Frontiers of Autonomous Systems

Digital Culture

Andrea Omicini
andrea.omicini@unibo.it

Master in Business Management / Bologna Business School

Academic Year 2017/2018

Next in Line...

Before Autonomous Systems: Machines I

Constructing for understanding

- building machines with
  - initiative
  - autonomy
  - knowledge

  for *understanding ourselves*, and the *world* where we live
- "playing God" to understand the world

Before Autonomous Systems: Machines II

Relieving humans from fatigue

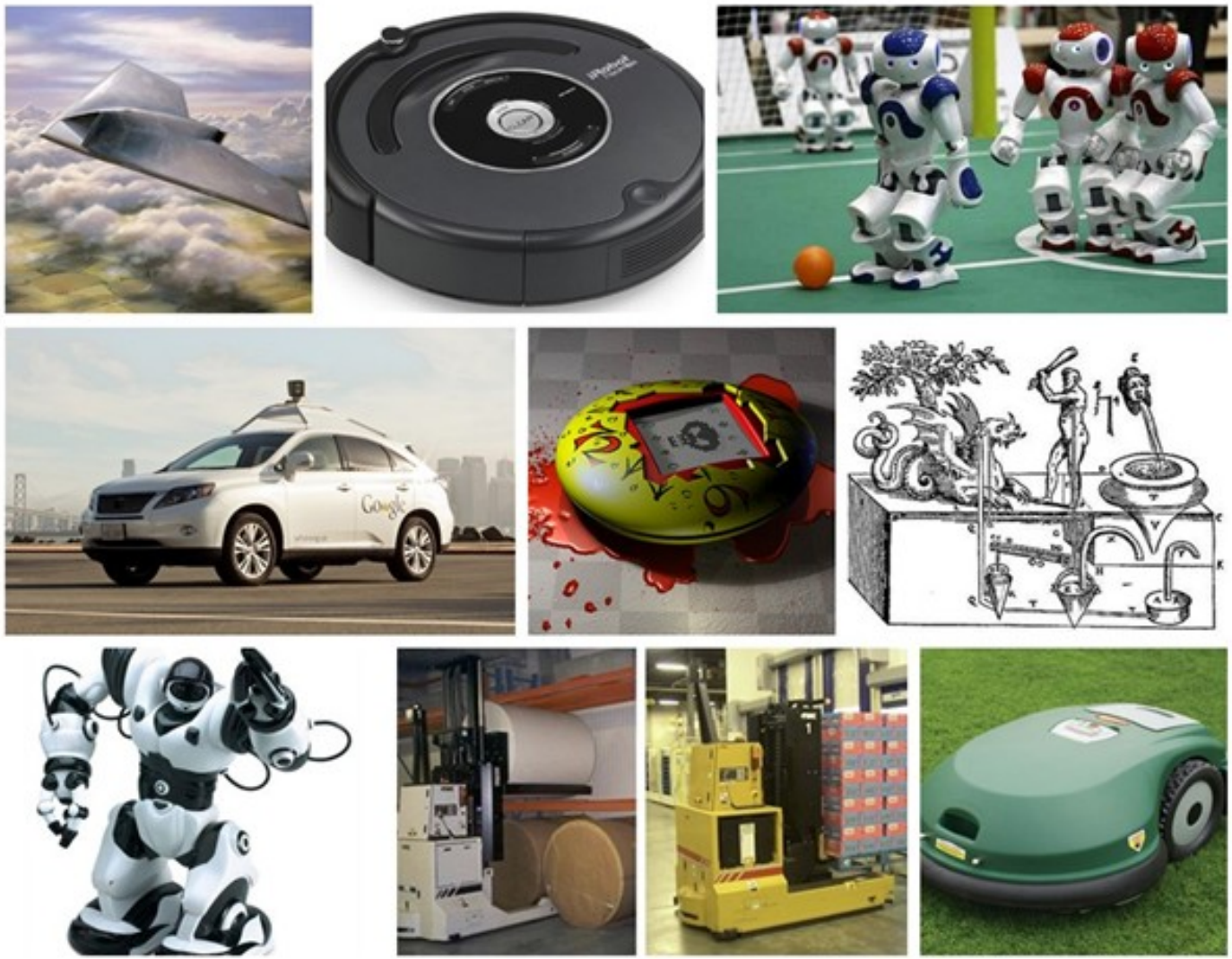
- goal: substituting human work in
  - quality
  - quantity
  - cost
- more, better, cheaper work done
  - new activities become feasible
- which work?
  - first, physical
  - then, repetitive, enduring
  - subsequently, intellectual, too
  - finally, simply more complex for any reason—or, all reasons together

Before Autonomous Systems: Machines III

Some steps beyond

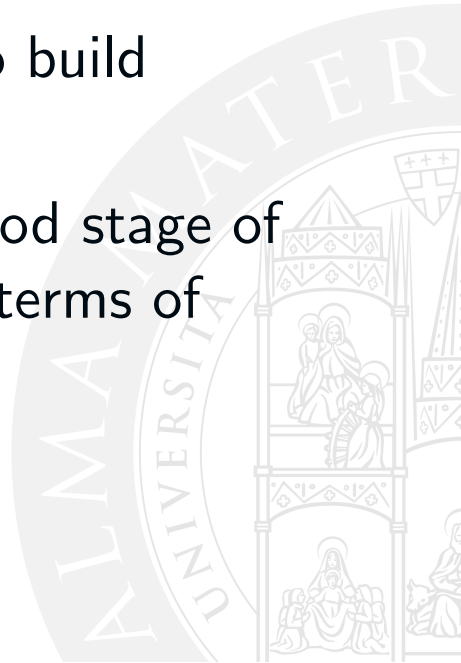
- delegating human functions to machines
  - within already existing social structures, organisations, and processes
- creating new functions
  - then, making new social structures, organisations, and processes possible
  - example: steam engines on wheels
- essentially, *changing the world we live in*

Autonomous Systems Around



Social Pressure

- activities that might be *delegated* to *artificial systems* grow in *number* and *complexity*
- people & organisations are already experiencing / conceiving "somehow autonomous" systems, and ask for more
- engineers are not yet trained on *general approaches* to build *autonomous systems*
- however, the theory of autonomous systems is at a good stage of development, and technologies are rapidly growing in terms of *availability* and *reliability*



Next in Line...

Oxford Dictionary of English (2nd Edition revised 2005)

Etimology

Early 17th cent.: from Greek *autonomia*, from *autonomos* 'having its own laws', from *autos* 'self' + *nomos* 'law'.

Dictionary

autonomy

- the right or condition of self-government
- a self-governing country or region
- freedom from external control or influence; independence.
- (in Kantian moral philosophy) the capacity of an agent to act in accordance with objective morality rather than under the influence of desires

Oxford Thesaurus of English (2nd Edition revised 2008)

Thesaurus

autonomy

- self-government, independence, self-rule, home rule, sovereignty, self-determination, freedom, autarchy;
- self-sufficiency, individualism.

Merriam-Webster I

Dictionary

autonomy

1. the quality or state of being self-governing; especially: the right of self-government
2. self-directing freedom and especially moral independence
3. a self-governing state

synonyms accord, free will, choice, self-determination, volition, will

antonyms dependence (also dependance), heteronomy, subjection, unfreedom

Merriam-Webster II

Thesaurus

autonomy

1. the act or power of making one's own choices or decisions: accord, free will, choice, self-determination, volition, will
2. the state of being free from the control or power of another: freedom, independence, independency, liberty, self-determination, self-governance, self-government, sovereignty

Autonomy and Biological Systems

Living systems. . .

. . . are the first systems provided of (some level of) autonomy that we have knowledge of

- they *work* as autonomous systems
- they *evolved* to become autonomous

The study of autonomy in biological systems

- the hierarchy of living systems provide examples of many *different* levels of autonomy—from lower to higher levels of autonomy
- the evolutionary view over biological systems potentially sheds light over the *role* of autonomy in living systems

→ the study of living system may help us understanding the many different sorts of autonomy, and their role in (artificial) systems in general

Biology & Evolutionary Biology I

Biology

. . . is the study of *living organisms*, or – perhaps more generally – of living systems: their structure, function, growth, origin, evolution, and distribution

Evolutionary biology

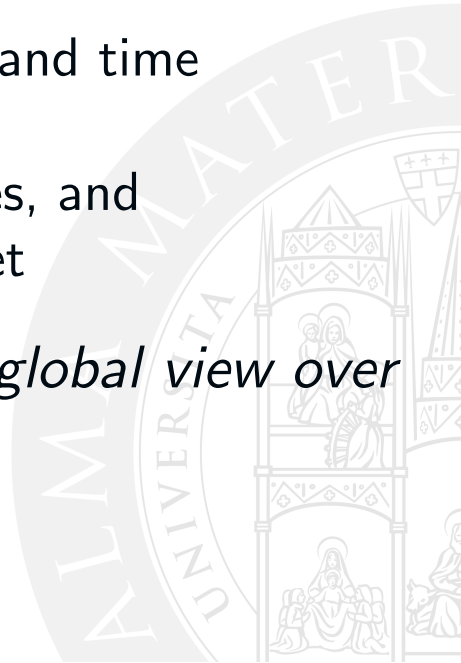
. . . studies how evolutionary processes produced diversity of life on Earth; that is, *how biological systems evolved over the ages* [Gould, 2002]. As a result, the view of evolutionary biology over biological systems includes

- not just how they are made, and how they do *work*
- but also – and mainly – how they do evolved towards their current form—the one we can presently observe

Biology & Evolutionary Biology II

- *evolutionary biology* is indeed a (nowadays constitutive) part of biology
- however, the way evolutionary biologists look over biological systems tends to be less analytical – yet wider – than the one by biologists
- including
  - an overall view of the life on Earth overcoming space and time boundaries
  - the role and mutual influences of all organisms, species, and ecosystems that inhabit (or, have inhabited) our planet

Roughly speaking, evolutionary biology provides us with a *global view over biological systems* at every conceivable level



Evolution & Diversity I

Not just genetic expression

- genes (and their products) are at the core of ours evolutionary model(s)
- however, they alone cannot explain the full range of variation and diversity of living systems
- according to molecular biology, even distantly-related organisms use similar processes for cellular function, development, and metabolism [Rosslenbroich, 2014]
  - bacteria and humans share part of the same metabolism
  - microscopic fungi and humans exhibit a very similar basic cell organisation and functions
  - and so on and so forth

  $\rightarrow$ most processes are *conserved* during evolution

Evolution & Diversity II

A key question in evolutionary biology

So

- how can we explain the *huge diversity* of life despite its deep and pervasively similar molecular architecture? [Rosslenbroich, 2014]
- how do organisms, species, and life overall *progress* along their evolutionary path?
- and, *do* they actually progress?

Progress in Evolution I

A controversial issue

According to [Rosslenbroich, 2014], the acceptation of the word *progress* conveys a number of diverse meanings—from Darwin to contemporary biologists

1. *change* leads to new (*higher*) organisms
2. which are somehow *improved*
3. progression is *linear*
4. evolution is an *intrinsic force* driving such a process
5. progression has a *goal* / end / culmination / perfection

Nowadays, we just need the first acceptation: however, we need to understand what is *higher*, what is *change*, and *what* is actually progressing along with evolution

Progress in Evolution II

Progress of what?

- increased *potential* for survival?
- increased *efficiency* of some form, like, energy consumption?
- increased *amount of information* in genes?
- increased *differentiation*?
- increased *complexity*?
- increased *emancipation from the environment*?

Nowadays, *complexity* is often used instead of *progress*: this, however, does not offer any explanation—no clear large-scale *patterns* in evolution, here

Autopoiesis I

Autopoiesis is the ability of a complex system of maintaining its own overall coherence, in terms of structure and organisation, through the mutual *interactions* of its components

Autopoiesis II

Living systems as autopoietic systems

[Maturana and Varela, 1980, Varela et al., 1974]

- living systems as autopoietic units capable of sustaining themselves based on an inner network of reactions that generate and regenerate all the system components
- all pertinent processes required have an inner efficient cause
- *structures* – based on a flow of molecules and energy – produce the components that, in turn, continue to maintain the organised bounded structure that gives rise to these components
- self-reference and self-maintenance are core notions here
- coherent and ordered global system behaviour of the system constrains / governs the behaviour of the individual components, while the component behaviour sustains the global order (*circular causality*)

Autopoiesis & Autonomy [Thompson, 2010] I

According to [Maturana and Varela, 1980], autonomous systems

- *acquire* the property of specifying their own rules of behaviour
- do *not* work as transducers or functions for converting input instructions into output products
- are the sources of their own activity, which specify their own domains of interaction

Autopoiesis & Autonomy [Thompson, 2010] II

"In fact, the notion of autopoiesis can be described as a characterisation of the mechanisms which endow living systems with the property of being autonomous; autopoiesis is an explication of the autonomy of the living" *[Maturana and Varela, 1980]*

Autonomy of a Cell [Thompson, 2010, Rosslenbroich, 2014]

An example of biological autonomy: the cell

- the cell stands out of a molecular soup by *actively creating the boundaries* that
  - set the cell apart from what it is not the cell
  - and simultaneously regulate cell interaction with the environment
- metabolic processes within the cell construct those boundaries, but the metabolic processes themselves are made possible by those boundaries
- thus, the cell emerges as a figure *standing out of a chemical background*
- should this process of self-production be interrupted, the cellular components no longer form a unit, gradually diffusing back into a molecular soup—*death*

Autopoiesis & Autonomy [Thompson, 2010] I

Boundary

- boundary is a central element of autonomy
- it is a constitutive element of the identity of a system
- in a cell, the *membrane* works as a boundary both containing processes/components and regulating the interaction with the environment
- *boundaries* – strict physical ones, not necessarily material – are *essential for an autonomous system*

Autopoiesis & Autonomy [Thompson, 2010] II

Autonomous systems are *closed* [Rosslenbroich, 2014]

- *organisationally* closed in the sense that their organisation is characterised by their internal network processes
- which recursively *depend on each other*, thus constitute the system *as a unit*

Heteronymous vs. Autonomous Systems

[Thompson, 2010, Rosslenbroich, 2014] I

Heteronymous systems

A heteronymous system is one whose organisation is defined by *input-output* information flow and *external* mechanisms of *control*

- traditional computational systems and many network views, for example, are heteronymous
  - they have an input layer and an output layer
  - the inputs are initially assigned by the observer outside the system
  - output performance is evaluated in relation to an externally imposed task

e.g., a Turing Machine typically represents computation by a heteronomous system

Heteronymous vs. Autonomous Systems

[Thompson, 2010, Rosslenbroich, 2014] II

Autonomous systems

An autonomous system is defined by its *endogenous*, self-organising, and self-controlling *dynamics*, and determines the domain in which it operates

- it has *input* and *output*—which, alone, do not determine the system
- it is the internal *self-production* process that *controls* and *regulates* the system's *interaction* with the outside *environment*

Autonomy and Environment I

Autonomy is not autarky

- living systems are *not independent* of their *environment*
- the *interchange* occurs though the *physical boundary*

Autonomy and Environment II

Plants vs. animals

- plants exhibit a predominantly *open* relation to their environment
- instead, animals have a more *closed* form of organisation
  - the exchange surfaces for metabolism are turned to the inside
  - special internal organs and internal cavities appear
  - exchange surfaces on the outside are reduced
- the loss of a direct environmental relation corresponds to a gain in degrees of freedom
- *stimulus-response* relationships in animals tend to be less tightly connected
- in animals, signals can internally be enforced, compared to other signals, and memorised
- thus, not a rigid, but rather a flexible relation between organism and environment emerges when "moving up" from plants to animals

Robustness I

Stability in front of change

- many structures and functions – as well as proteins and genes – have certain stability in the face of *environmental variations* and *genetic changes*
- they are resistant, robust, to *perturbations*, producing relatively invariant outputs [Kitano, 2002]

Robustness II

Robustness in living systems

- is understood as a property that allows a system to *maintain its functions* against internal and external perturbations and uncertainties
- encompasses a broad range of traits: from macroscopic, visible traits, to molecular traits, such as the expression level of a gene, or, the three-dimensional conformation of a protein
- is widely recognised as an *inherent property* of all biological systems

Robustness III

Autonomy & robustness

- autonomy and robustness somehow overlap, but they are not the same
- robustness may be seen as a *pre-requisite* for autonomy: for instance, self-maintenance requires robustness
- or, robustness is a *part of* autonomy, as it maintains the identity, structure, and organisation of a living system against well-separated surroundings

Robustness IV

Principles & strategies for robustness [Kitano, 2002]

redundancy of components to protect *against failure* of a specific component by providing for alternative ways to carry out the function the component performs

feedback circuits to *monitor* a system function so as to *regulate* it

modularity as the *encapsulation* of functions, for robustness and evolvability

layering in hierarchical systems to enhance *control* and robustness

Homeostasis

Homeostasis is the ability of a system to regulate its internal conditions to keep some or several functions stable

e.g., properties such as temperature or blood composition in animals

- separating internal and external environments
- where *internal environment* is kept relatively stable with respect to *external perturbations*

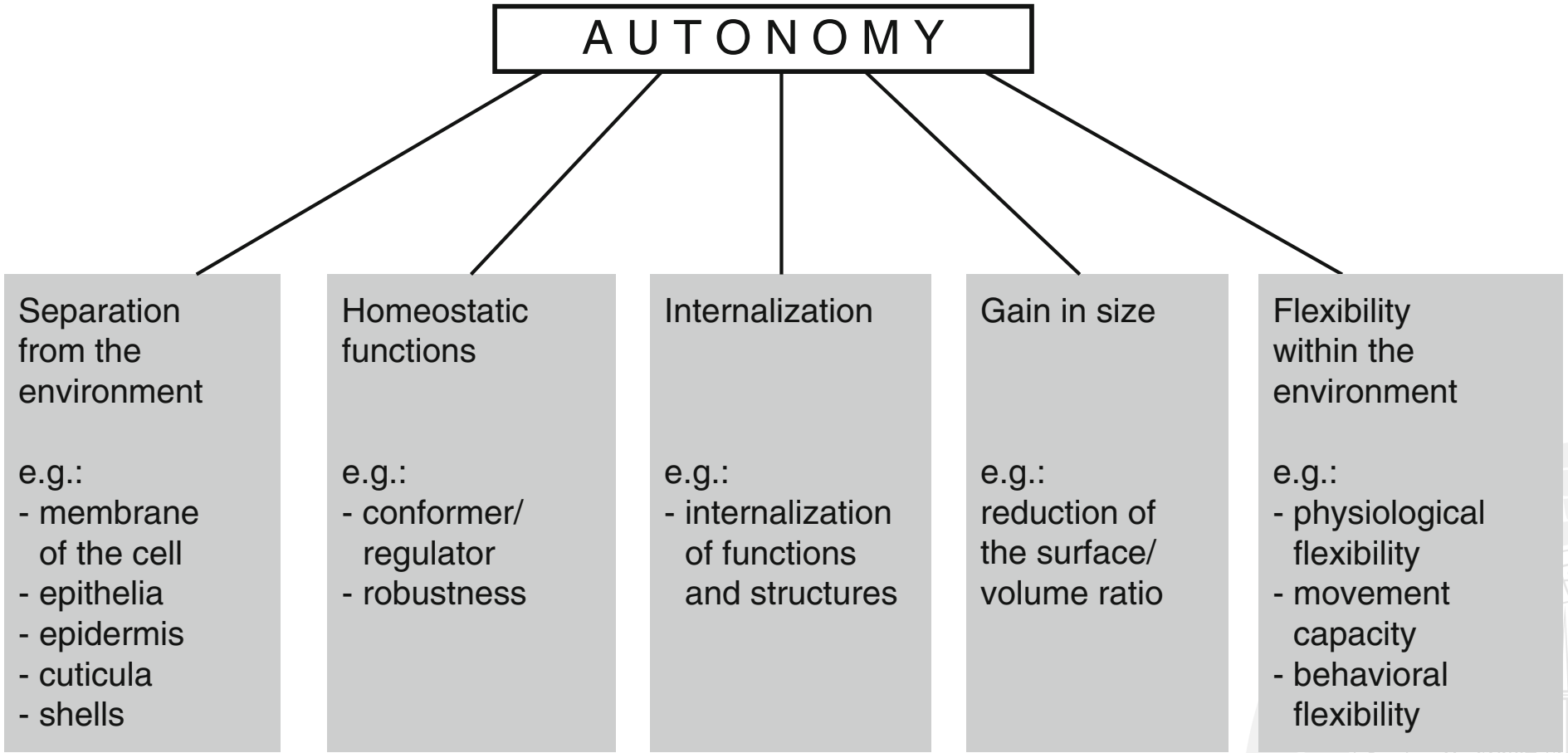
Time Autonomy

- living entities establish *their own* cycles in *time*
- e.g., metabolism, rest-activity cycle, development, reproduction
- involving all biochemical, cellular, and organic processes
- from reaction rates, frequencies are endogenous, lead to *autonomous cycles*, which only later synchronise with external cycles

Evolutions through Increasing Autonomy

Increasing autonomy [Rosslenbroich, 2014]

- which features are able to contribute to *changes* in autonomy of an individual organism?
- how can autonomy be *defined* accordingly in a more formal way?



Resources to change autonomy—from [Rosslenbroich, 2014]

A Definition for Autonomy in Living Systems

General autonomy [Rosslenbroich, 2014]

Living systems are *autonomous* in the sense that they *maintain themselves in form and function* within time and achieve a *self-determined flexibility*

1. they generate, maintain, and regulate an *inner network* of interdependent, energy-consuming processes, which in turn generate and maintain the system
2. they establish a *boundary* and actively regulate their *interaction* and exchange with the *environment*
3. they specify their *own rules of behaviour* and react to external stimuli in a *self-determined* way, according to their internal disposition and condition
4. they establish an *interdependence* between the system and its parts within the organism, which includes a *differentiation* in subsystems
5. they establish a *time autonomy*
6. they maintain a phenotypic stability (*robustness*) in the face of diverse perturbations arising from environmental changes, internal variability, and genetic variations

Autonomy in Evolution

- autonomy is an essential trait of living systems
- many authors observe that evolution progresses through biological systems with *diverse degrees of autonomy*
- many additional functions resulting from evolutionary processes tends to improve autonomy
- how can autonomy help in understanding evolution?
- is it a *pattern of evolution*?

Increasing Autonomy [Rosslenbroich, 2014] I

- increasing autonomy is defined as an *evolutionary shift* in the *system-environment relationship*, such that

  interactive autonomy the direct influences of the environment on the respective individual systems are gradually reduced

  constitutive autonomy stability and flexibility of self-referential, intrinsic functions within the systems are generated and enhanced

- with respect to the environment
  - *autonomy* of living systems is *relative*
  - while retaining numerous interconnections with and dependencies on the external environment
  - thus, organisms can undergo relative *emancipation from environmental fluctuations*, gaining self-determination and flexibility of behaviour

Increasing Autonomy [Rosslenbroich, 2014] II

- a set of resources can be involved to change autonomous capacities
  1. changes in *spatial separation* from the environment
  2. changes in *homeostatic* capacities and robustness
  3. *internalisation* of structures or functions
  4. increase in body *size*
  5. changes in the *flexibility* within the environment, including behavioural flexibility

Overall: Autonomy as a Driver for Evolution

- autonomy is an essential trait of living systems
- biological systems evolve towards increasing degrees of autonomy
- autonomy is an evolutionary pattern

Internet Encyclopedia of Philosophy I

Many acceptations of autonomy

general	an individual's capacity for self-determination or self-governance
folk	inchoate desire for freedom in some area of one's life
personal	the capacity to decide for oneself and pursue a course of action in one's life
moral	the capacity to deliberate and to give oneself the moral law, rather than merely heeding the injunctions of others
political	the property of having one's decisions respected, honored, and heeded within a political context

Internet Encyclopedia of Philosophy II

Individual autonomy

- after Kant, autonomy is an essential trait of the individual, and strictly related with its morality, represented by some high-level ethical principles
- then, with the relation between its inner self and its individual actions
- that is, mind and behaviour

Internet Encyclopedia of Philosophy III

Independence from oneself

- a more demanding notion of autonomy requires not only self-determination, but also independence from oneself
- this conception is connected with notions of freedom and choice, and (maybe) non-determinism
- and requires the ability of reasoning on (and possibly changing) not just one own course of actions, but one own goals

Unmanned Systems Integrated Roadmap FY 2011-2036 I

Automatic vs. autonomous

automatic systems are fully pre-programmed and act repeatedly and independently of external influence or control. An automatic system can be described as self-steering or self-regulating and is able to follow an externally given path while compensating for small deviations caused by external disturbances. However, the automatic system is not able to define the path according to some given goal or to choose the goal dictating its path.

autonomous systems are self-directed toward a *goal* in that they do not require outside control, but rather are governed by laws and strategies that direct their behavior. Initially, these control algorithms are created and tested by teams of human operators and software developers. However, if *machine learning* is utilized, autonomous systems can develop modified strategies for themselves by which they select their behavior. An autonomous system is self-directed by choosing the behavior it follows to reach a human-directed goal.

Unmanned Systems Integrated Roadmap FY 2011-2036 II

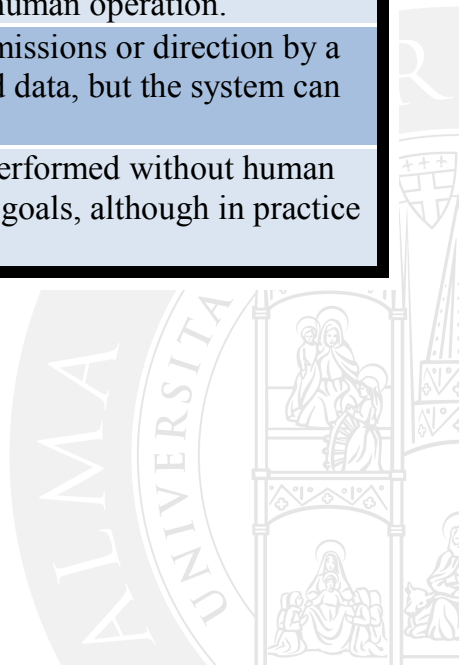
Four levels of autonomy for unmanned systems [Edwards, 2013]

Various levels of autonomy in any system guide how much and how often humans need to interact or intervene with the autonomous system:

- human operated
- human delegated
- human supervised
- fully autonomous

Unmanned Systems Integrated Roadmap FY 2011-2036 III

Level	Name	Description
1	Human Operated	A human operator makes all decisions. The system has no autonomous control of its environment although it may have information-only responses to sensed data.
2	Human Delegated	The vehicle can perform many functions independently of human control when delegated to do so. This level encompasses automatic controls, engine controls, and other low-level automation that must be activated or deactivated by human input and must act in mutual exclusion of human operation.
3	Human Supervised	The system can perform a wide variety of activities when given top-level permissions or direction by a human. Both the human and the system can initiate behaviors based on sensed data, but the system can do so only if within the scope of its currently directed tasks.
4	Fully Autonomous	The system receives goals from humans and translates them into tasks to be performed without human interaction. A human could still enter the loop in an emergency or change the goals, although in practice there may be significant time delays before human intervention occurs.



Unmanned Systems Integrated Roadmap FY 2011-2036 IV

Autonomy & unpredictability

- the special feature of an autonomous system is its ability to be goal-directed in *unpredictable situations*.
- this ability is a significant improvement in capability compared to the capabilities of automatic systems.
- an autonomous system is able to *make a decision* based on a set of rules and/or limitations.
- it is able to determine *what information is important* in making a decision.
- it is capable of a *higher level of performance* compared to the performance of a system operating in a predetermined manner.

Autonomy in AWS I

Who takes the decision?

- autonomy is at least about *deliberation* as much as about *action*
- *e.g.*, for artificial weapons, the question is not just

  who pulls the trigger?

  but also / rather

  who decides to pull the trigger?

  and

  based on what evidence?

Autonomy in AWS II

Autonomy & responsibility

- *responsibility* is both an ethical and a legal issue
- autonomy of AWS changes the overall picture
- *multi-level autonomy* has the potential to make things even much more complicated [Sartor and Omicini, 2016]

Autonomy as a Relational Concept [Castelfranchi, 1995] I

Autonomy as a social concept

- an agent is autonomous mostly in relation to other agents
- autonomy has no meaning for an agent in isolation

Autonomy from environment

- the Descartes' problem: (human, agent) behaviour is affected by the environment, but is not depending on the environment
- situatedness, reactiveness, adaptiveness *do not* imply lack of autonomy

Autonomous Goals [Castelfranchi, 1995] I

Agency

- agents as teleonomic, goal-oriented entities
- that is, whose *behaviour* is *not casual* under any acceptation of the term

Autonomous Goals [Castelfranchi, 1995] II

Agents & goals [Conte and Castelfranchi, 1995]

- agents in a society can be generally conceived as either *goal-governed* or *goal-oriented* entities
  - goal-governed entities refer to the strong notion of agency, i.e. agents with some forms of cognitive capabilities, which make it possible to explicitly represent their goals, driving the selection of agent actions
  - goal-oriented entities refer to the weak notion of agency, i.e. agents whose behaviour is directly designed and programmed to achieve some goal, which is not explicitly represented
- in both cases, agent goals are *internal*

Autonomous Goals [Castelfranchi, 1995] III

Executive *vs.* motivational autonomy

executive autonomy — given a goal, the agent is autonomous in achieving it by itself

motivational autonomy the agent's goals are somehow self-generated, not externally imposed

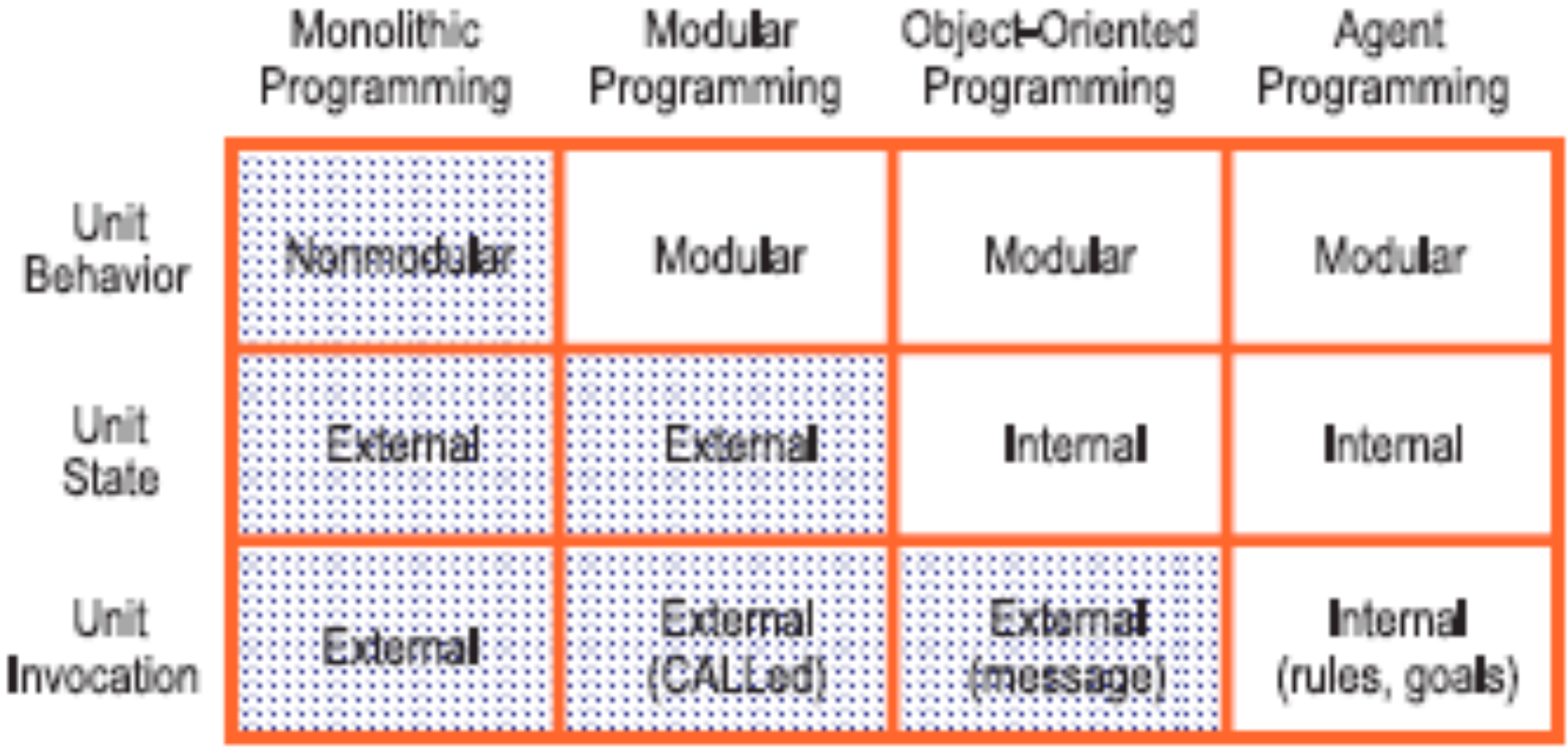
Autonomy & autonomous goals

- autonomy requires *autonomous goals*
- executive autonomy is not enough for real autonomy

Goal-autonomous agent

A *goal-autonomous agent* is an agent endowed with its own *goals*

Evolution of Programming Languages: The Picture



[Odell, 2002]

Evolution of Programming Languages: Dimensions

Historical evolution

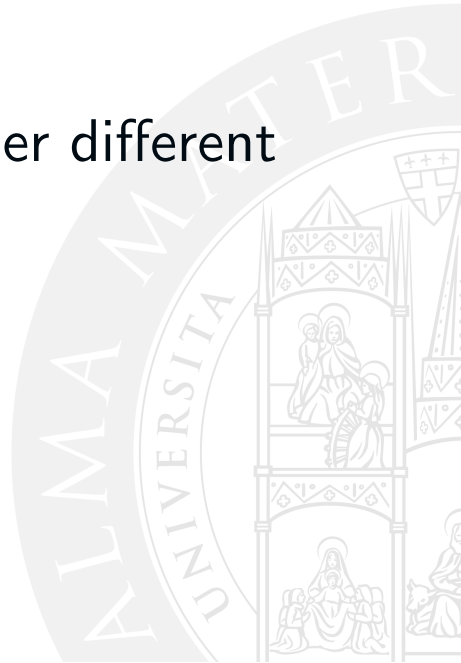
- monolithic programming
- modular programming
- object-oriented programming
- agent programming

Degree of modularity & encapsulation

- unit behaviour
- unit state
- unit invocation

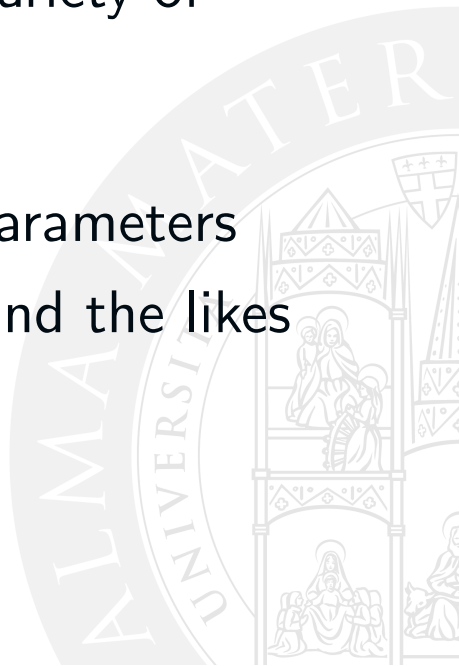
Monolithic Programming

- the basic unit of software is the whole program
- programmer has full control
- program's state is responsibility of the programmer
- program invocation determined by system's operator
- behaviour could not be invoked as a reusable unit under different circumstances
  - modularity does not apply to unit behaviour



Modular Programming

- the basic unit of software are structured loops / subroutines / *procedures* / ...
  - this is the era of procedures as the primary unit of decomposition
- small units of code could actually be *reused* under a variety of situations
  - modularity applies to subroutine's code
- program's *state* is *determined* by externally supplied parameters
- program invocation determined by CALL statements and the likes



Object-Oriented Programming

- the basic unit of software are *objects & classes*
- structured units of code could actually be *reused* under a variety of situations
- objects
  - have *local control* over variables manipulated by their own methods
    - variable state is *persistent* through subsequent invocations
    - object's state is *encapsulated*
  - are *passive*—methods are invoked by external entities
    - modularity does not apply to unit invocation
    - object's *control* is *not encapsulated*

Agent-Oriented Programming

- the basic unit of software are *agents*
  - encapsulating everything, in principle
    - by simply following the pattern of the evolution
  - whatever an agent is
    - we do not need to define them now, just to understand their desired features
- agents
  - could in principle be *reused* under a variety of situations
  - have *control* over their own *state*
  - are *active*
    - they cannot be invoked
    - agent's control is encapsulated

Autonomy as the Foundation of the Definition of Agent

Lex Parsimoniae: Autonomy

- autonomy as the only fundamental and defining feature of agents
- let us see whether other typical agent features follow / descend from this somehow

Computational Autonomy

- agents are autonomous as they encapsulate (the thread of) control
- control does not pass through agent boundaries
  - only data (knowledge, information) crosses agent boundaries
- agents have no interface, cannot be controlled, nor can they be invoked
- looking at agents, MAS can be conceived as an aggregation of multiple distinct *loci* of control interacting with each other by exchanging information

Agents as Autonomous Components

Definition (Agent)

Agents are *autonomous computational entities*

genus agents are computational entities

differentia agents are autonomous, in that they encapsulate control along with a criterion to govern it

Agents are *autonomous*

- from autonomy, many other features stem
  - autonomous agents *are* interactive, social, proactive, and situated;
  - they *might* have goals or tasks, or be reactive, intelligent, mobile
  - they live within MAS, and *interact* with other agents through *communication actions*, and with the environment with *pragmatical actions*

Next in Line...

Intelligent Agents I

According to a classical definition, an intelligent agent is a computational system capable of autonomous action and perception in some environment

Reminder: computational autonomy

- agents are autonomous as they encapsulate (the thread of) control
- control does not pass through agent boundaries
  - only data (knowledge, information) crosses agent boundaries
- agents have no interface, cannot be controlled, nor can they be invoked
- looking at agents, MAS can be conceived as an aggregation of multiple distinct loci of control interacting with each other by exchanging information

Intelligent Agents II

Question: what about the other notions of autonomy?

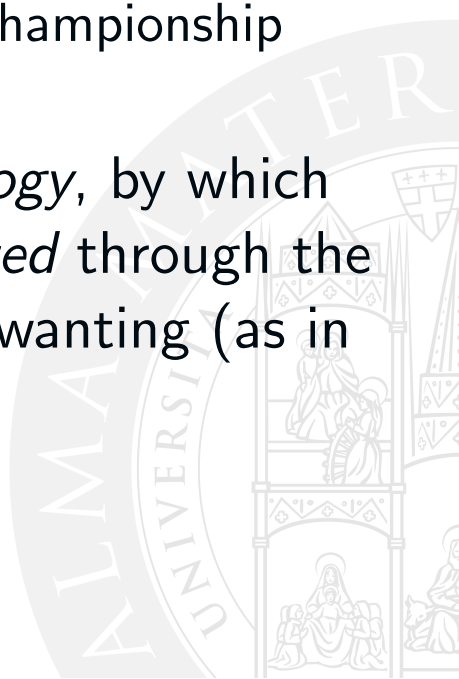
- autonomy with respect to other agents – social autonomy
- autonomy with respect to environment – interactive autonomy
- autonomy with respect to humans – artificial autonomy
- autonomy with respect to oneself – moral autonomy
- ...

Question: what is intelligence to autonomy?

- any sort of intelligence?
- which intelligence for which autonomy?
- which intelligent architecture for which autonomy?
- ...

Intentional Systems

- an idea is to refer to human attitudes as *intentional notions*
- when explaining human activity, it is often useful to make statements such as the following:
  - Seb got rain tires because he believed it was going to rain
  - Kimi is working hard because he wants to win world championship again
- these statements can be read in terms of *folk psychology*, by which human behaviour can be explained and can be *predicted* through the attribution of mental attitudes, such as believing and wanting (as in the above examples), hoping, fearing, and so on.



The Intentional Stance

- the philosopher – cognitive scientist – Daniel Dennett coined the term intentional system to describe entities *'whose behaviour can be predicted by the method of attributing to it belief, desires and rational acumen'* [Dennett, 1971]
- Dennett identifies several grades of intentional systems:
  1. a first-order intentional system has beliefs, desires, etc.
     - Seb believes P
  2. a second-order intentional system has beliefs, desires, etc. about beliefs, desires, etc. both its own and of others
     - Seb believes that Kimi believes P
  3. a third-order intentional system is then something like
     - Seb believes that Kimi believes that Seb believes P
  4. ...

The Intentional Stance in Computing

What entities can be described in terms of *intentional stance*?

- human beings are prone to provide an intentional stance to almost anything
  - sacrifices for ingratiating gods benevolence
  - animism
  - ...

Ascribing mental qualities to machines [McCarthy, 1979]

Ascribing mental qualities like beliefs, intentions and wants to a machine is sometimes correct if done conservatively and is sometimes necessary to express what is known about its state [...] it is useful when the ascription helps us understand the structure of the machine, its past or future behaviour, or how to repair or improve it

Agents as Intentional Systems

Strong notion of agency

- early agent theorists start from a (strong) notion of agents as intentional systems
- agents were explained in terms of mental attitudes, or mental states
- in their social abilities, agents simplest consistent description implied the intentional stance
- agents contain an explicitly-represented – *symbolic* – model of the world (written somewhere in the working memory)
- agents make decision on what action to take in order to achieve their goals via *symbolic* reasoning

Which Domains for Intention Systems? I

Mental states are a worth *abstraction* for developing agents to effectively act in a class of application *domains* characterised by various practical limitations and *requirements* [Rao and Georgeff, 1995]

Which Domains for Intention Systems? II

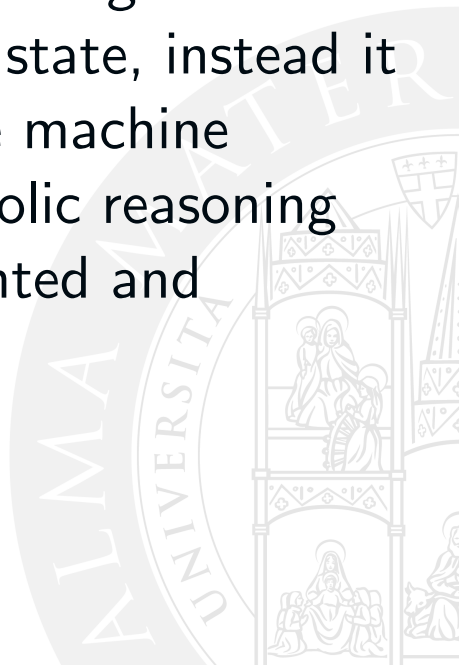
- at any instant of time there are many different ways in which an environment may evolve—the environment is not deterministic
- at any instant of time there are many actions or procedures the agent may execute—the agent is not deterministic, too
- at any instant of time the agent may want to achieve several objectives
- the actions or procedures that (best) achieve the various objectives are dependent on the state of the environment—i.e., on the particular situation, context
- the environment can only be sensed locally
- the rate at which computations and actions can be carried out is within *reasonable bounds* to the rate at which the environment evolves

Goal-Oriented & Goal-Directed Systems

- there are two main families of architectures for agents with mental states

  teleo-reactive / goal-oriented agents are based on their own design model and internal control mechanism. The *goal* is not explicitly represented within the internal state, instead it is an 'end state' for agents internal state machine

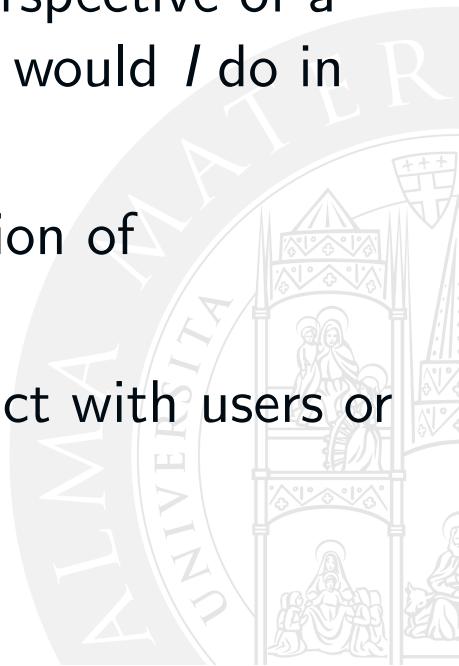
  deliberative / goal-directed agents are based on symbolic reasoning about goals, which are explicitly represented and processed aside the control loop



Modelling Agents with Mental States I

Modelling agents based on mental states. . .

- eases the development of agents exhibiting complex behaviour
- provides us with a familiar, non-technical way of understanding and explaining agents
- allows the developer to build MAS by adopting the perspective of a cognitive entity engaged in complex tasks—e.g., what would *I* do in the same situation?
- simplifies the construction, maintenance, and verification of agent-based applications
- is useful when the agent has to comunicate and interact with users or other system entities



Modelling Agents with Mental States II

The intentional stance [Dennett, 2007]

The intentional stance is the strategy of interpreting the behaviour of an entity (person, animal, artifact, whatever) by treating it as if it were a rational agent who governed its 'choice' of 'action' by a 'consideration' of its 'beliefs' and 'desires'.

The scare-quotes around all these terms draw attention to the fact that some of their standard connotations may be set aside in the interests of exploiting their central features: their role in practical reasoning, and hence in the prediction of the behaviour of practical reasoners.

Modelling Agents with Mental States III

Agents with mental states

- agents governing their behaviour based on *internal states* that mimic cognitive (human) *mental states*

  epistemic states representing agents *knowledge*—their knowledge on the world
  - i.e., percepts, beliefs

  motivational states representing agents *objectives*—what they aim to achieve
  - i.e., goals, desires
- the process of selecting one action to execute among the many available based on the actual mental states is called *practical reasoning*
  - i.e., $action(next(i, perception(e))$

Practical vs. Epistemic Reasoning

practical reasoning is reasoning directed towards actions—the process of figuring out what to do in order to achieve what is desired

[Bratman, 1987]

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.

epistemic reasoning is reasoning directed towards knowledge—the process of updating information, replacing old information (no longer consistent with the world state) with new information

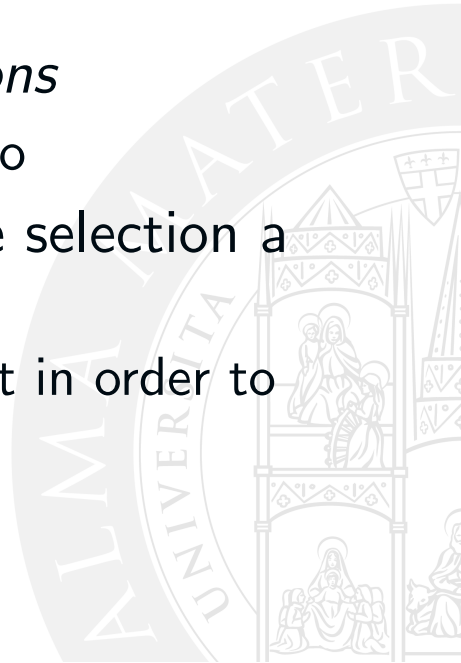
Practical Reasoning

- practical reasoning consists of two main cognitive activities

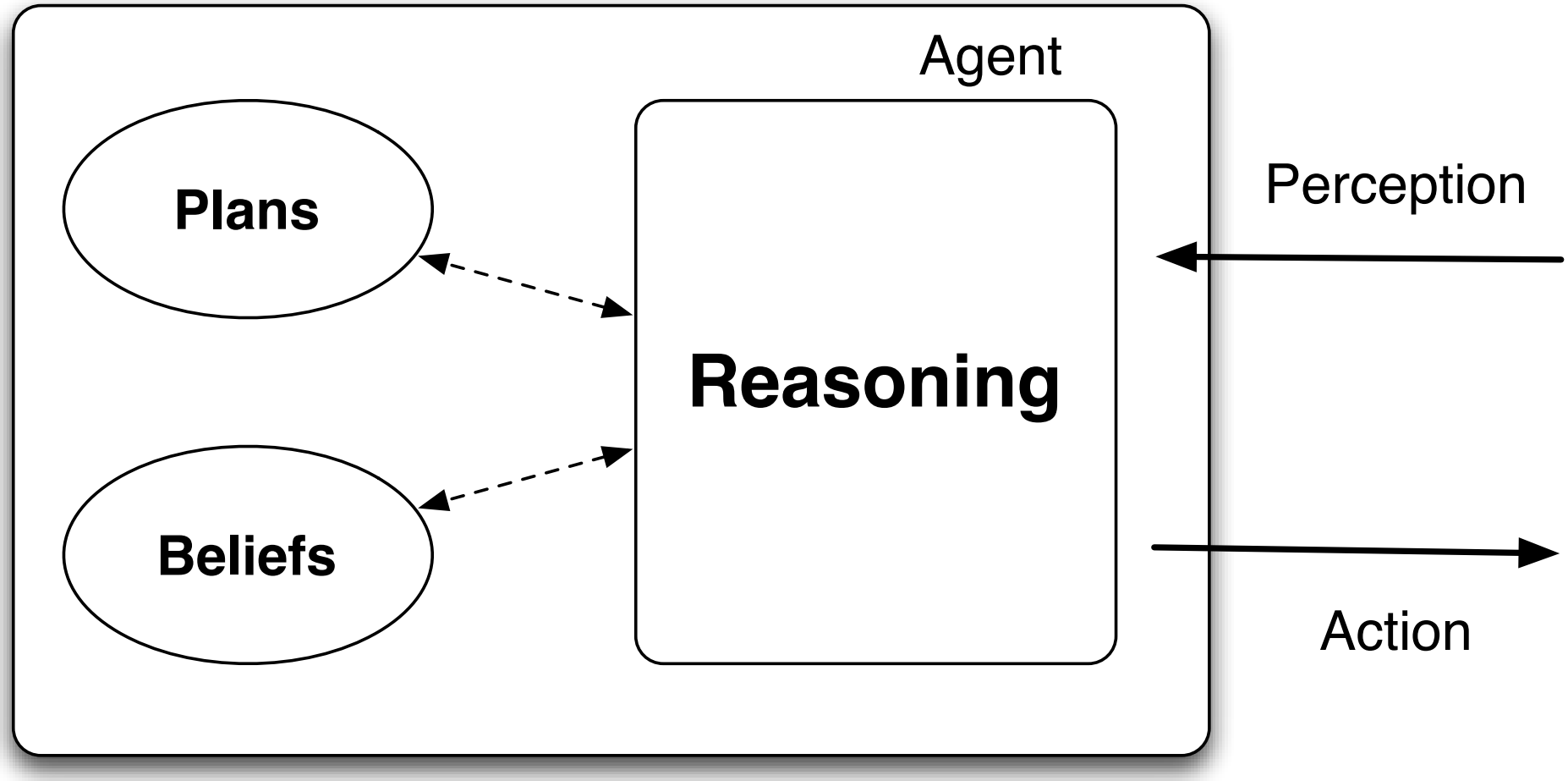
  deliberation when the agent makes decision on what state of affairs the agent desire to achieve

  means-ends reasoning when the agent makes decisions on how to achieve these state of affairs

- the outcome of the deliberation phase are the *intentions*
  - what agent desires to achieve, or what he desires to do
- the outcome of the means-ends reasoning phase is the selection a given course of actions
  - the workflow of the actions the agent intends to adopt in order to achieve its own goals expressed as intentions

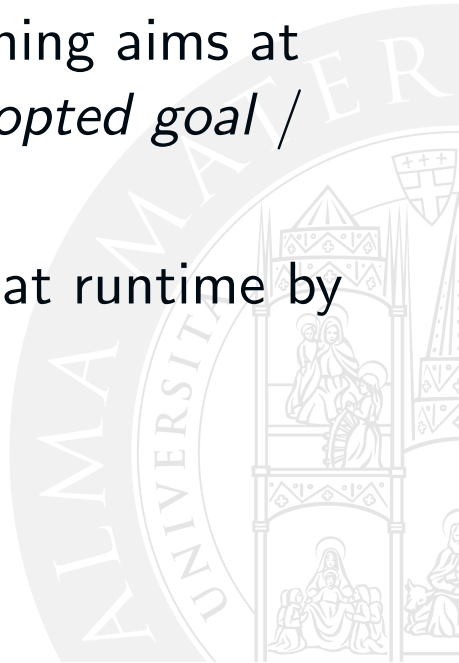


Basic Architecture of a Mentalistic Agent

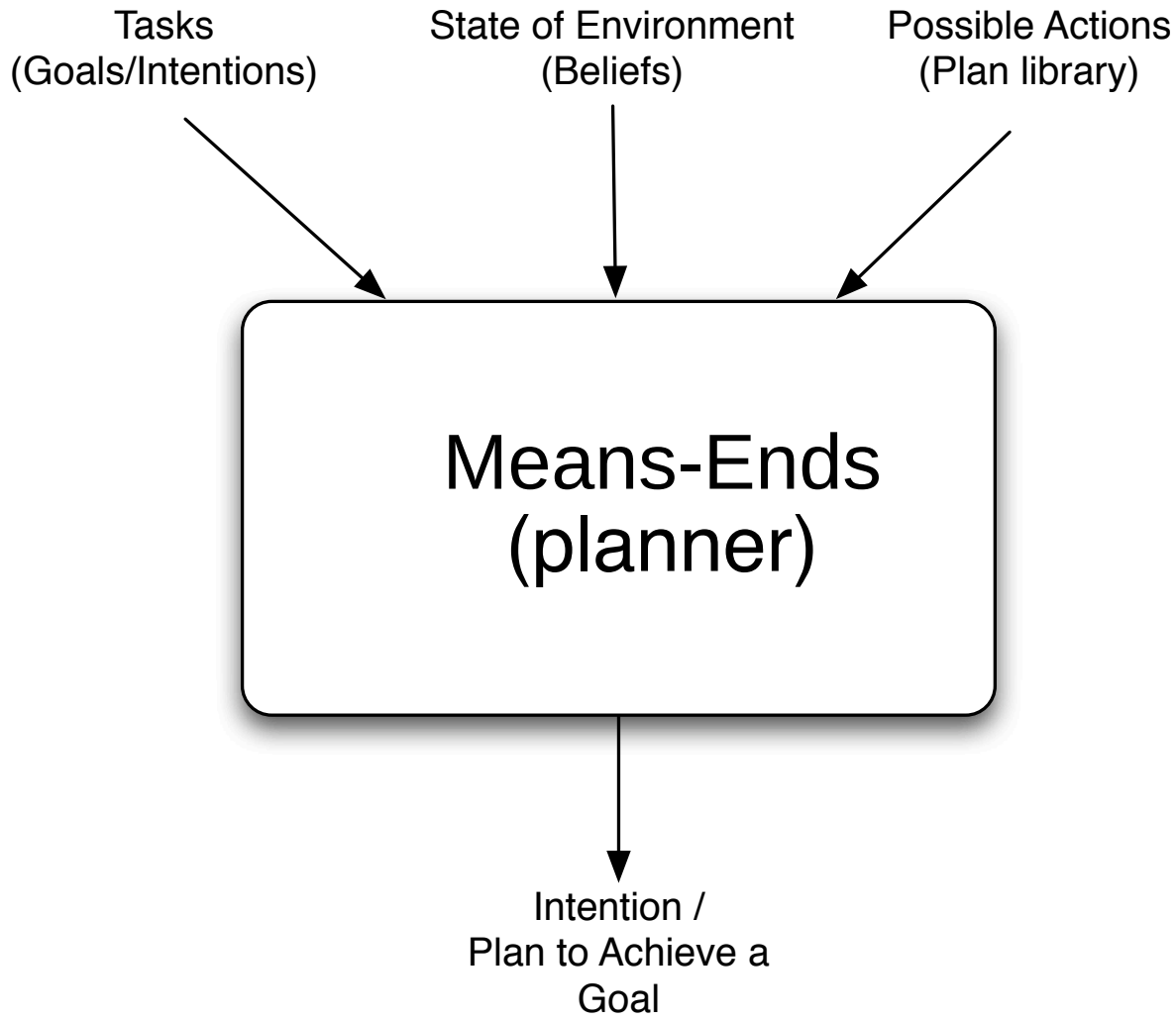


Means-Ends Reasoning I

- the basic idea is to provide agents with three sorts of representations
  - representation of goal / intention to achieve
  - representation of actions / plans – in repertoire
  - representation of the environment
- given the environmental conditions, means-ends reasoning aims at *devising out a plan* that could possibly *achieve the adopted goal* / intention
- the selected intention is an emergent property, reified at runtime by selecting a given plan for achieving a given goal



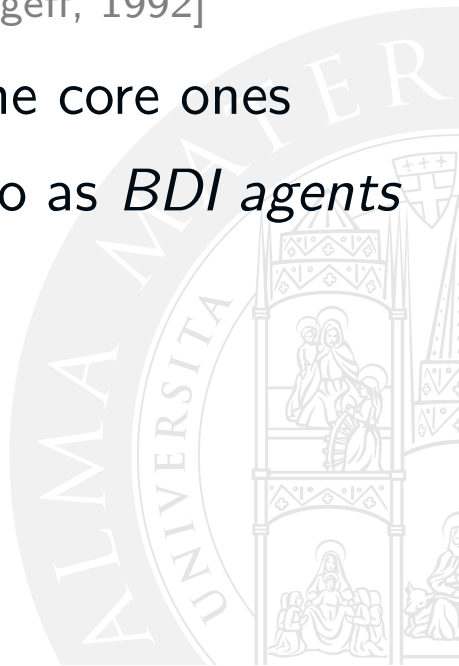
Means-Ends Reasoning II



The BDI Framework I

According to [Dasgupta and Ghose, 2011]

- one of the most popular and successful framework for agent technology is defined by Rao and Georgeff [Rao and Georgeff, 1992]
- there, the notions of *belief*, *desire*, and *intention* are the core ones
- hence, agents in this framework are typically referred to as *BDI agents*



The BDI Framework II

beliefs represent at any time the agent's current *knowledge about the world*, including

- information about the current state of the environment inferred from perception devices
- messages from other agents
- internal information

desires represent a *state of the world* the agent is trying *to achieve*

intentions are the chosen means to achieve the agent's desires, and are generally implemented as plans and post-conditions

- as in general it may have *multiple desires*, an agent can have *a number of intentions active at any one time*
- these intentions may be thought of as *running concurrently*, with *one chosen intention active at any one time*

The BDI Framework III

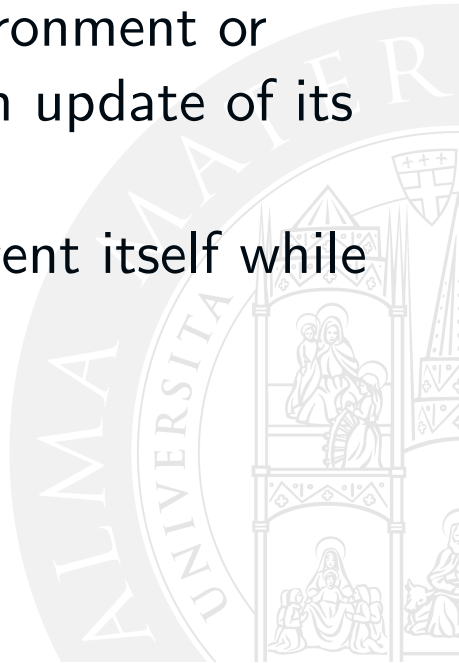
Besides these components, the BDI model includes

plan library — namely, a set of "recipes" representing the *procedural knowledge* of the agent

event queue — where

- *events* — either perceived from the environment or generated by the agent itself to notify an update of its belief base
- *internal subgoals* — generated by the agent itself while trying to achieve a desire

are stored.



The BDI Framework IV

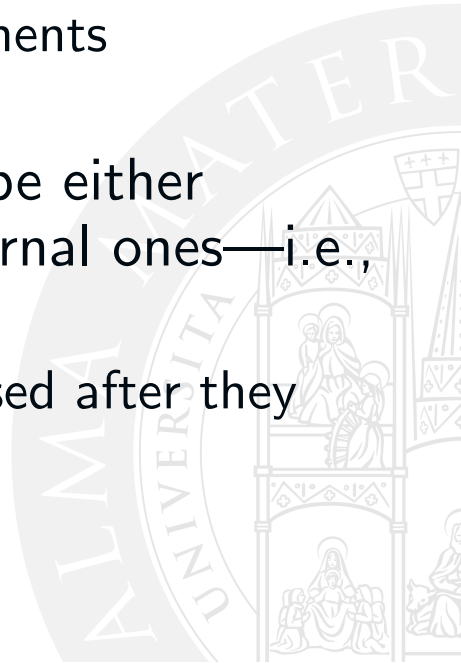
Plans & plan library

- usually, BDI-style agents do no adopt first principles planning at all
- all plans must be *generated* by the agent programmer *at design time*, which are then *selected* for execution at run time
- pre-programmed plans are collected in the plan library
- the planning done by agents consists entirely of context-sensitive subgoal expansion, which is deferred until a subgoal is selected for execution

The BDI Abstract Architecture

Accordingly, the *abstract architecture* proposed by [Rao and Georgeff, 1992] comprise three dynamic and global structures representing agent beliefs, desires, and intentions (BDI), along with an input queue of events

- *update* (write) and *query* (read) operations are possible upon the three structures
  - update operation are subject to compatibility requirements
  - formalised constraints hold upon the mental attitudes
- the events that the system is able to recognise could be either external – i.e., coming from the environment – or internal ones—i.e., coming from some reflexive action
  - events are assumed to be *atomic*, and can be recognised after they have occurred

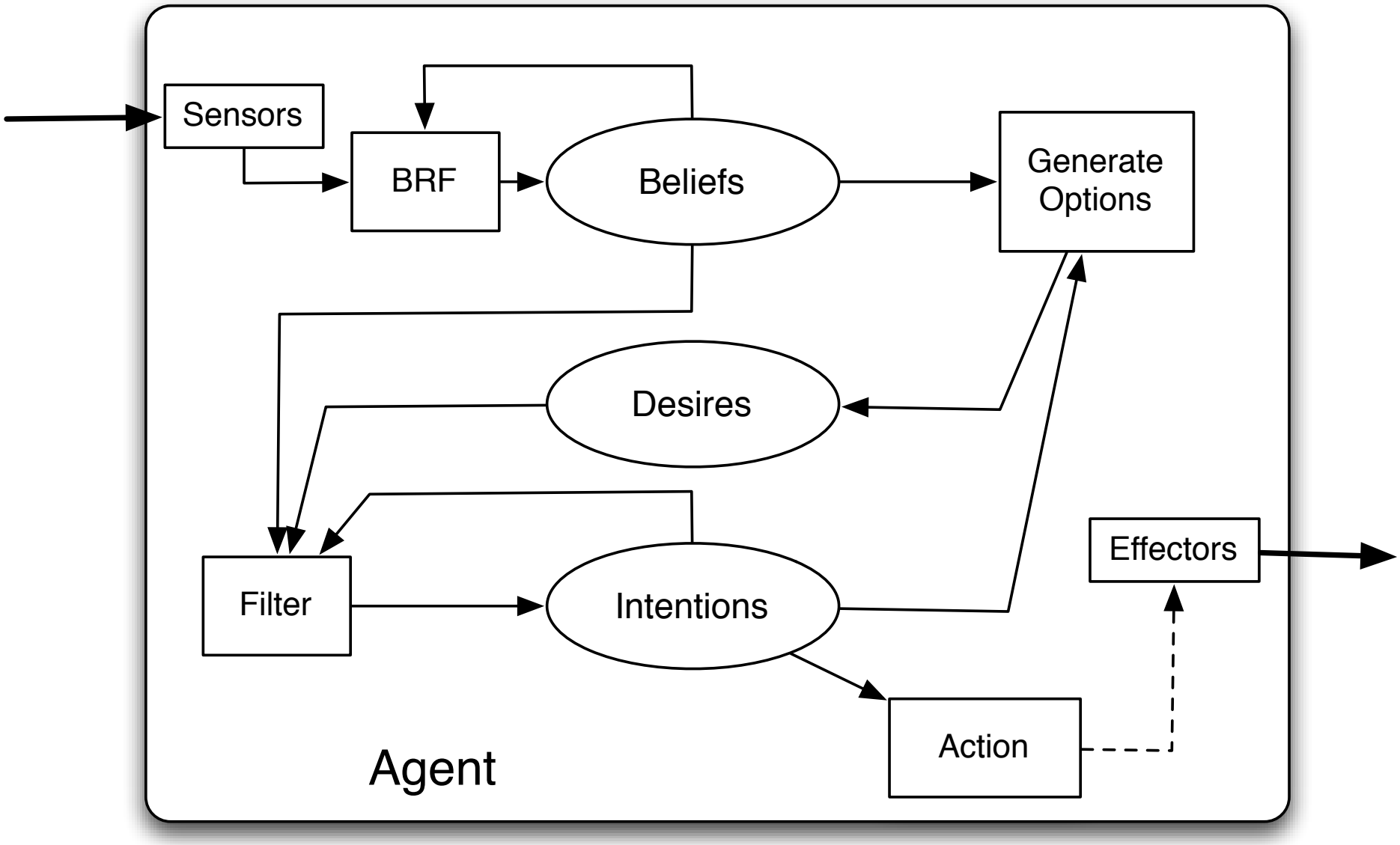


Structure of BDI Systems

BDI architectures are based on the following constructs

- a set of beliefs
- a set of desires (or goals)
- a set of intentions
  - or better, a subset of the goals with an associated stack of plans for achieving them; these are the intended actions
- a set of internal events
  - elicited by a belief change (i.e., updates, addition, deletion) or by goal events (i.e. a goal achievement, or a new goal adoption)
- a set of external events
  - perceptive events coming form the interaction with external entities (i.e. message arrival, signals, etc.)
- a plan library (repertoire of actions) as a further (static) component

Basic Architecture of a BDI Agent [Wooldridge, 2002]



BDI Agents Programming Platforms

Jason (Brasil) http://jason.sourceforge.net/
Agent platform and language for BDI agents based on AgentSpeak(L)

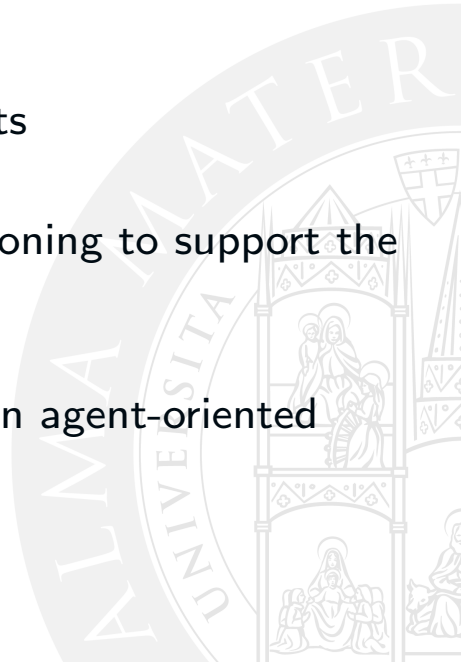
JADEX (Germany) http://www.activecomponents.org/
Agent platform for BDI and Goal-Directed Agents

2APL (Netherlands) http://www.cs.uu.nl/2apl/
Agent platform providing programming constructs to implement cognitive agents based on the BDI architecture

3APL (Netherlands) http://www.cs.uu.nl/3apl/
A programming language for implementing cognitive agents

PRACTIONIST (Italy) http://practionist.eng.it/
Framework built on the Bratman's theory of practical reasoning to support the development of BDI agents

ASTRA http://astralanguage.com/
A distributed / concurrent programming language based on agent-oriented programming

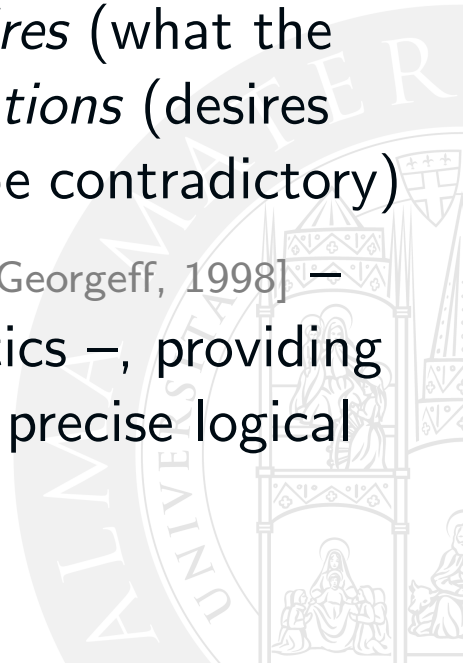


BDI Viewpoints I

There are three main viewpoints over the BDI Model [Mazal et al., 2008]:

philosophical based on the work of philosopher Bratman [Bratman, 1987], using uses terms of folk psychology to view humans as planning agents: the main concepts in his work are *beliefs* (what an agent knows about the world), *desires* (what the agent wants, can be contradictory) and *intentions* (desires that the agent has decided to reach, cannot be contradictory)

logical mainly Rao and Georgeff's BDI CTL [Rao and Georgeff, 1998] – multimodal logics with possible world semantics –, providing beliefs, goals (desires), and intentions with a precise logical semantics



BDI Viewpoints II

implementation there are a huge number of systems and technologies that are said to conform to the BDI model—between the BDI CTL logics (very expressive) and the implementing systems. which then treat the main modalities rather as data structures, and mostly focus on plans

Next in Line...

Complex Systems

. . . by a complex system I mean one made up of a large number of parts that interact in a non simple way [Simon, 1962]

Which "parts" for complex systems?

- is autonomy of "parts" a necessary precondition?
- is it also sufficient?

Which kind of systems are we looking for?

- what is *autonomy* for a system as a whole?
- where could we find significant examples?

Nature-inspired Models

Complex natural systems

- such as physical, chemical, biochemical, biological, social systems
- natural system exhibit *features*
  - such as distribution, opennes, situation, fault tolerance, robustness, adaptiveness, ...
- which we would like to understand, capture, then bring to *computational* systems

Nature-inspired computing (NIC)

- for instance, NIC [Liu and Tsui, 2006] summarises decades of research activities
- putting emphasis on
  - autonomy of components
  - self-organisation of systems

Autonomy & Interaction I

Self-organisation

- autonomy of *systems* requires essentially the same features that self-organising systems exhibit
- . . . such as opennes, situation, fault tolerance, robustness, adaptiveness, . . .

(say it again)

> *. . . by a complex system I mean one made up of a large number of parts that interact in a non simple way* [Simon, 1962]

Autonomy & interaction

- "parts" should be *autonomous*
- *interaction* is essential as well

MAS & Complexity I

Autonomy & interaction

- multi-agent systems (MAS) are built around autonomous components
- how can MAS deal with self-organisation, properly handling interaction among components?

The observation of self-organising termite societies led to the following observation:

> *The coordination of tasks and the regulation of constructions are not directly dependent from the workers, but from constructions themselves* [Grassé, 1959]

MAS & Complexity II

Coordination as the key issue

- many well-known examples of natural systems – and, more generally, of complex systems – seemingly rely on simple yet powerful coordination mechanisms for their key features—such as self-organisation
- it makes sense to focus on coordination models as the core of complex MAS
- . . . since they are conceived to deal with the complexity of interaction [Omicini, 2013]

Basic Issues of Nature-inspired Coordination I

Environment

- environment is essential in nature-inspired coordination
  - it works as a mediator for component interaction — through which the components of a distributed system can communicate and coordinate indirectly
  - it is active — featuring autonomous dynamics, and affecting component coordination
  - it has a structure — requiring a notion of *locality*, and allowing components of any sort to *move* through a topology

Basic Issues of Nature-inspired Coordination II

Stochastic behaviour

- complex systems typically require probabilistic models
  - *don't know* / *don't care* non-deterministic mechanisms are not expressive enough to capture all the properties of complex systems such as biochemical and social systems
  - probabilistic mechanisms are required to fully capture the dynamics of coordination in nature-inspired systems
  - coordination models should feature (possibly simple yet) expressive mechanisms to provide coordinated systems with stochastic behaviours

Coordination Middleware I

Middleware

- coordination middleware to build complex software environment
- coordination abstractions to embed situatedness and stochastic behaviours
- coordination abstractions to embed social rules–such as *norms*, and *laws*

Coordination Middleware II

Nature-inspired middleware

- starting from early chemical and stigmergic approaches, nature-inspired models of coordination evolved to become the *core* of *complex distributed systems*—such as pervasive, knowledge-intensive, intelligent, and self-* systems
- this particularly holds for tuple-based coordination models

  [Omicini and Viroli, 2011]
- tuple-based middleware as a perspective technology for complex autonomous systems

References I

Bratman, M. E. (1987).
Intention, Plans, and Practical Reason.
Harvard University Press.

Castelfranchi, C. (1995).
Guarantees for autonomy in cognitive agent architecture.
In Wooldridge, M. J. and Jennings, N. R., editors, *Intelligent Agents*, volume 890 of *Lecture Notes in Computer Science*, pages 56–70. Springer Berlin Heidelberg.
ECAI-94 Workshop on Agent Theories, Architectures, and Languages (ATAL) Amsterdam, The Netherlands 8–9 August 1994. Proceedings.

Conte, R. and Castelfranchi, C. (1995).
Cognitive and Social Action.
UCL Press.

Dasgupta, A. and Ghose, A. K. (2011).
BDI agents with objectives and preferences.
In Omicini, A., Sardina, S., and Vasconcelos, W., editors, *Declarative Agent Languages and Technologies VIII*, volume 6619 of *Lecture Notes in Computer Science*, pages 22–39. Springer Berlin Heidelberg.

References II

Dennett, D. (1971).
Intentional systems.
Journal of Philosophy, 68:87–106.

Dennett, D. (2007).
Intentional systems theory.
In Beckermann, A. and Walter, S., editors, *Oxford Handbook of the Philosophy of Mind*, chapter 19. Oxford University Press.

Edwards, G. (2013).
Military autonomous & robotic systems. Considerations for the way forward from a UK military perspective.
Air Power Review, 16(3):50–71.

Gould, S. J. (2002).
The Structure of Evolutionary Theory.
The Belknap Press of Harvard University Press.

References III

Grassé, P.-P. (1959).
La reconstruction du nid et les coordinations interindividuelles chez Bellicositermes natalensis et Cubitermes sp. la théorie de la stigmergie: Essai d'interprétation du comportement des termites constructeurs.
Insectes Sociaux, 6(1):41–80.

Kitano, H. (2002).
Foundations of Systems Biology.
MIT Press.

Liu, J. and Tsui, K. C. (2006).
Toward nature-inspired computing.
Communications of the ACM, 49(10):59–64.

Maturana, H. R. and Varela, F. G. (1980).
Autopoiesis and Cognition: The Realization of the Living, volume 42 of *Boston Studies in the Philosophy of Science*.
D. Reidel Publishing Company, Dordrecht, Holland.

References IV

Mazal, Z., Kočí, R., Janoušek, V., and Zbořil, F. (2008).
PNagent: A framework for modelling BDI agents using object oriented Petri nets.
In *2008 8th International Conference on Intelligent Systems Design and Applications (ISDA 2008)*, volume 2, pages 420–425. IEEE.

McCarthy, J. (1979).
Ascribing mental qualities to machines.
In Ringle, M., editor, *Philosophical Perspectives in Artificial Intelligence*, Harvester Studies in Cognitive Science, pages 161–195. Harvester Press, Brighton.

Odell, J. (2002).
Objects and agents compared.
Journal of Object Technologies, 1(1):41–53.

Omicini, A. (2013).
Nature-inspired coordination models: Current status, future trends.
ISRN Software Engineering, 2013.
Article ID 384903, Review Article.

References V

Omicini, A. and Viroli, M. (2011).
Coordination models and languages: From parallel computing to self-organisation.
The Knowledge Engineering Review, 26(1):53–59.
Special Issue 01 (25th Anniversary Issue).

Rao, A. S. and Georgeff, M. P. (1992).
An abstract architecture for rational agents.
In Nebel, B., Rich, C., and Swartout, W. R., editors, *3rd International Conference on Principles of Knowledge Representation and Reasoning (KR '92)*, pages 439–449, Cambridge, MA, USA. Morgan Kaufmann.
Proceedings.

Rao, A. S. and Georgeff, M. P. (1995).
BDI agents: From theory to practice.
In Lesser, V. R. and Gasser, L., editors, *1st International Conference on Multi Agent Systems (ICMAS 1995)*, pages 312–319, San Francisco, CA, USA. The MIT Press.

Rao, A. S. and Georgeff, M. P. (1998).
Decision procedures for BDI logics.
Journal of Logic and Computation, 8(3):293–342.

References VI

Rosslenbroich, B. (2014).
On the Origin of Autonomy. A New Look at the Major Transitions in Evolution.
History, Philosophy and Theory of the Life Sciences. Springer International Publishing.

Sartor, G. and Omicini, A. (2016).
The autonomy of technological systems and responsibilities for their use.
In Bhuta, N., Beck, S., Geiß, R., Liu, H.-Y., and Kreß, C., editors, *Autonomous Weapon Systems. Law, Ethics, Policy*, chapter 3, pages 39–74. Cambridge University Press, Cambridge, UK.

Simon, H. A. (1962).
The architecture of complexity.
Proceedings of the American Philosophical Society, 106(6):467–482.

Thompson, E. (2010).
Mind in Life: Biology, Phenomenology, and the Sciences of Mind.
Belknap Press.

Varela, F. G., Maturana, H. R., and Uribe, R. (1974).
Autopoiesis: The organization of living systems, its characterization and a model.
Biosystems, 5(4):187–196.

References VII

Wooldridge, M. J. (2002).
An Introduction to MultiAgent Systems.
John Wiley & Sons Ltd., Chichester, UK.

Frontiers of Autonomous Systems

Digital Culture

Andrea Omicini
andrea.omicini@unibo.it

Master in Business Management / Bologna Business School

Academic Year 2017/2018