## Philosophical Foundations

- Aristoteles: practical syllogisms
- Dennett's intentional stance : B & D
- Bratman : intentions of rational agents
- Cohen & Levesque : theory of intentions
- Rao & Georgeff : BDI theory/logic/arch.
- Brooks: non-cognitive reactive systems
- Sloman : emotional agents
- Asimov : social robots: laws of robotics

46

## 'Practical' Reasoning (Aristoteles)

- An 'ordinary' practical syllogism

*Exercise would be good for me.*
*Jogging is exercise.*

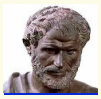
*Therefore, jogging would be good for me.*

- 'Just' deductive / logical reasoning
  - "Da's nogal logisch…"

47

## 'Practical' Reasoning (Aristoteles)

- More interesting practical syllogism



*Would that I exercise.*
*Jogging is exercise.*

*Therefore, I shall go jogging.*

- No deduction, but rather a specification of the action selection / decision of the agent!

48

## Dennett's intentional stance

- The *intentional stance* is the strategy of interpreting the behaviour of an entity by treating it *as if it were* a *rational agent* that governed its *choice* of *action* by a *consideration* of its *beliefs* and *desires*

  - Anthropomorphic instance of the *design (functionality) stance*, contra the *physical stance*
  - Instrumental / operational use of beliefs and desires of *human beings*: no causally active inner states of people, just calculational devices

49

## Background: Dennett's philosophy of consciousness

- *Reduction* of human consciousness to an illusionary feature of a 'virtual machine' running on the brain as hardware
- This hardware can be something different, provided *sufficiently complicated*
- In principle it might be a *computer*, so that a computer may run a 'consciousness program' → a *'conscious computer'*
- Dennett is a proponent of *Strong AI*

50

## Bratman : the role of intentions

- *Rational* behavior needs, besides *beliefs* and *desires*, also *intentions*
- Two justifications for this:
  - (Resource-bounded)agents need to settle on some desire(s) and commit themselves
  - Co-ordination of future actions after commitment(s)

51

## Bratman

- Intentions, *unlike mere desires*, play the following functional roles:
  - Intentions normally pose *problems* for the agent; the agent needs to determine a way to achieve them → *focus on solving concrete problems*
  - Intentions provide a "screen of admissibility" for *adopting other intentions*
  - Agents "track" the success of their attempts to achieve their intentions -- *replanning*

52

## Cohen & Levesque : intentions

- A 'tiered' formalism
  - Atomic layer: beliefs, goals, actions
  - Molecular layer: concepts defined in terms of primitives, e.g. intention
- Intention = choice + commitment

Hector Levesque 53

## Cohen & Levesque

- Intention should satisfy the following: if an agent intends to achieve p, then:
  - The agent believes p is *possible*
  - The agent does *not* believe he will *not* bring about p
  - Under certain conditions, the agent believes he *will* bring about p
  - Agents need *not* intend all the expected *side-effects* of their intentions

54

## Rao & Georgeff : BDI theory

- *"Rational agent* possesses mental attitudes of *beliefs, desires and intentions*, representing the *information, motivational, and deliberative states of an agent*, respectively"
- "These mental attitudes determine the system's behaviour and are critical for achieving *adequate* or *optimal* performance when deliberation is subject to *resource bounds*" --- *computational* perspective!

Michael Georgeff

## Rodney Brooks: '*anti*-cogn.robotics'

- Disappointed with GOFAI/reasoning approach to robotics
- Brooks' approach:
  - *bottom-up* instead of top-down
  - *Subsumption architecture : 'reactive system', no modelling ('world is best model'), no 'thinking'*
  - 'cockroach AI' (Genghis)
  - humanoid robot (Cog)
  - 'Having a mind'/ thinking is an *emergent* property of sufficiently complex systems?

56

## Beyond rationality: BDI+ Brooks on emotions

emotions!

- Tradition: *emotions* versus *rationality*
- Can *machines* have emotions?
- Brooks:

  Humans are machines
  Humans have emotions
  Ergo: there are machines with emotions

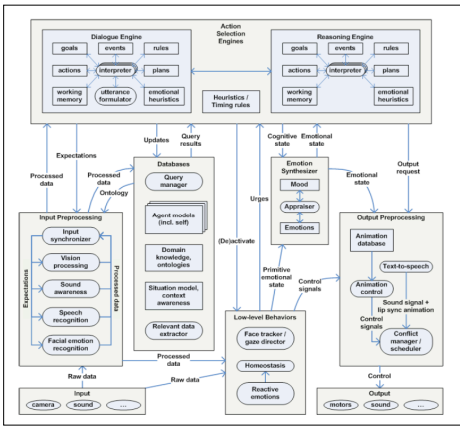
- Once you have concluded this, there is no problem of *ascribing emotions to machines*!

57

## Aaron Sloman : how to build emotional agents?


- Humans are machines, though not artefacts
- What kind of machines?
- In particular, people have *emotions*
- *Are emotions and rationality opposites?*
- What kind of machines has *emotions*?
- Implications for AI and I(ntell.)A(gents)?
  - How do we *employ* emotions fruitfully in agent design?

58




Boon Companion Architecture

59

## Towards social agents / robots

- Individual agent has *mental / cognitive* attitudes
- It is even more interesting to put *several* agents into the same environment
  - *Multi*-agent systems (MAS), or even
  - Agent *societies*
- *Social* attitudes become important
  - 'Agent *ethics*' (what is 'good' behaviour?)
  - Artificial Normative Systems

60

## Asimov's laws of robotics


- *A robot may not injure a human being, or, through inaction, allow a human being to come to harm*
- *A robot must obey the orders given it by human beings except where such orders would conflict with the above*
- *A robot must protect its own existence as long as such protection does not conflict with the above rules*

61

## Asimov's laws of robotics

- Questions:
  - How realistic are these laws?
  - How easily *implementable / realizable* are these laws?
    - What AI techniques are needed for implementation?

62

## Theory of Mind

- Notion from biology / ethology and child psychology
- Means that an agent has a theory and can *reason about the possible mind of another agent*
- Seems also useful for some applications of artificial agents
- Currently under investigation in our group in context of applications such as companion robots and virtiual characters in video games

63