Stronger Notion, mentalistic notions: *anthropomorphism*

- Information attitudes
  - Knowledge
  - Belief
- Pro–attitudes (guide the agent's actions)
  - Intention
  - Goal
  - Desire
  - Obligation (Shoham: Agent–Oriented programming)
  - Emotion (Joe Bates: Believable agents)

- What is an agent?
- Weak Notion:
  - Autonomy (operate without intervention, own control)
  - Social ability (agent communication language)
  - Reactivity (perception of the environment, action)
  - Pro–activeness (goal–directed, *taking the initiative*)
- Self–contained, concurrently executed software process, that encapsulates some state and is able to communicate with other agents via message passing.
- Object–Based concurrent programming (Agha,86: ACTORS)

- AGENTS:
  - Theories (specification)
  - Architectures (implementation)
  - Languages (programming)
- INTENTIONAL SYSTEMS
  - Very complex systems: animistic, intentional explanations, abstraction versus mechanistic interpretations
  - Information attitudes: belief, knowledge
  - Pro–attitudes: desire, intention, obligation, commitment, choice, ...

- Visual representation (Maes: Interface Agents)
- Mobility
- Veracity (no communicate false information)
- Benevolence (it try to do what is asked of it)
- Rationality (it will act to achieve its goals)
- Solitary, parasite, social, selfish ...

SOLUTIONS:

- Syntactic:
  - Modal language: non–truth–functional modal operators
  - Meta–language: a first–order language containing formulae of some other *object–language*

$$Bel(Janine, \lceil Father(Zeus, Cronos) \rceil)$$

- Semantic:
  - Possible worlds
  - Sentencial models

Janine believes Cronos is the father of Zeus

$$Bel(Janine, Father(Zeus, Cronos))$$

Two problems:

1. Syntactic: not well–founded formula of FOL
2. Semantic: Suposse $(Zeus = Jupiter)$, then $Bel(Janine, Father(Jupiter, Cronos))$ can not be inferred. Belief is not truth functional.

The operators are duals:

$$\Box\varphi \Leftrightarrow \neg\Diamond\neg\varphi$$

$$\Diamond\varphi \Leftrightarrow \neg\Box\neg\varphi$$

- Axiom K (Kripke): $\Box(\varphi \Rightarrow \psi) \Rightarrow (\Box\varphi \Rightarrow \Box\psi)$
- Necessitation rule: if φ is valid, then $\Box\varphi$ is valid

Four axioms:

T (reflexive) $\Box\varphi \Rightarrow \varphi$

D (serial) $\Box\varphi \Rightarrow \Diamond\varphi$

4 (transitive) $\Box\varphi \Rightarrow \Box\Box\varphi$

5 (euclidean) $\Diamond\varphi \Rightarrow \Box\Diamond\varphi$

NORMAL MODAL LOGIC

Classic propositional logic + two operators: $\Box$ (necessarily) and $\Diamond$ (possibly)

Atomic propositions $Prop = \{p, q, \ldots\}$

1. if $p \in Prop$ then p is a formula
2. if φ, ψ are formulae, then so are $\neg\varphi$ and $\varphi \vee \psi$
3. if φ is a formula, then so are $\Box\varphi$ and $\Diamond\varphi$

Accesibility relation: what world are accesible from every other world

- $\Box\varphi$ is true if φ is true in every accesible world
- $\Diamond\varphi$ is true if φ is true in at least one accesible world

T (knowledge) $K_i\varphi \Rightarrow \varphi$ (distintion between knowledge and belief)

D (non-contradictory) $K_i\varphi \Rightarrow \neg K_i \neg \varphi$

4 (positive introspection) $K_i\varphi \Rightarrow K_i K_i \varphi$

5 (negative introspection) $\neg K_i\varphi \Rightarrow K_i \neg K_i \varphi$

Epistemic Logic: KTD45 (idealised) knowledge

Doxastic Logic: KD45 (idealised) Belief

EPISTEMIC LOGIC:

$\Box\varphi$, $K\varphi$, *it is known that* φ

Agents: $K_i\varphi$, *i knows that* φ

Logical omniscience problem:

Necessitation rule: if φ is valid, then $K_i\varphi$ is valid. An agent knows all valid formulae.

Axiom K: $K_i(\varphi \Rightarrow \psi) \Rightarrow (K_i\varphi \Rightarrow K_i\psi)$. Agent's knowledge is closed under implication.

INTENTIONS: COHEN AND LEVESQUE

- Primary Modalities
  - (BEL x p) Agent x believes p
  - (GOAL x p) Agent x has a goal of p
  - (DONE x a) Agent x has just performed action a
  - (DOING x a) Agent x is just performing action a
- Event Sequences
  - $a; b$ Action a followed by action b
  - $a?$ Test action
- Standard future operations of temporal logic
  - $\Box$ Always
  - $\Diamond$ Sometimes

THEORIES OF AGENCY

- Dynamic aspects
- Relation between information attitudes and pro–attitudes
- Changes of the cognitive state over time
- Environment changes de cognitive state
- How to perform actions

INTENTION

(INTENT x p q) =

(PGOAL x

(DONE x

[UNTIL (DONE x a) (BEL x (DOING x a))]?; a)

q)

PERSISTENT GOALS: Agent x, Goal p, Motive q

(PGOAL x p q) =

1. (BEL x $\neg p$) $\wedge$
2. (GOAL x $\Diamond p$) $\wedge$
3. (UNTIL [(BEL x p) $\vee$ (BEL x $\neg\Box p$) $\vee$ (BEL x $\neg q$)] (GOAL x $\Diamond p$))

1. Agent x believes goal p is currently false
2. Agent x wants goal p to be eventually true
3. Continue until agent x believes that p is true or will never be true, or that the motivation q is no longer present

	OOP	AOP
Basic unit	object	agent
State	unconstrained	beliefs, commitments, capabilities, choices, ...
Computation	message passing, methods	message passing, methods
Type of messages	unconstrained	inform, request, offer, promise, decline, ...
Constrains on methods	none	honesty, consistency, ...

AGENT ORIENTED PROGRAMMING (AOP)

YOAV SHOHAM

STANFORD UNIVERSITY

- specialization OOP
- mental state: beliefs, decisions, capabilities, obligations
- speech act theory: informing, requesting, offering, ...

Decision (choice) freedom central to the notion of agenthood

$$DEC_a^t\varphi \stackrel{def}{=} OBL_{a,b}^t\varphi$$

commitment to oneself

Capability at time t agent a is capable of φ

$$CAN_a^t\varphi$$

$$CAN_{robot}^5 open(door)^8$$

immediate version of CAN

$$ABLE_a\varphi \stackrel{def}{=} CAN_a^{time(\varphi)}\varphi$$

$$ABLE_{robot} open(door)^5 \stackrel{def}{=} CAN_{robot}^5 open(door)^5$$

LANGUAGE

Time simple point–based temporal language

$$holding(robot, cup)^t$$

Action instantaneous

$$raise\_arm(robot)^t$$

Belief at time t agent a believes a (recursively defined) sentence φ

$$B_a^t\varphi$$

$$B_a^3 B_b^{10} like(a, b)^7$$

Obligation at time t agent a is obligated (committed) to agent b about φ

$$OBL_{a,b}^t\varphi$$

Introspection no total introspective capabilities, obligation

- for any t, a, b, φ: $OBL^t_{a,b}\varphi \equiv B^t_a OBL^t_{a,b}\varphi$
- for any t, a, b, φ: $\neg OBL^t_{a,b}\varphi \equiv B^t_a \neg OBL^t_{a,b}\varphi$

Persistence of mental state mental states persist over time

- beliefs persist (until learn contradiction)
- absence of beliefs persist (until learn)
- obligations (and decisions) persist (until they are revoked)
- capabilities are fixed

PROPERTIES

Internal consistency beliefs and obligations are internally consistent

- for any t, a: $\{\varphi : B^t_a\varphi\}$ is consistent
- for any t, a: $\{\varphi : OBL^t_{a,b}\varphi \text{ for some } b\}$ is consistent

Good faith agents commit only to what they believe themselves capable of and only if they really mean it

- for any t, a, b, φ: $OBL^t_{a,b}\varphi \Rightarrow B^t_a((ABLE_a\varphi) \wedge \varphi)$

GENERIC INTERPRETER

GENERIC AGENT INTERPRETER

- Basic loop
  1. read current messages, update mental state (belief and commitments)
  2. execute commitments for the current time, possibly resulting in further belief change
- message passing: addressable by name
- clock: synchronization

```
(UNREQUEST t a action)
(REFRAIN action)

(REQUEST 1 a (REQUEST 5 b (INFORM 10 c fact)))
```

– conditional: mental condition, logical combination of mental patterns

```
(B fact)                    ((CMT a) action)
```

conditional action

```
(IF mntlcond action)


(REQUEST t b (IF (B fact) (INFORM t+1 a fact)))
(REQUEST t b (IF (B NOT(fact))
                 (INFORM t+1 a NOT(fact))))
```


AGENT0

- Fact statements: atomic objective sentences

```
(t (employee (smith acme)))
(NOT (t (employee (smith acme))))
```

- Private and communicative action statements: private or communicative; conditional or unconditional

  – private

```
(DO t p-action)
```

  – communicative: informing, requesting, cancelling

```
(INFORM t a fact)
(REQUEST t a action)
```

AGENT0 PROGRAM:

- Time grain
- Agent's capabilities
- Initial beliefs
- A sequence of commitment rules

- Variables

```
(IF (NOT ((CMT ?x) (REFRAIN sing))) sing)
(IF (B (t (emp ?!x acme)))
      (INFORM a (t (emp ?!x acme))))
```

- Commitment rules
  - conditional: message condition, logical combination of message patterns

    `(From Type Contents)` `(a REQUEST (DO t walk))`
  - commitment rule

```
(COMMIT msgcond mntlcond (agent action)*)
(COMMIT (?a REQUEST ?action)
        (B (now (myfriend ?a)))
   (?a ?action))
```

- the mental condition part allows one to prevent commitment to incompatible actions

```
((?!time (rotate wheelbase ?degrees))
   (NOT ((CMT ?x) ?!time (service wheelbase))))
```

- belief change affect capabilities, private actions depend of mental preconditions
- belief update: examines the current commitments preconditions
- `UNREQUEST`: removes the commitment

AGENT0 INTERPRETER, DATABASES:

beliefs

- updated when informed or as a result of a private action
- they incorporate any fact of which they are informed, retracting the contradictory atomic belief if that were previously held

commitments

- structure: `(agent action)`
- removed as a result of a belief change or a `UNREQUEST` message

capabilities

- structure: `(privateaction mntlcond)`
- fixed