select between them via a *filter* function.

We will refine this further later...



Functional Components of Deliberation



Intentions

- ▶ In ordinary speech: intentions refer to **actions** or to **states of mind**;
 - ▶ here we consider the latter!
 - ▶ E.g., **I may adopt/have the intention to be an academic.**
- ▶ Focus on **future-directed** intentions i.e. **pro-attitudes** leading to actions.
 - ▶ Intentions are about the (desired) future.
- ▶ We make **reasonable attempts to fulfill** intentions once we form them, but they may change if circumstances do.
 - ▶ Behavior arises to fulfill intentions.
 - ▶ Intentions affect action choice.

Main Properties of Intentions

1 Intentions drive means-end reasoning.

If I adopt the intention to become an academic, then I will decide how to achieve it, for example, by applying to a PhD programme. Moreover, if I fail to gain a PhD place at one university, I might try another university.

Main Properties of Intentions

1 Intentions drive means-end reasoning.

If I adopt the intention to become an academic, then I will decide how to achieve it, for example, by applying to a PhD programme. Moreover, if I fail to gain a PhD place at one university, I might try another university.

2 Intentions constrain future deliberation (i.e., provide a “filter”).

If I intend to be an academic, I would not consider being rich.

Main Properties of Intentions

1 Intentions drive means-end reasoning.

If I adopt the intention to become an academic, then I will decide how to achieve it, for example, by applying to a PhD programme. Moreover, if I fail to gain a PhD place at one university, I might try another university.

2 Intentions constrain future deliberation (i.e., provide a “filter”).

If I intend to be an academic, I would not consider being rich.

3 Intentions persist.

If I intend to become an academic, I will not give up without a good reason—e.g., if I already obtained an academic position or realized that it will be impossible to be an academic.

Main Properties of Intentions

1 Intentions drive means-end reasoning.

If I adopt the intention to become an academic, then I will decide how to achieve it, for example, by applying to a PhD programme. Moreover, if I fail to gain a PhD place at one university, I might try another university.

2 Intentions constrain future deliberation (i.e., provide a “filter”).

If I intend to be an academic, I would not consider being rich.

3 Intentions persist.

If I intend to become an academic, I will not give up without a good reason—e.g., if I already obtained an academic position or realized that it will be impossible to be an academic.

4 Intentions influence beliefs concerning future practical reasoning.

If I adopt the intention to be an academic, then I can plan for the future on the assumption that I will be an academic.

Main Properties of Intentions (cont.)

- 5 Agents believe their intentions are possible.

If I intend to be an academic, there must be some conceivable way for me to become an academic (e.g., getting a PhD and applying for a university position).

Main Properties of Intentions (cont.)

- 5 Agents believe their intentions are possible.

If I intend to be an academic, there must be some conceivable way for me to become an academic (e.g., getting a PhD and applying for a university position).

- 6 Agents do not believe they will not bring about their intentions.

It would not be rational of me to adopt an intention to become an academic if I believed I would fail becoming one.

Main Properties of Intentions (cont.)

- 5 Agents believe their intentions are possible.

If I intend to be an academic, there must be some conceivable way for me to become an academic (e.g., getting a PhD and applying for a university position).

- 6 Agents do not believe they will not bring about their intentions.

It would not be rational of me to adopt an intention to become an academic if I believed I would fail becoming one.

- 7 Under certain circumstances, agents believe they will bring about their intentions

If I intend to be an academic, then I believe that under "normal circumstances" I will indeed become an academic.

Main Properties of Intentions (cont.)

- 8 Agents need not intend all the expected side effects of their intentions

If I believe $\phi \implies \psi$ and I intend that ϕ , I do not necessarily intend ψ also.

So, intentions are not closed under implication!

This last problem is known as the **side effect** or **package** deal problem:

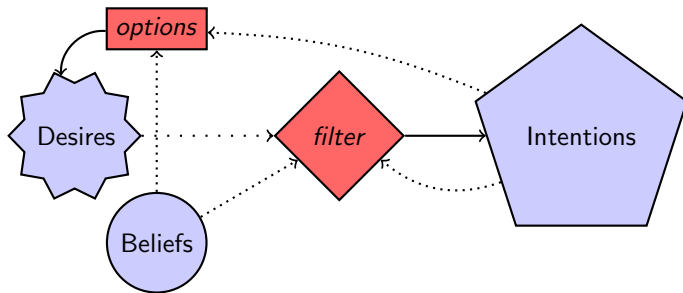
I may believe that going to the dentist involves pain, and I may also intend to go to the dentist — but this does not imply that I intend to suffer pain!

Plans

Human practical reasoning consists of two activities:

- 1 **Deliberation**: deciding **what** to do. Forms **intentions**.
- 2 **Means-ends reasoning**: deciding **how** to do it. Forms plans. Forms **plans**.

Intentions drive means-ends reasoning: *If I adopt an intention, I will attempt to achieve it, this affects action choice.*

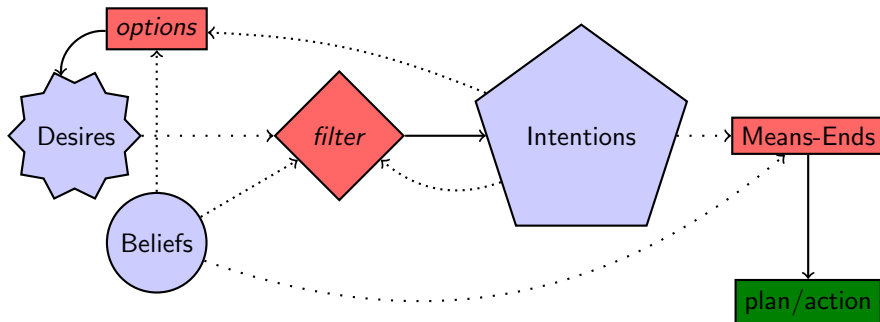


Plans

Human practical reasoning consists of two activities:

- 1 **Deliberation**: deciding **what** to do. Forms **intentions**.
- 2 **Means-ends reasoning**: deciding **how** to do it. Forms plans. Forms **plans**.

Intentions drive means-ends reasoning: *If I adopt an intention, I will attempt to achieve it, this affects action choice.*



Means-End Reasoning: Obtaining Plans & Actions

How does the agent obtain plans/actions to realize our intentions?

✓ Planning: design a course of action that will achieve the goal. Given:

- 1 (representation of) goal/intention to achieve;
- 2 (representation of) actions it can perform; and
- 3 (representation of) the environment;

... have it generate a plan to achieve the goal.

This is **automatic** programming
This is **hard** (PSPACE-complete)!

Means-End Reasoning: Obtaining Plans & Actions

How does the agent obtain plans/actions to realize our intentions?

✓ Planning: design a course of action that will achieve the goal. Given:

- 1 (representation of) goal/intention to achieve;
- 2 (representation of) actions it can perform; and
- 3 (representation of) the environment;

... have it generate a plan to achieve the goal.

This is **automatic** programming

This is **hard** (PSPACE-complete)!

✓ High-level programming (e.g., Golog, ConGolog, IndiGolog).

This is **semi-automatic/hybrid** programming.

Means-End Reasoning: Obtaining Plans & Actions

How does the agent obtain plans/actions to realize our intentions?

✓ Planning: design a course of action that will achieve the goal. Given:

- 1 (representation of) goal/intention to achieve;
- 2 (representation of) actions it can perform; and
- 3 (representation of) the environment;

... have it generate a plan to achieve the goal.

This is **automatic** programming

This is **hard** (PSPACE-complete)!

✓ High-level programming (e.g., Golog, ConGolog, IndiGolog).

This is **semi-automatic/hybrid** programming.

✓ BDI-style programming (e.g., AgentSpeak, CAN, Jason, JACK, etc.)

This is **implicit** programming.

Next Topic:

Commitment

Willie: The Committed Assistant

Some time in the not-so-distant future, you are having trouble with your new household robot. You say “Willie, bring me a beer.” The robot replies “OK boss.” Twenty minutes later, you screech “Willie, why didn’t you bring me that beer?” It answers “Well, I intended to get you the beer, but I decided to do something else.” Miffed, you send the wise guy back to the manufacturer, complaining about a **lack of commitment**.

After retrofitting, Willie is returned, marked “Model C: The Committed Assistant.” Again, you ask Willie to bring you a beer. Again, it accedes, replying “Sure thing.” Then you ask: “What kind of beer did you buy?” It answers: “Genessee.” **You say “Never mind.” One minute later, Willie trundles over with a Genessee in its gripper.**

Willie: The Committed Assistant (cont.)

[...] After still more tinkering, the manufacturer sends Willie back, promising no more problems with its commitments. So, being a somewhat trusting customer, you accept the rascal back into your household, but as a test, you ask it to bring you your last beer. [...] The robot gets the beer and starts towards you. As it approaches, it lifts its arm, wheels around, **deliberately smashes the bottle, and trundles off.**

Back at the plant, when interrogated by customer service as to why it had abandoned its commitments, the robot replies that according to its specifications, it kept its commitments as long as required — commitments must be dropped when fulfilled or impossible to achieve. By smashing the bottle, the **commitment became unachievable.**

What is Wrong with Willie?

Some time in the not-so-distant future, you are having trouble with your new household robot. You say “Willie, bring me a beer.” The robot replies “OK boss.” Twenty minutes later, you screech “Willie, why didn’t you bring me that beer?” It answers “Well, I intended to get you the beer, but I decided to do something else.” Miffed, you send the wise guy back to the manufacturer, complaining about a **lack of commitment**.

After retrofitting, Willie is returned, marked “Model C: The Committed Assistant.” Again, you ask Willie to bring you a beer. Again, it accedes, replying “Sure thing.” Then you ask: “What kind of beer did you buy?” It answers: “Genessee.” **You say “Never mind.” One minute later, Willie trundles over with a Genessee in its gripper.**

After still more tinkering, the manufacturer sends Willie back, promising no more problems with its commitments. So, being a somewhat trusting customer, you accept the rascal back into your household, but as a test, you ask it to bring you your last beer. [...] The robot gets the beer and starts towards you. As it approaches, it lifts its arm, wheels around, **deliberately smashes the bottle, and trundles off.**

What is Wrong with Willie?

Some time in the not-so-distant future, you are having trouble with your new household robot. You say “Willie, bring me a beer.” The robot replies “OK boss.” Twenty minutes later, you say “Willie, I don’t see the beer. Can’t you bring me that beer?” It answers “Well, I was going to bring you the beer, but I decided to do something else.” Miffed, you send the guy back to the manufacturer, complaining about a lack of commitment.

Under-committed!

After retrofitting, Willie is returned, marked “Model C: The Committed Assistant.” Again, you ask Willie to bring you a beer. Again, it accedes, replying “Sure thing.” Then you ask: “What kind of beer did you buy?” It answers: “Genessee.” You say “Never mind.” One minute later, Willie trundles over with a Genessee in its gripper.

After still more tinkering, the manufacturer sends Willie back, promising no more problems with its commitments. So, being a somewhat trusting customer, you accept the rascal back into your household, but as a test, you ask it to bring you your last beer. [...] The robot gets the beer and starts towards you. As it approaches, it lifts its arm, wheels around, deliberately smashes the bottle, and trundles off.

What is Wrong with Willie?

Some time in the not-so-distant future, you are having trouble with your new household robot. You say “Willie, bring me a beer.” The robot replies “OK boss.” Twenty minutes later, you say “Willie, I don’t see the beer. Can’t you bring me that beer?” It answers “Well, I was going to bring you the beer, but I decided to do something else.” Miffed, you send the guy back to the manufacturer, complaining about a **lack of commitment**.

Under-committed!

After retrofitting, Willie is returned and marked “Model C: The Committed Assistant.” Again, you ask it to bring you a beer. Again, it accedes, replying “Sure thing.” You say “What kind of beer did you buy?” It answers: “I don’t know. I’ll never mind.” **One minute later, Willie trundles over with a beer in its gripper.**

Over-committed!

After still more tinkering, the manufacturer sends Willie back, promising no more problems with its commitments. So, being a somewhat trusting customer, you accept the rascal back into your household, but as a test, you ask it to bring you your last beer. [...] The robot gets the beer and starts towards you. As it approaches, it lifts its arm, wheels around, **deliberately smashes the bottle, and trundles off.**

What is Wrong with Willie?

Some time in the not-so-distant future, you are having trouble with your new household robot. You say “Willie, bring me a beer.” The robot replies “OK boss.” Twenty minutes later, you say “Willie, I don’t see the beer. Can’t you bring me that beer?” It answers “Well, I was going to bring you the beer, but I decided to do something else.” Miffed, you send the guy back to the manufacturer, complaining about a **lack of commitment**.

Under-committed!

After retrofitting, Willie is returned and marked “Model C: The Committed Assistant.” Again, you ask it to bring you a beer. Again, it accedes, replying “Sure thing.” You say “What kind of beer did you buy?” It answers: “I don’t know. Never mind.” One minute later, Willie **trundles over with a beer in its gripper**.

Over-committed!

After still more tinkering, the manufacturer sends you a new robot back, promising no more problems with its commitments. So, you, a somewhat trusting customer, you accept the rascal and name it “Willie.” As a test, you ask it to bring you your last beer. It gets the beer and starts towards you. As it approaches, it drops its arm, wheels around, **deliberately smashes the bottle**, and **trundles off**.

???

What is Wrong with Willie?

Some time in the not-so-distant future, you are having trouble with your new household robot. You say “Willie, bring me a beer.” The robot replies “OK boss.” Twenty minutes later, you say “Willie, I don’t see the beer. Can’t you bring me that beer?” It answers “Well, I was going to bring you the beer, but I decided to do something else.” Miffed, you take the guy back to the manufacturer, complaining about a **lack of commitment**.

Under-committed!

After retrofitting, Willie is returned and marked “Model C: The Committed Assistant.” Again, you ask it to bring you a beer. Again, it accedes, replying “Sure thing.” You say “What kind of beer did you buy?” It answers: “I don’t know. I’ll never mind.” One minute later, Willie **trundles over with a beer in its gripper**.

Over-committed!

After still more tinkering, the manufacturer sends you a new robot back, promising no more problems with its commitments. You say “Willie, I don’t see the beer. Can’t you bring me that beer?” It replies “Sure thing.” You say “What kind of beer did you buy?” It answers: “I don’t know. I’ll never mind.” One minute later, Willie **trundles over with a beer in its gripper**.

Wrongly committed?

Commitments

We may think that deliberation and planning are sufficient to achieve desired behavior, unfortunately things are more complex...

After filter function, agent makes a **commitment** to chosen option:

- ▶ Commitment: *an agreement or pledge to do something in the future*;
- ▶ \therefore it implies temporal persistence.

Questions:

- 1 how long should an intention persist?
- 2 what is the commitment on?

Degrees of Commitments

Rao and Georgeff (1991) described the following **commitment strategies**:

Blind/Fanatical commitment A blindly committed agent will continue to maintain an intention until it believes the intention has actually been achieved.

Degrees of Commitments

Rao and Georgeff (1991) described the following **commitment strategies**:

Blind/Fanatical commitment A blindly committed agent will continue to maintain an intention until it believes the intention has actually been achieved.

Single-minded commitment A single-minded agent will continue to maintain an intention until it believes that either the intention has been achieved, or else that it is no longer possible to achieve the intention.

Degrees of Commitments

Rao and Georgeff (1991) described the following **commitment strategies**:

Blind/Fanatical commitment A blindly committed agent will continue to maintain an intention until it believes the intention has actually been achieved.

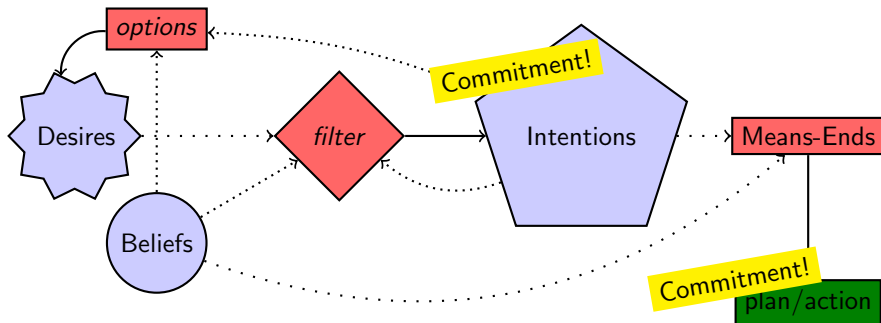
Single-minded commitment A single-minded agent will continue to maintain an intention until it believes that either the intention has been achieved, or else that it is no longer possible to achieve the intention.

Open-minded commitment An open-minded agent to remains committed to its intentions until it succeeds or the goal is dropped.

Commitments to Ends and Means

An agent has commitment both to **ends** (intentions), and **means** (plans).

- ▶ I am committed to meet/see my friend John this week (an intention);
- ▶ I am committed to drop-by John's place on Thursday afternoon (a mean).



Intention Reconsideration

When should we stop to think whether an intention should be dropped?

- ▶ Intention reconsideration is costly! (but necessary!)

A dilemma:

- ▶ an agent that does not stop to reconsider its intentions sufficiently often will continue attempting to achieve its intentions even after it is clear that they cannot be achieved, or that there is no longer any reason for achieving them;
- ▶ an agent that constantly reconsiders its intentions may spend insufficient time actually working to achieve them, and hence runs the risk of never actually achieving them.

Solution: incorporate an explicit meta-level control component, that decides whether or not to reconsider.

When is an IR Strategy Optimal?

- IR strategy is **optimal** if it would have changed intentions had he/she deliberated again (this assumes IR itself is cheap...)

Situation	IR?	Changed intentions?	Would have changed?	Optimal?
1	NO	–	NO	YES
2	NO	–	YES	NO
3	YES	NO	–	NO
4	YES	YES	–	YES

When is an IR Strategy Optimal?

- IR strategy is **optimal** if it would have changed intentions had he/she deliberated again (this assumes IR itself is cheap...)

Situation	IR?	Changed intentions?	Would have changed?	Optimal?
1	NO	—	NO	YES
2	NO	—	YES	NO
3	YES	NO	—	NO
4	YES	YES	—	YES

In situation (1), the agent did not choose to deliberate, and thus, did not change intentions. Still, if it had deliberated, it would not have changed.

∴ IR was not worth it! :-)

When is an IR Strategy Optimal?

- IR strategy is **optimal** if it would have changed intentions had he/she deliberated again (this assumes IR itself is cheap...)

Situation	IR?	Changed intentions?	Would have changed?	Optimal?
1	NO	–	NO	YES
2	NO	–	YES	NO
3	YES	NO	–	NO
4	YES	YES	–	YES

In situation (2), the agent did not choose to deliberate, but if it had done so, it would have changed intentions.

∴ IR must have been done! :-)

When is an IR Strategy Optimal?

- IR strategy is **optimal** if it would have changed intentions had he/she deliberated again (this assumes IR itself is cheap...)

Situation	IR?	Changed intentions?	Would have changed?	Optimal?
1	NO	–	NO	YES
2	NO	–	YES	NO
3	YES	NO	–	NO
4	YES	YES	–	YES

In situation (3), the agent chose to deliberate, but did not change intentions.
 ∴ IR was a waste of time! :-)

When is an IR Strategy Optimal?

- IR strategy is **optimal** if it would have changed intentions had he/she deliberated again (this assumes IR itself is cheap...)

Situation	IR?	Changed intentions?	Would have changed?	Optimal?
1	NO	–	NO	YES
2	NO	–	YES	NO
3	YES	NO	–	NO
4	YES	YES	–	YES

In situation (4), the agent chose to deliberate, and did change intentions.
 ∴ IR was worth it! :-)

When is an IR Strategy Optimal?

Kinny and Georgeff experimentally investigated effectiveness of intention reconsideration strategies:

Bold agents never pause to reconsider intentions.

Cautious agents stop to reconsider after every action.

Dynamism in the environment is represented by the rate of world γ change:

When is an IR Strategy Optimal?

Kinny and Georgeff experimentally investigated effectiveness of intention reconsideration strategies:

Bold agents never pause to reconsider intentions.

Cautious agents stop to reconsider after every action.

Dynamism in the environment is represented by the rate of world γ change:

- ▶ If γ is low (i.e., the environment does not change quickly), then **bold agents do well compared to cautious ones**.
 - ▶ cautious ones waste time reconsidering their commitments while bold agents are busy working towards—and achieving—their intentions.

When is an IR Strategy Optimal?

Kinny and Georgeff experimentally investigated effectiveness of intention reconsideration strategies:

Bold agents never pause to reconsider intentions.

Cautious agents stop to reconsider after every action.

Dynamism in the environment is represented by the rate of world γ change:

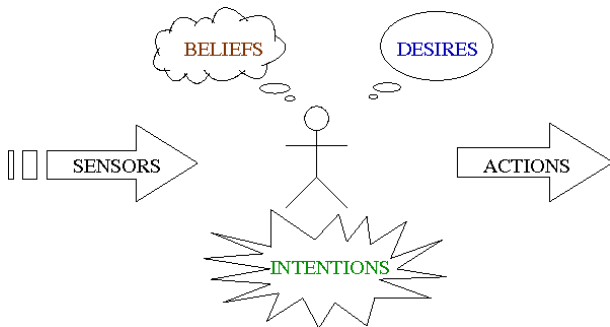
- ▶ If γ is low (i.e., the environment does not change quickly), then **bold agents do well compared to cautious ones**.
 - ▶ cautious ones waste time reconsidering their commitments while bold agents are busy working towards—and achieving—their intentions.
- ▶ If γ is high (i.e., the environment changes frequently), then **cautious agents tend to outperform bold agents**.
 - ▶ cautious agents are able to recognize when intentions are doomed, and also to take advantage new opportunities when they arise.

Next Topic:

Agent Architectures

Belief-Desire-Intention (BDI) Model of Agency

A model of agency with roots in practical reasoning and intentional systems
 \therefore a set of concepts for thinking about and building agents.



Behavior arises due to the agent committing to some of its desires, and selecting actions that achieve its intentions given its beliefs.

Abstract Rational Architecture

Rao and Georgeff (1992) proposed an abstract interpreter for agents:

Algorithm 1 BDI Rational Agent Interpreter

```

1: initialize_state()
2: while  $\neg \text{quit}$  do
3:   options  $\Leftarrow$  option_generator(event_queue,  $\mathcal{B}$ ,  $\mathcal{G}$ ,  $\mathcal{I}$ );
4:   selected_options  $\Leftarrow$  deliberate(options,  $\mathcal{B}$ ,  $\mathcal{G}$ ,  $\mathcal{I}$ );
5:   update_intentions(selected_options,  $\mathcal{I}$ );
6:   execute( $\mathcal{I}$ );
7:   drop_successful_attitudes( $\mathcal{B}$ ,  $\mathcal{G}$ ,  $\mathcal{I}$ );
8:   drop_impossible_attitudes( $\mathcal{B}$ ,  $\mathcal{G}$ ,  $\mathcal{I}$ );
9: end while
  
```

Key to this architecture is the notion of **events**:

- ▶ the inputs of the system;
- ▶ both externals (from the environment) and internal (within the system);
- ▶ will play a major role in BDI agent programming languages!.

IRMA: Intelligent Resource-bounded Machine Architecture

IRMA has four key symbolic data structures:

- 1 a **plan library**: normal operations;
- 2 **beliefs**: information available to the agent — may be represented symbolically, but may be as simple as PASCAL variables;
- 3 **desires**: those things the agent would like to make true — think of desires as tasks that the agent has been allocated; in humans, not necessarily logically consistent, but our agents will be! (goals);
- 4 **intentions**: desires that the agent has chosen and committed to.

IRMA: Other Components

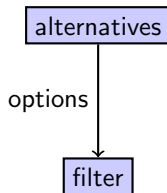
Additionally, the architecture has:

- ▶ a **reasoner** for reasoning about the world; an inference engine;
- ▶ **means-ends analyzer** determines which plans might be used to achieve intentions;
- ▶ an **opportunity analyzer** monitors the environment, and as a result of changes, generates new options;
- ▶ a **filtering process** determines which options are compatible with current intentions; and
- ▶ a **deliberation process** responsible for deciding upon the 'best' intentions to adopt.

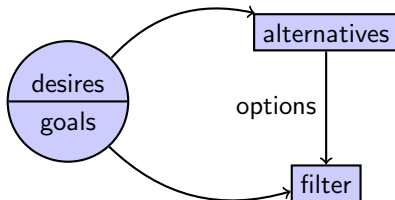
IRMA: The Architecture

alternatives

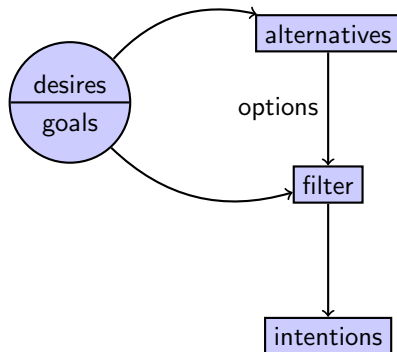
IRMA: The Architecture



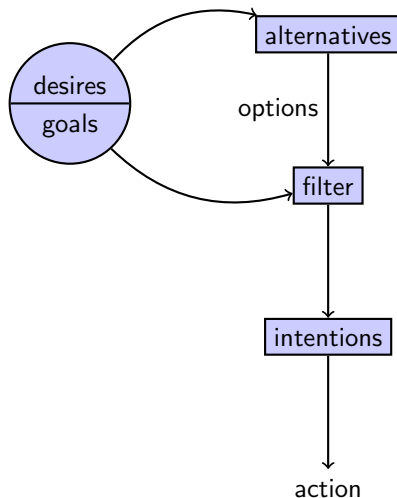
IRMA: The Architecture



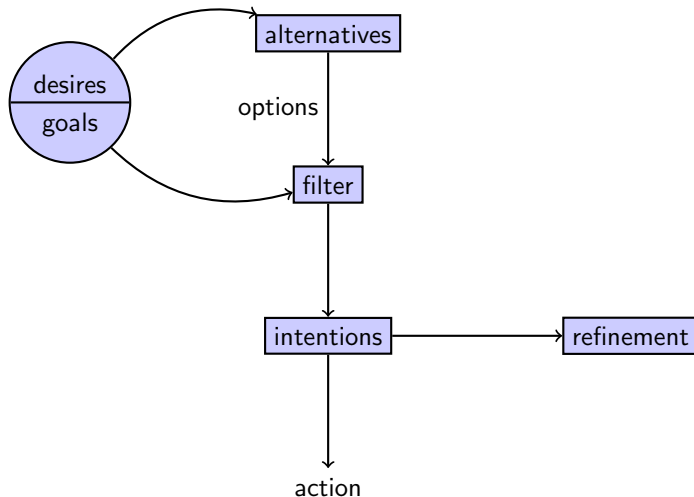
IRMA: The Architecture



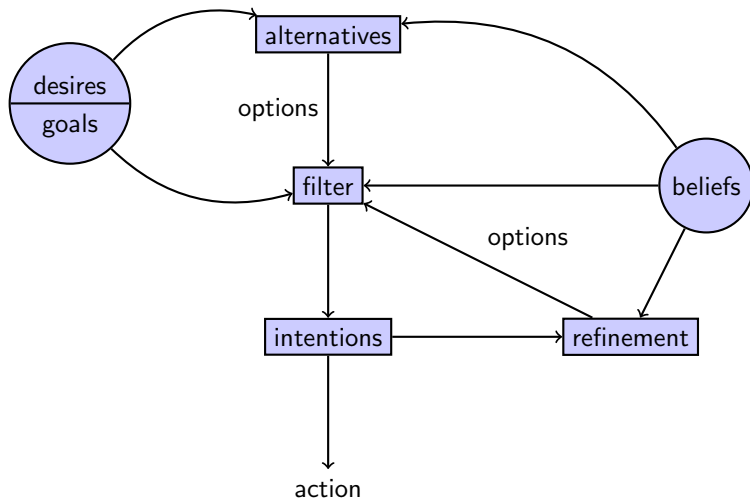
IRMA: The Architecture



IRMA: The Architecture



IRMA: The Architecture



Procedural Reasoning System: PRS

PRS is an actual implemented agent system!
(Designed primarily in the Australian AI Institute & SRI International)

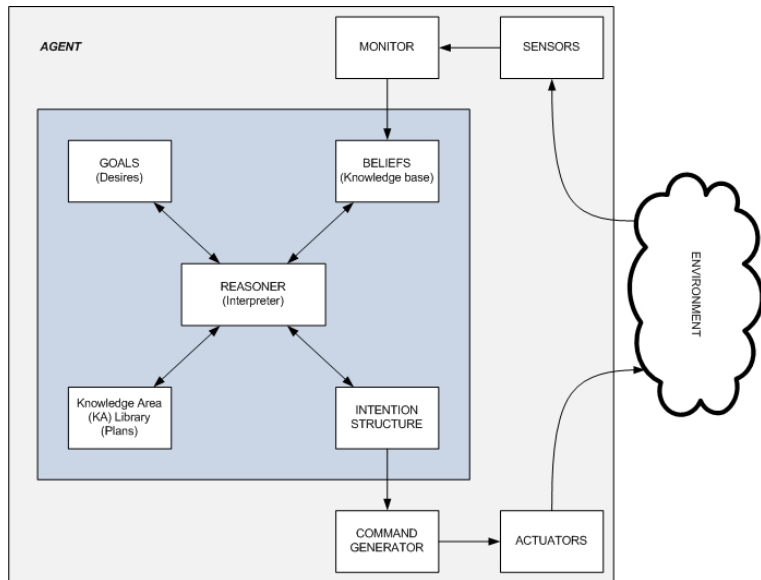
In the PRS, each agent is equipped with a **plan library**, representing that **agent's procedural knowledge**: knowledge about the mechanisms that can be used by the agent in order to realize its intentions.

The **options** available to an agent are **directly determined by the plans an agent has**: an agent with no plans has no options.

In addition, PRS agents have explicit representations of beliefs, desires, and intentions, as in IRMA.

Many variations of PRS: dMars, SPARKS, PRS-CL, JAM, **JACK**, etc.

Procedural Reasoning System: PRS



Plans in PRS: Stacking a Block

```

Plan: {
NAME: "Stack blocks that are already clear"
GOAL:
    ACHIEVE ON $OBJ1 $OBJ2;
CONTEXT:
BODY:
    EXECUTE print "Making sure " $OBJ1 " is clear\n";
    ACHIEVE CLEAR $OBJ1;
    EXECUTE print "Making sure " $OBJ2 " is clear.\n";
    ACHIEVE CLEAR $OBJ2;
    EXECUTE print "Moving " $OBJ1 " on top of " $OBJ2 ".\n";
    PERFORM move $OBJ1 $OBJ2;
UTILITY: 10;
FAILURE:
    EXECUTE print "\n\nStack blocks failed!\n\n";
}

```

Plans in PRS: Clearing a Block

```

Plan: {
NAME: "Clear a block"
GOAL:
    ACHIEVE CLEAR $OBJ;
CONTEXT:
    FACT ON $OBJ2 $OBJ;
BODY:
    EXECUTE print "Clearing " $OBJ2 " from on top of " $OBJ "\n";
    EXECUTE print "Moving " $OBJ2 " to table.\n";
    ACHIEVE ON $OBJ2 "Table";
EFFECTS:
    EXECUTE print "CLEAR: Retracting ON " $OBJ2 " " $OBJ "\n";
    RETRACT ON $OBJ1 $OBJ;
FAILURE:
    EXECUTE print "\n\nClearing block " $OBJ " failed!\n\n";
}

```

Next Topic:

Agent Theory

Agent Theories

Formal specifications of agent properties — what kind of mental states they have and how they are related to each other and to action; should support reasoning about agents.

We now consider the **semantics** of BDI architectures: to what extent does a BDI agent satisfy a theory of agency?

Two major seminal works:

- 1 Cohen & Levesque: “Intentions = Choice + Commitment”
- 2 Rao & Georgeff’s BDI logics: non-classical logics with modal connectives for representing beliefs, desires, and intentions.

Cohen-Levesque: Intention = Choice+Commitment

A logic formalism used to develop a **theory of intention** (as in 'I intend to ...') based on 2 primitive mental attitudes:

(BEL x p) p follows from the agent x 's goals.

(GOAL x p) p follows from the agent x 's goals.

Cohen-Levesque: Intention = Choice+Commitment

A logic formalism used to develop a **theory of intention** (as in 'I intend to ...') based on 2 primitive mental attitudes:

$(BEL\ x\ p)$ p follows from the agent x 's goals.

$(GOAL\ x\ p)$ p follows from the agent x 's goals.

The logic has a **possible world semantics**:

$(BEL\ x\ p)$ In all the worlds the agent x believes possible, p holds.

$(BEL\ x\ p)$ Agent x is choosing a world in which p is true.

The world is defined as a discrete sequence of events, temporally extended into past and future

Then, further attitudes, such as intentions, are defined in terms of these.

Persistent Goals

The first major derived attitude is the persistent goal.

An agent has a **persistent goal** of p (P-GOAL \times p):

- 1 It has a goal that p eventually becomes true.
- 2 Believes that p is not currently true.
- 3 Before it drops the goal p , one of the following conditions must hold:
 - ▶ the agent believes p has been satisfied; or
 - ▶ the agent believes p will never be satisfied.

$$\begin{aligned}
 (\text{P-GOAL } \times p) &\stackrel{\text{def}}{=} \\
 &(\text{GOAL } \times (\text{LATER } p)) \wedge \\
 &(\text{BEL } \times \neg p) \wedge \\
 &(\text{BEFORE } [(\text{BEL } \times p) \vee (\text{BEL } \times \Box \neg p)] \neg (\text{GOAL } \times (\text{LATER } p)))
 \end{aligned}$$

Intentions as Persistent Goals

Definition

An agent **intends to do action a** **iff** it has a **persistent goal** to have brought about a state wherein **it believed it was about to do a** , and then **did a** .

$$(\text{INTEND } x \ a) \stackrel{\text{def}}{=} (\text{P-GOAL } x \ [(\text{DONE } x \ (\text{BEL } x \ (\text{HAPPENS } a)))?; a])$$

(HAPPENS a) Action expression a will happen next.

(DONE $x \ p$) Agent x has just performed action expression a .

Consistency with Bratman's Requirements

Intentions provide a “screen of admissibility” for adopting other intentions:

$$\models \forall x. (\text{INTEND } x \ b) \wedge \Box((\text{BEL } x \ [(\text{DONE } x \ a) \supset \Box \neg(\text{DONE } x \ b)])) \supset \neg(\text{INTEND } x \ a; b)$$

Consistency with Bratman's Requirements

Intentions provide a “screen of admissibility” for adopting other intentions:

$$\models \forall x. (\text{INTEND } x \ b) \wedge \Box((\text{BEL } x \ [(\text{DONE } x \ a) \supset \Box \neg(\text{DONE } x \ b)]) \supset \neg(\text{INTEND } x \ a; b))$$

Other properties proved that match Bratman's framework:

- 1 Intentions pose problems for agents, who need to determine ways of achieving them.
- 2 Agents track the success of their intentions, and are inclined to try again if their attempts fail.
- 3 Agents believe their intentions are possible.
- 4 Agents do not believe they will not bring about their intentions.
- 5 Under certain circumstances, agents believe they will bring about their intentions.
- 6 Agents need not intend all the expected side effects of their intentions.

Rao and Georgeff's BDI Logic

In order to give a semantics to BDI architectures, Rao & Georgeff have developed BDI logics: non-classical logics with **modal connectives** for representing beliefs, desires, and intentions:

$(\text{Bel } i \ \phi)$ i believes ϕ

$(\text{Des } i \ \phi)$ i desires ϕ

$(\text{int } i \ \phi)$ i intends π

The 'basic BDI logic' is a quantified extension of the expressive branching time logic CTL*.

Properties of Rao and Georgeff's BDI Logic

- ▶ **Belief goal compatibility:** $(\text{Des } \alpha) \Rightarrow (\text{Bel } \alpha)$
States that if the agent has a goal to optionally achieve something, this thing must be an option.

This axiom is operationalized in the function *option_generator*
- ▶ **Goal-intention compatibility:** $(\text{Int } \alpha) \Rightarrow (\text{Des } \alpha)$
States that having an intention to optionally achieve something implies having it as a goal (i.e., there are no intentions that are not goals).

Operationalized in the *deliberate* function.
- ▶ **Volitional commitment:** $(\text{Int } \text{does}(a)) \Rightarrow \text{does}(a)$
If you intend to perform some action *a* next, then you do *a* next.

Operationalized in the *execute* function.

Properties of Rao and Georgeff's BDI Logic (cont.)

- Awareness of goals & intentions:

$$(\text{Des } \phi) \Rightarrow (\text{Bel } (\text{Des } \phi))$$

$$(\text{Int } \phi) \Rightarrow (\text{Bel } (\text{Int } \phi))$$

Requires that new intentions and goals be posted as events.

- No unconscious actions: $\text{done}(a) \Rightarrow (\text{Bel } \text{done}(a))$

If an agent does some action, then it is aware that it has done the action.

Operationalized in the *execute* function.

A stronger requirement would be for the success or failure of the action to be posted.

- No infinite deferral: $(\text{Int } \phi) \Rightarrow A\Diamond(\neg(\text{Int } \phi))$

An agent will eventually either act for an intention, or else drop it.

Next Topic:

Conclusions

Review

In this lecture we have seen:

- 1 Philosophical background on (agent) rational action/behavior
 - ▶ Daniel Dennet's intentional system theory;
 - ▶ Michael Bratman's theory of practical reasoning.
- 2 Three major issues/topics of practical reasoning:
 - ▶ the deliberation process;
 - ▶ intentions and their role in practical reasoning;
 - ▶ commitment (on ends & means).
- 3 Three major abstract agent architectures:
 - ▶ Abstract BDI interpreter (Rao & Georgeff);
 - ▶ IRMA (Bratman, Israel & Pollack);
 - ▶ PRS (Georgeff & Lansky).
- 4 Two seminal works on agent theory:
 - ▶ Cohen & Levesque's theory of intentions;
 - ▶ Rao & Georgeff's BDI logics.

Next Lecture

BDI Agent-Oriented Programming

For Further Reading: BDI Theory



Michael Bratman.

Intentions, Plans, and Practical Reason.

Harvard University Press, 1987.



Daniel Dennett.

The Intentional Stance.

MIT Press, 1987.



Philip R. Cohen and Hector J. Levesque.

Intention is choice with commitment.

Artificial Intelligence Journal, 42:213–261, 1990.





A.S. Rao and M.P. Georgeff.


Modeling rational agents within a BDI-architecture.


In *Proceedings of KR-91*, pages 473–484, 1991.


For Further Reading: BDI Architectures

 Bratman, M. E., Israel, D. J., and Pollack, M. E.
Plans and resource-bounded practical reasoning.
Computational Intelligence, 4:349-355, 1988.

 Anand S. Rao and Michael P. Georgeff.
An abstract architecture for rational agents.
In *Proceedings of KR-92*, pages 438–449, 1992.

 Georgeff and Lansky.
Reactive reasoning and planning.
In *Proceedings of AAAI-87*, pages 677-682, 1987.

 Georgeff, M. P. and Ingrand, F. F.
Decision-making in an embedded reasoning system.
In *Proceedings of IJCAI-89*, 972-978, 1989.

 F. Ingrand, M. Georgeff, and A Rao.
An architecture for real-time reasoning and system control.
IEEE Expert, 7(6), 1992.