Symbolic AI: History and Foundations

Emiliano Lorini

IRIT-CNRS, Université de Toulouse

What is AI?

"...AI (Artificial Intelligence) deals with some of the phenomena surrounding computers, hence is a part of computer science. It is also a part of psychology and cognitive science. It deals, in particular, with the phenomena that appear when computers perform tasks that, if performed by people, would be regarded as requiring intelligence and thinking" [H. Simon, 1993].

Symbolic vs sub-symbolic AI

- Symbolic models
  - Physical symbol hypothesis (Newell & Simon, 1976): intelligence can be captured by operations on symbols
  - "Language of thought" (Fodor, 1975)
  - Compositionality + local representations: logic
- Sub-symbolic models: artificial neural networks (ANNs) and reinforcement learning (RL) models
  - In ANNs information is encoded at the level of the connection weights
  - Representations in ANNs can be either local (concept => node) or distributed (concept => pattern of activation involving many nodes)
  - RL: implicit (non-explicit) expectation of a reward/punishment

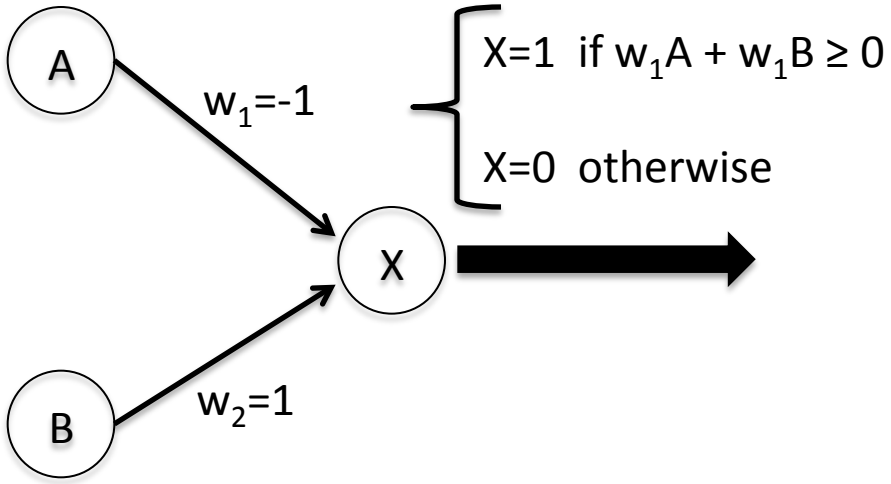
Symbolic vs sub-symbolic AI

Symbolic representation of implication

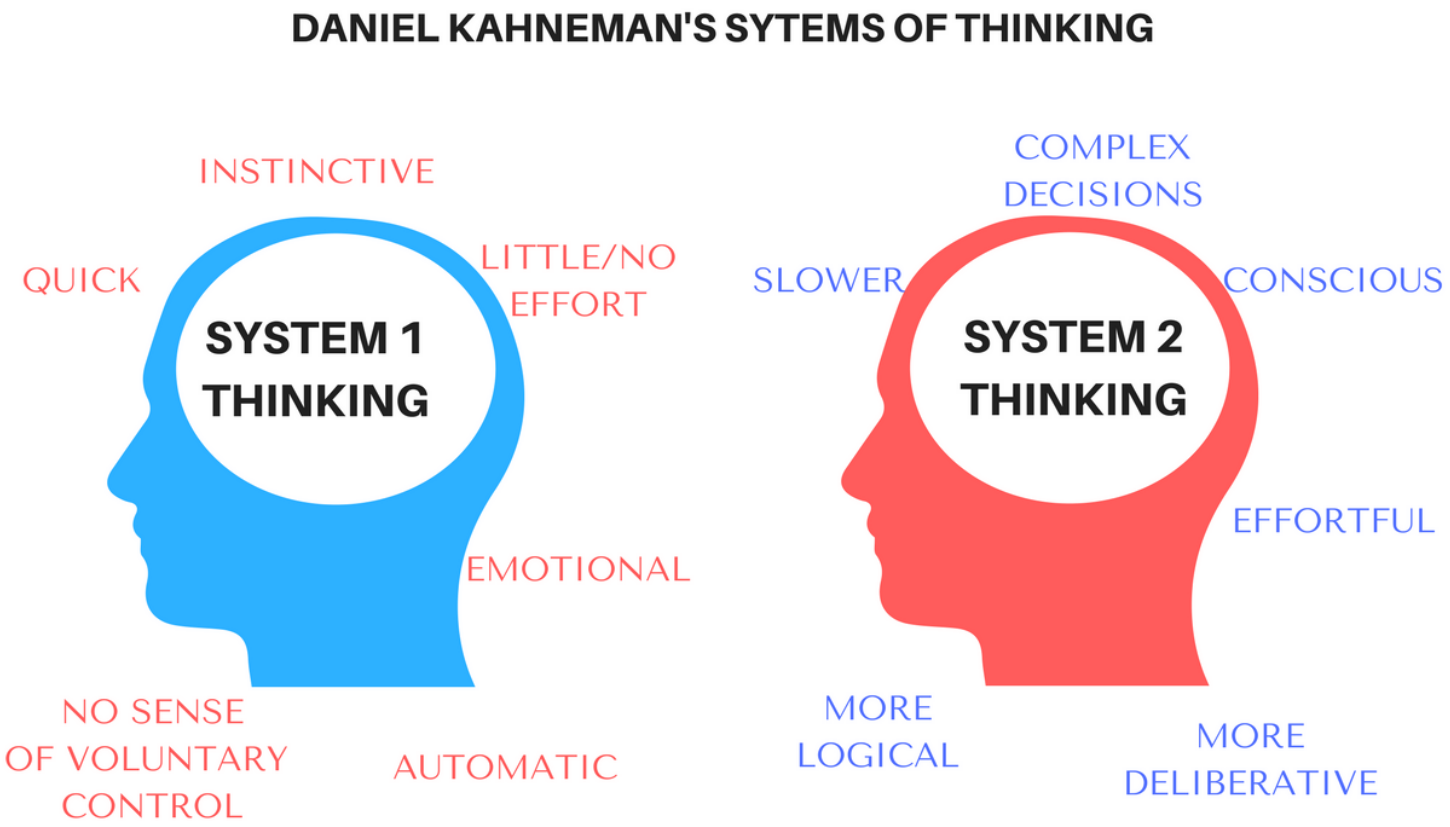
Sub-symbolic representation of implication (via perceptron)

A => B (*"A implies B"*)

A	B	A => B
0	0	1
0	1	1
1	0	0
1	1	1



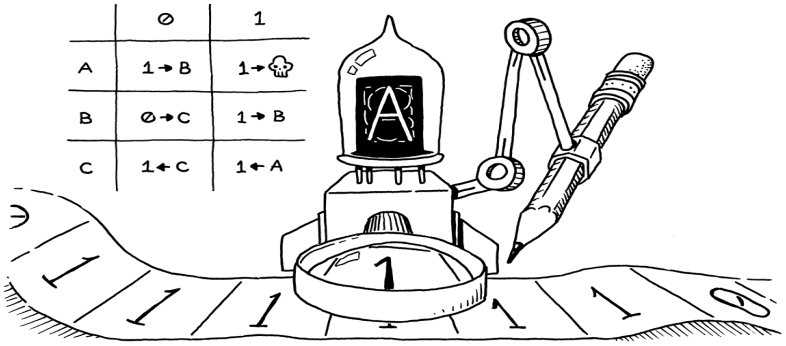
System 1 and system 2



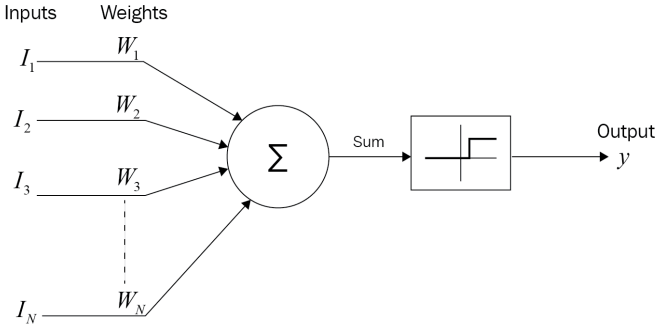
D. Kahneman (2003). A perspective on judgement and choice. American Psychologist,58:697-720.

Symbolic AI: the beginning

– Turing machine (Turing, 1936)



– McCulloch & Pitts (1943): first connectionist model



Symbolic AI: 60ies => 70ies

- "Programs with Common Sense" (McCarthy, 1959): first paper on logical AI
- Resolution for propositional logic (Robinson, 1965)
- Perceptron (Minsky & Papert, 1969): learning for single-layer ANNs
- Classical planning: STRIPS (Stanford Research Institute Problem Solver) by Fikes & Nilsson (1971)
- Structured knowledge for ontologies:
  - Semantic networks (Quillian, 1968)
  - Frames (Minsky, 1974)
  - Scripts (Schank & Abelson, 1977)
- Expert systems: DENDRAL, MYCIN,…

Symbolic AI: 80ies => 90ies

- Belief revision, non-monotonic reasoning, reasoning about action and change
  - Belief revision (Alchourròn et al., 1985)
  - Circumscription (McCarthy, 1980)
  - Default logic (Reiter, 1980)
  - Situation calculus (Reiter, 1991)
  - Autoepistemic logic (Moore, 1985)
  - Conditional logics (Kraus et al., 1990)
  - Stable model (answer set) semantics of logic programming (Gelfond & Lifschitz, 1988)
- Back-propagation algorithm (Rumelhart et al., 1986): learning algorithm for multi-layer feedforward ANNs
- Bayesian networks (Pearl, 1988)

Symbolic AI: 90ies => 2000s

- From single-agent reinforcement learning to multi-agent learning
  - Q-learning (Watkins, 1989)
  - Markov games (Littman, 1994)
- Logics for distributed AI and multi-agent systems: formal verification approach (Fagin et al., 1995)
- Cognitive agents
  - Qualitative decision theory (Boutilier, 1994)
  - BDI (belief, desire, intention) models and architectures (Bratman, Israel and Pollack, 1988)
  - CP-nets (Boutilier et al., 2004)
  - Affective computing (Picard, 1997)
- Argumentation (Dung, 1995)
- Description logic (Baader et al., 2003): modern view of ontology for semantic web applications

Symbolic AI: 2000s => today

- Social AI
  - Algorithmic game theory and computational social choice (Brandt et al., 2016): boolean and parity games, judgement aggregation
  - Machine ethics (Wallach & Allen, 2008): deontic logics, logics of responsibility and causality
  - Theory of mind (ToM) modeling (Albrecht et al., 2020): epistemic/ cognitive planning, recursive modeling
- Modern ANNs
  - Deep ANNs (Goodfellow et al., 2016)
  - Generative adversarial networks (GANs) (Goodfellow et al., 2014)
  - Spiking neural networks (Maass, 1997)
- Neuro-symbolic integration (Garcez el., 2009)

Formal methods for symbolic AI

- Logics in/for AI
  - Propositional logic and predicate logic
  - Logical approaches to belief revision, non-monotonic reasoning and abduction
  - Logics of uncertainty: logics of probability, possibilistic logic, real-valued logics
  - Modal logics: epistemic logic, logic of preference, deontic logic, BDI logics, logics for strategic reasoning, logic of agency,…
  - Description logics
- Compact representation of probabilistic and causal knowledge and preferences
  - Bayesian networks
  - CP-nets
  - Structural equation models

FORMAL METHODS FOR SYMBOLIC AI

Formal logic

Formal logic:

(1) Logical language, e.g.: $\forall x$ (Human(x) $\Rightarrow$ Mortal(x))

(2) Model theory ('semantics')

(3) Proof theory

(4) Provers or proof assistants

Advantages:

(1)+(2) ➔ Form and sense of propositions clear and well-defined (≠ natural language)

(3)+(4) ➔ Reasoning can be performed by a machine

Two basic logical systems:

Propositional logic

Predicate logic

Propositional logic: language

- Logical language
  - Atomic propositions ('facts'): itRains, itSnows, roadWet, lightOn,…
  - Boolean connectives:

    ¬itRains = "it does not rain" **negation**

    itRains ∧ itSnows = "it rains and it snows" **conjunction**

    ItRains ∨ itSnows = "it rains or it snows" **disjunction**

    itRains ⟹ ¬itSnows = "if it rains then it does not snow" **implication**

    itRains ⟺ itSnows = "it rains if and only if it snows" **equivalence**
  - Complex propositions

    (itRains ∨ ¬itRains) ⟺ TRUE

    ((itSnows ⟹ itRains) ∧ (itSnows ⟹ ¬itRains)) ⟹ ¬itSnows

Propositional logic: model theory

- Model theory
  - Interpretation = function **I** associating 0 ('true') or 1 ('false') to each atomic proposition
  - Associate truth values to complex formulas:

    $\mathbf{I}(\neg\varphi) = 1$ if $\mathbf{I}(\varphi) = 0$

    $\mathbf{I}(\varphi \wedge \psi) = 1$ if $\mathbf{I}(\varphi) = 1$ and $\mathbf{I}(\psi) = 1$

    $\mathbf{I}(\varphi \vee \psi) = 1$ if $\mathbf{I}(\varphi) = 1$ or $\mathbf{I}(\psi) = 1$

    $\mathbf{I}(\varphi \Rightarrow \psi) = 1$ if $\mathbf{I}(\varphi) = 0$ or $\mathbf{I}(\psi) = 1$
- Reasoning problems
  - Model checking: given **I** and φ, do we have $\mathbf{I}(\varphi) = 1$?
  - Validity: given φ, do we have $\mathbf{I}(\varphi) = 1$ for every **I**?
  - Satisfiability: given φ, do we have $\mathbf{I}(\varphi) = 1$ for some **I**?
  - Consequence: given φ and ψ, do we have that $\mathbf{I}(\varphi) = 1$ implies $\mathbf{I}(\psi) = 1$?

    notation : $\varphi \models \psi$

    φ is valid iff TRUE $\models \varphi$

Propositional logic: proof theory and provers

- Proof theory
  - Truth table
  - Sequent calculus
  - Resolution
- Mathematical properties
  - SAT problem: "is φ satisfiable?"
  - **Decidable**: there are algorithms which for every formula φ answer (in finite time) if φ has a model or not
  - **NP complete**: the algorithm requires time polynomial in the input if the 'right' non-deterministic choices are made
- Provers
  - More and more 'efficient', a lot of progress since ~1995
  - Recently integration of learning methods

Predicate logic: language

- Predicates:
  - No argument = atomic propositions ('facts', e.g.: itRains)
  - One argument: Human(x), Mortal(x) **'property'**
  - Two arguments: MarriedWith(Ann,Bob), EmployeeOf(Ann,CNRS) **'relation'**
  - Three arguments: Between(1,x,10)
  - …
- The arguments of predicates are *terms*:
  - Variables: x, y,…
  - Constants: cat, table, Ann, Bob…
  - Complex terms built with function symbols: age(EL),…
- Boolean connectives of propositional logic
- Variables and quantifiers:
  - $\forall x$ (Human(x) $\Rightarrow$ Mortal(x)) **'for all'**
  - $\exists y$ (Human(y) $\wedge \neg$Mortal(y)) **'there is'**

Predicate logic: model theory

- Domain D = set of objects ('individuals') under concern
- Interpretation = function **I** associating 0 ('true') or 1 ('false')
  to each atomic formula without variables
  - **I**(Male(Bob)) = 1, **I**(FriendOf(Bob,Father(John))) = 1
  - **I**(Male(Ann)) = 0, **I**(Female(Father(John))) = 0
- Allows us to associate a truth value to any formula

  I($\neg\varphi$) = 1 if …

  I($\varphi \wedge \psi$) = 1 if …

  I($\varphi \vee \psi$) = 1 if …

  I($\varphi \Rightarrow \psi$) = 1 if …

  I($\forall x\, \varphi(x)$) = 1 if *for every* object d in D, **I**(φ(d)) = 1

  I($\exists x\, \varphi(x)$) = 1 if *there is* an object d in D such that **I**(φ(d)) = 1
- Reasoning problems:
  - … (as before)
  - Querying: given φ and $\psi(x)$, for which x do we have $\varphi \models \psi(x)$?

Predicate logic: proof theory and provers

- Proof theory
  - Resolution (and variants)
- Mathematical properties
  - SAT problem: "is φ satisfiable?"
  - **Undecidable**: for every algorithm there are formulas for which computation loops
    - … more precisely, SAT is semi-decidable: if φ is unsatisfiable/valid then the algorithm answers "unsatisfiable"/"valid" in finite time; otherwise there is no guarantee for termination
- Provers
  - Well established since the 1980ies

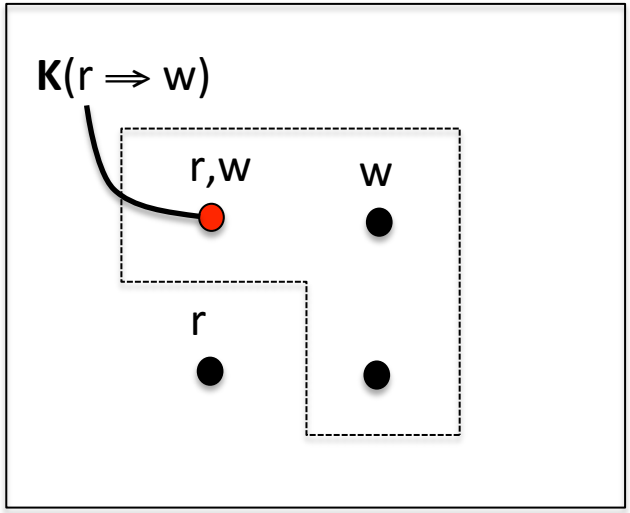
Modal logics: language

➔ Extensions of propositional logic with "modalities"

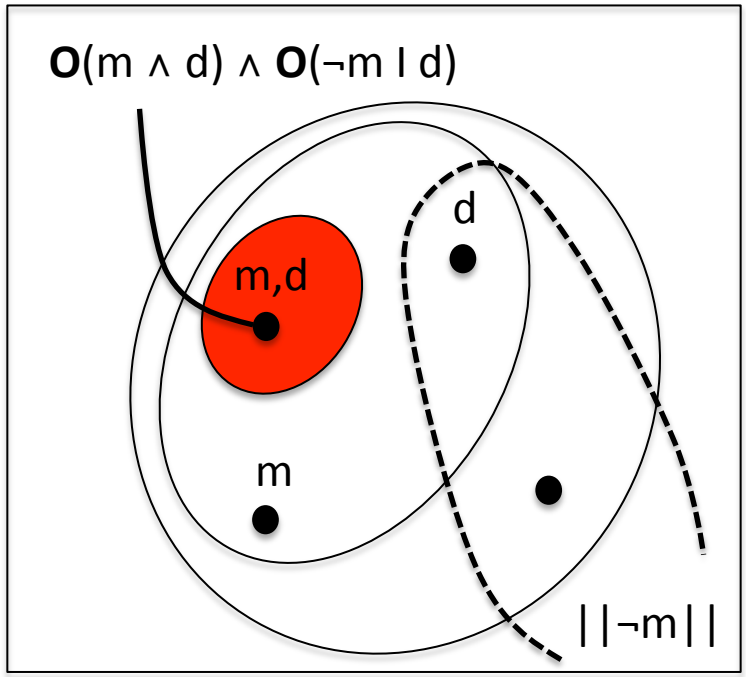
- Epistemic modalities: $\mathbf{K}_i\varphi$ (knowledge), $\mathbf{B}_i\varphi$ (belief), $\mathbf{A}_i\varphi$ (awareness), $\mathbf{T}_{i,j}\varphi$ (trust), $\mathbf{B}_i^k\varphi$ (graded belief)
- Volitional modalities: $\psi \leq_i \varphi$ (preference), $\mathbf{G}_i\varphi$ (goal), $\mathbf{I}_i\varphi$ (intention)
- Group modalities: $\mathbf{CB}_G\varphi$ (common belief), $\mathbf{DB}_G\varphi$ (distributed belief)
- Deontic modalities: $\mathbf{O}\varphi$ (obligation), $\mathbf{P}\varphi$ (permission), $\mathbf{O}(\varphi|\psi)$ (conditional obligation)
- Temporal modalities: $\mathbf{X}\varphi$ (next), $\mathbf{G}\varphi$ (always), $\varphi\mathbf{U}\psi$ (until)
- Action modalities: $\mathbf{STIT}_i\varphi$ ("seeing-to-it"), $\mathbf{CAN}_i\varphi$ (capability)
- Dynamic modalities: $[\varphi!]$ (public announcement), $[\varphi_G!]$ (private announcement)

Modal logics: semantics

r = it rains
w = the grass is wet
K: “the agent knows that…”



m = use mask
d = keep social distancing
O: “it ought to be the case that…”
O(…I…): “it ought to be the case that… under condition…”



Modal logics: complexity and theorem provers

- Rich variety of complexity results for satisfiability checking
  - NP: basic modal logic S5 (knowledge) and KD45 (belief)
  - PSPACE: multi-agent epistemic logic, linear temporal logic
  - EXPTIME: branching-time temporal logics (CTL), alternating-time temporal logics (ATL), propositional dynamic logic (PDL)
- Formal verification: use of compact semantics (e.g., reactive modules, belief bases)
- Provers: tableaux, sequent calculi, reduction to SAT and QBF
- Connection with description logics for ontologies

Logics of uncertainty

Probabilistic logic operators: **Prob**(φ) $\geq$ c and **Prob**($\varphi|\psi$) $\geq$ c

M $\models$ **Prob**(φ) $\geq$ c <==> $\pi(||\varphi||) = \Sigma_{s \in ||\varphi||} \pi(s) \geq c$

M $\models$ **Prob**($\varphi|\psi$) $\geq$ c <==> $M^{\varphi} \models$ **Prob**(ψ) $\geq$ c

with M^{φ} the model with updated probability distribution π^{φ}:

- $\pi^{\varphi}(s)=\pi(s)$ if $\pi(||\varphi||)=0$
- $\pi^{\varphi}(s)=0$ if $\pi(||\varphi||)>0$ and $s \in ||\neg\varphi||$
- $\pi^{\varphi}(s)=\pi(s)/\pi(||\varphi||)$ if $\pi(||\varphi||)>0$ and $s \in ||\varphi||$

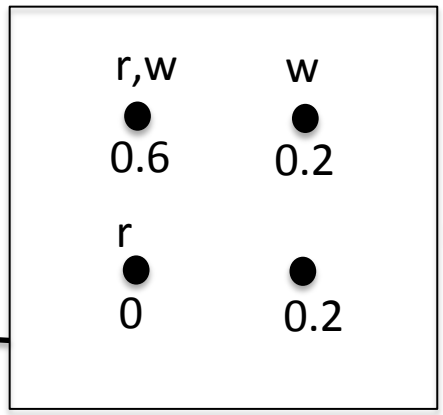
r = it rains

w = the grass is wet

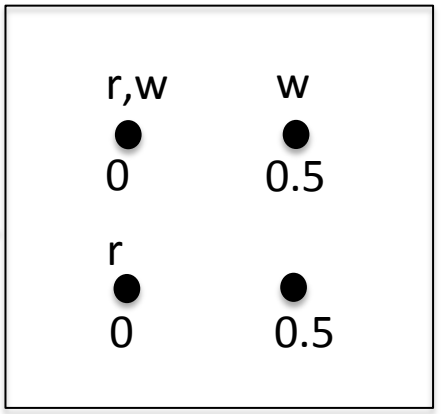
Prob(r) $\geq$ 0.6

Prob(r $\Rightarrow$ w) $\geq$ 1

Prob($\neg$r|w) $\geq$ 0.5



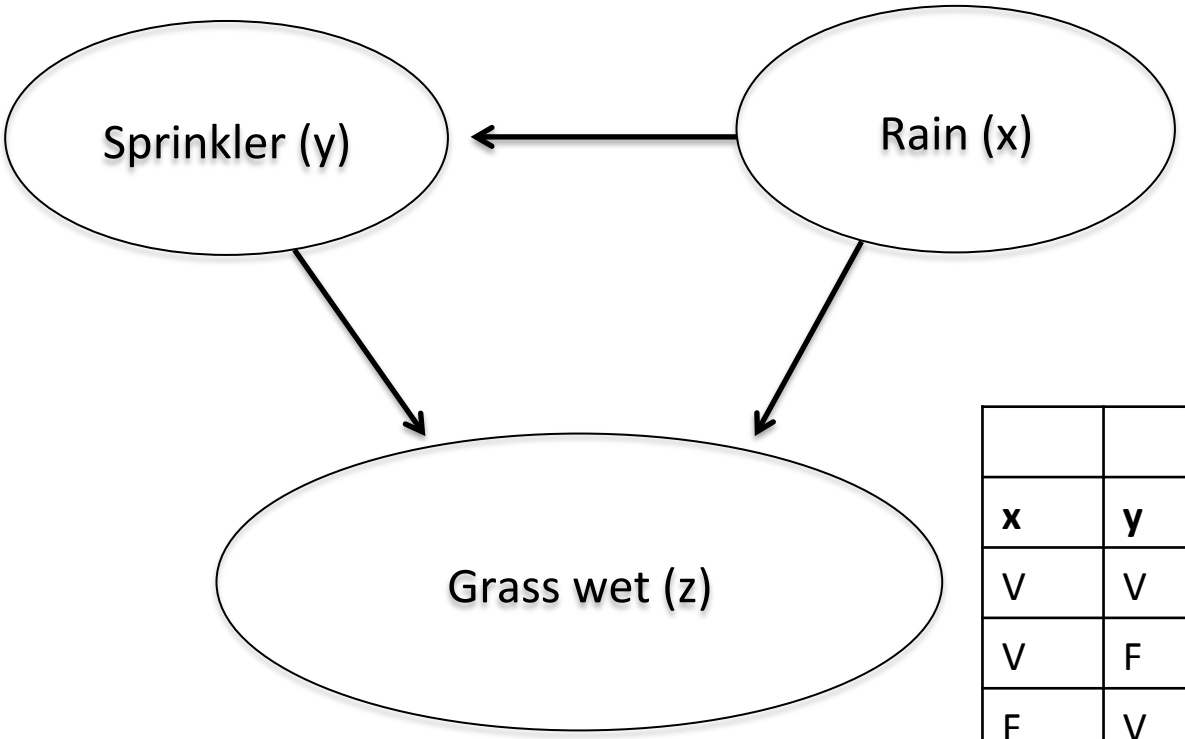
Update with $\neg$r



Bayesian network

Compact representation of probabilistic knowledge

	y	
x	V	F
V	0,01	0,99
F	0,4	0,6



x	
V	0,05
F	0,95

		z	
x	**y**	V	F
V	V	0,99	0,01
V	F	0,7	0,3
F	V	0,8	0,2
F	F	0	1,0

What is the probability that x is true when z is true P(x=V|z=V)?

Non-monotonic reasoning approaches

- In ‘real life’ knowledge is often by default

  $\forall x$ (Student(x) ~> Young(x)) (~> = “implies by default”)

  $\forall x$ ((Student(x) $\wedge$ RegisteredHabilitation(x)) ~> $\neg$Young(x))

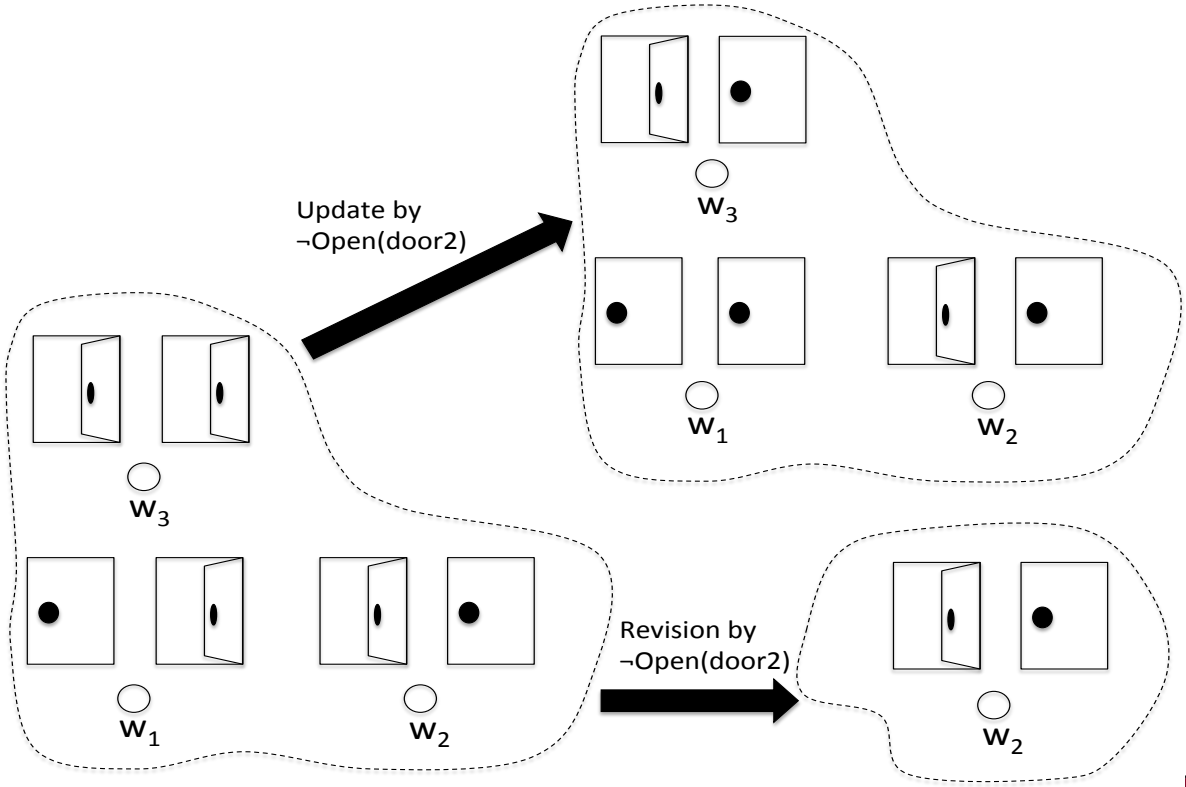
- Consequence relation becomes **nonmonotonic**:

  {Student(EL)} + DEFAULTS |= Young(EL)

  {Student(EL), RegisteredHabilitation(EL)} + DEFAULTS |= $\neg$Young(EL)

Belief revision vs update

- Belief revision (Alchourron et al., 1985)
  - World has changed (due to an action that took place)
  - Revision of belief base {Open(door1) v Open(door2)} by ¬Open(door2) |= Open(door1)
- Belief update (Katsuno & Mendelzon, 1992)
  - World didn't change (only beliefs changed)
  - Update of belief base {Open(door1) v Open(door2)} by ¬Open(door2)|≠ Open(door1)

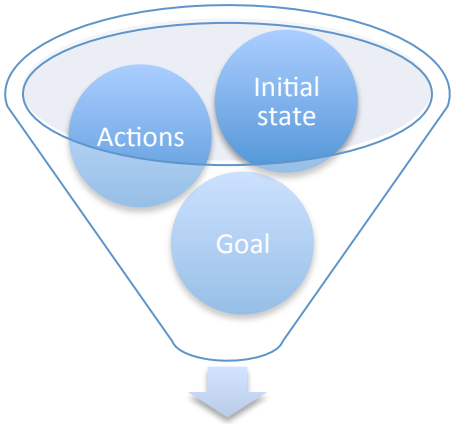


Belief revision vs update

- A belief revision operator should satisfy **rationality postulates**
  - revision of BB by φ is consistent
  - revision of BB by φ must contain φ (success)
  - revision of BB by φ is BB$\cup\{\varphi\}$ if φ is *consistent* with BB (persistence)
  - …
- Similar postulates for update
  - Main difference: update does not satisfy persistence
- Several operators revision are compatible with the rationality postulates:
  - Depends on extra-logical factors
  - in particular: epistemic importance ('entrenchment') of the different elements of BB (=> necessary for minimal change)

LOGIC-BASED MODELING IN AI

Planning approaches

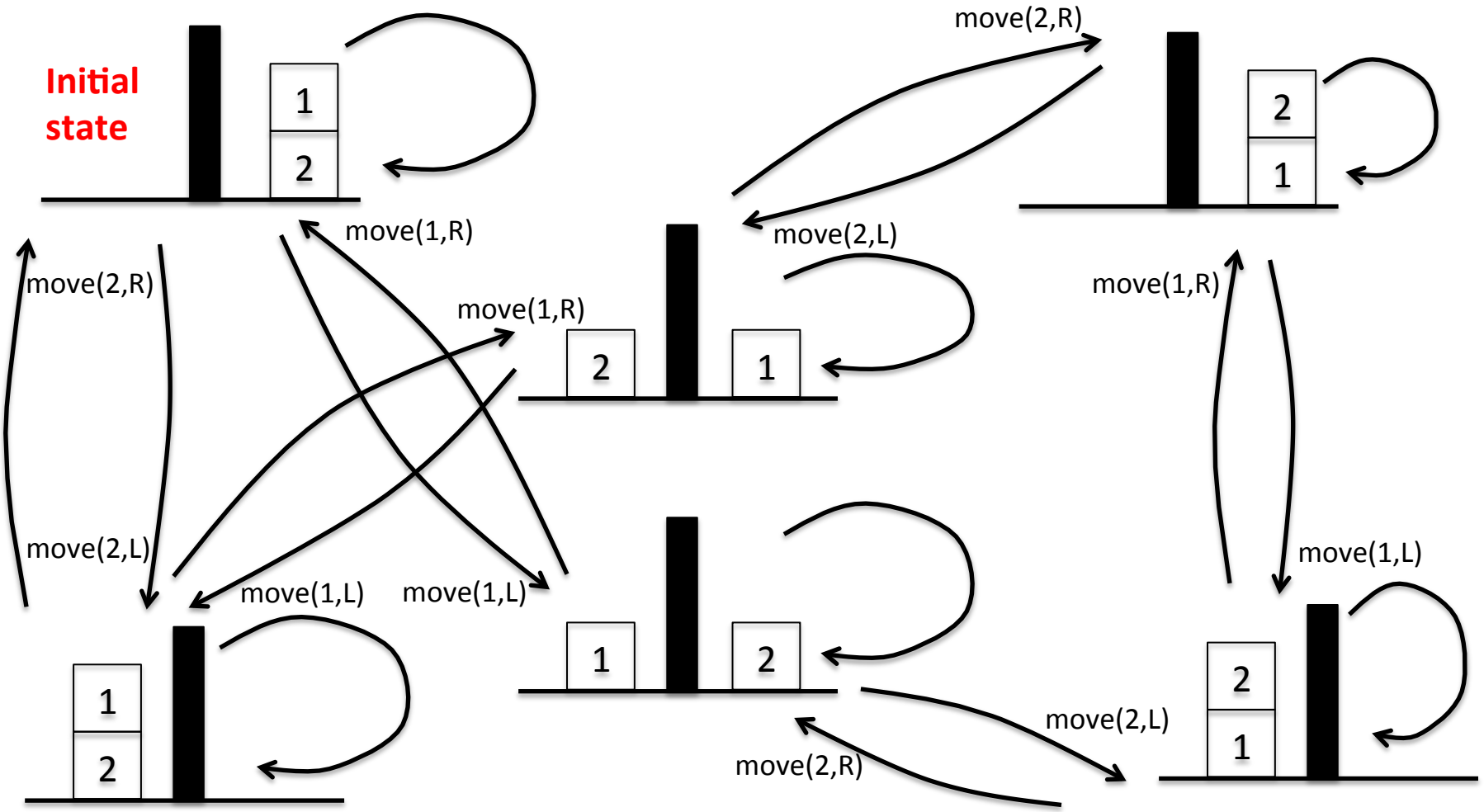


- Integration of different aspects
  - Multi-agent
  - Flexible
  - Temporal
  - Epistemic/ Cognitive
  - Contingent
- Algorithms
  - Hybrid approaches
  - Tractable classes
  - Relaxations
- Transformation into simpler versions
  - Solver (CSP, SAT, SMT, QBF)
  - Classical planning

Classical planning

- Planning problem $<S, s_{init}, \gamma, Act>$
  - Set of states S
  - Initial state s_{init}
  - Goal γ
  - Set of actions Act:
    - a : S -> S with $a \in Act$
- Solution: sequence of actions $a_1,\ldots,a_n \in Act$ such that $\exists\ s_1,\ldots,s_n \in S$ with:
  - $s_1 = s_{init}$
  - $a_1(s_1) = s_2$
  - $\forall\ 1 < k < n : a_k(s_k) = s_{k+1}$
  - $s_n \models \gamma$
- Plan existence: PSPACE-complete problem

Classical planning



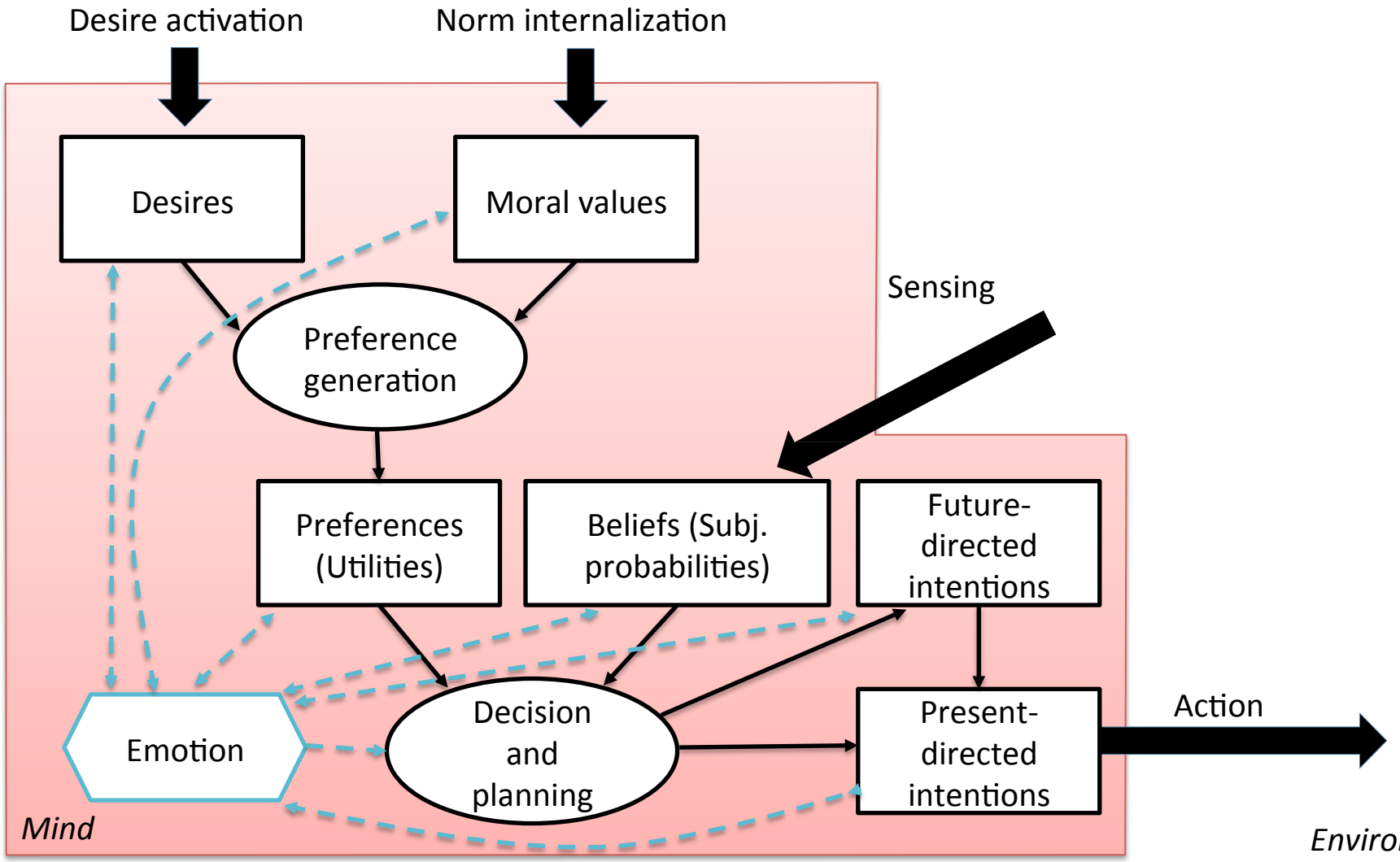
Goal: at(1,R) ∧ at(2,R) ∧ on(2,1)

Solution Plan: move(1,L),move(2,L),move(1,R)

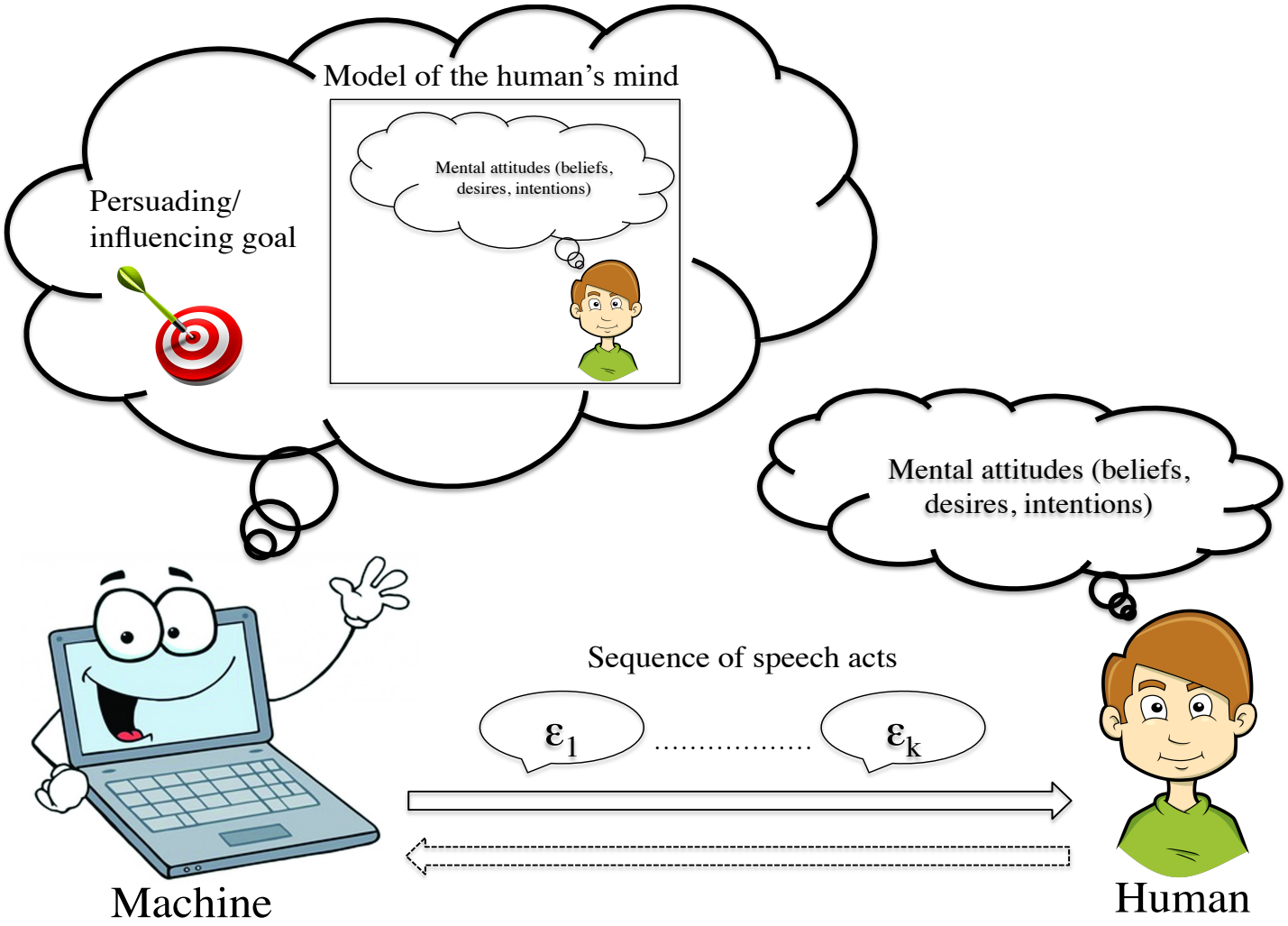
Cognitive agents

- Models of mental attitudes: folk (common sense) psychology and ‘‘intentional stance’’ (Dennett, 1987):
  - Beliefs, knowledge, uncertainty
  - Desires, goals, preferences, intentions
  - Values, norms, imperatives
- Functionalism
- Appraisal theories of emotion

Cognitive agent: architecture



Cognitive planning



ANR project CoPains 2019-2022

https://www.irit.fr/CoPains/

Social AI

- Logics for multi-agent systems
- Algorithmic game theory and computational social choice (Brandt et al., 2016)
  - Boolean games
  - Parity games
  - Judgement aggregation
  - Voting procedures and fair division algorithms
- Machine ethics
  - Deontic logic
  - Models of responsibility and causality
- Theory of mind (ToM) modeling

Boolean games

- Boolean game $\langle \mathbf{P}, \mathbf{N}, (\gamma_i)_{i \in Agt}, (C_i)_{i \in Agt} \rangle$
  - Set of propositional variables **P**
  - Set of agents **N** = {1,…,n}
  - Agent i's goal γ_i
  - Agent i's controlled variables $C_i \subseteq \mathbf{P}$:
    - $C_i \cap C_j = \emptyset$ if $i \neq j$
    - $C_1 \cup \ldots \cup C_n = \mathbf{P}$
- Agent i's strategy s_i: assignment for i's controlled variables
- Collective strategy $s=(s_1,\ldots,s_n)$
- Agent i's best response $BR(s_i,s_{-i})$:

  $\forall s_i' \in S_i$: if $(s_i',s_{-i}) \models \gamma_i$ then $(s_i,s_{-i}) \models \gamma_i$
- Nash equilibrium $Nash(s)$:

  $\forall i \in \mathbf{N}$: $BR(s_i,s_{-i})$

Boolean games

- $water_1$ = agent 1 waters the plant
- $water_2$ = agent 2 waters the plant
- plantDies $=_{def} (water_1 \wedge water_2) \vee (\neg water_1 \wedge \neg water_2)$
- $C_1 = \{water_1\}$
- $C_2 = \{water_2\}$
- $\gamma_1 = \gamma_2 = \neg$plantDies

Nash(1 => $water_1$, 2 => $\neg water_2$)

Nash(1 => $\neg water_1$, 2 => $water_2$)

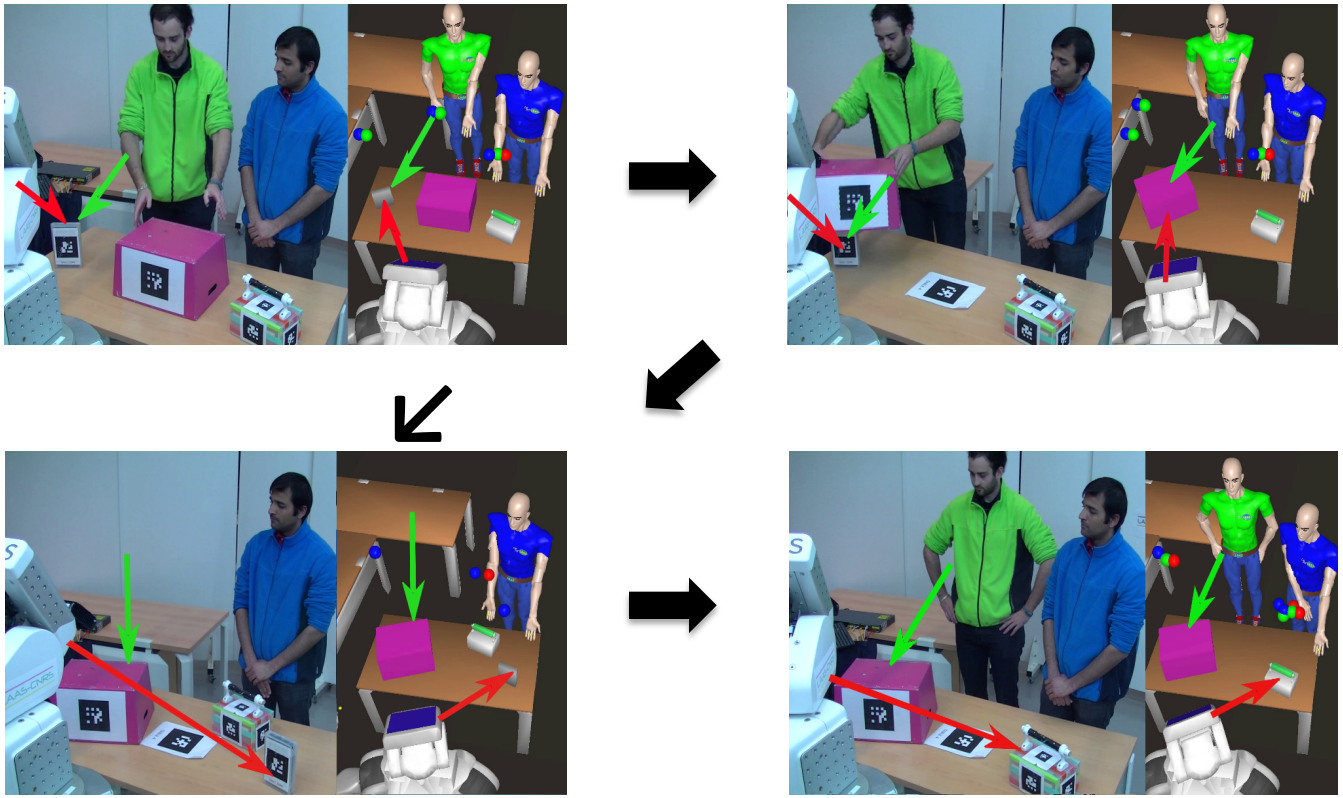
	$water_1$	$\neg water_1$
$water_2$	plantDies	$\neg$plantDies
$\neg water_2$	$\neg$plantDies	plantDies

Reasoning about others' beliefs

- Important: *false belief tasks* (Baron Cohen et al., 1985)
  - *Sally-Ann Task* : https://www.youtube.com/watch?v=jbL34F81Rz0
  - *Chocolate Task*
  - …
- Typically fail:
  - Less than 3 years old
  - Autistic subjects
- Hypothesis: specific human capacity of reasoning about others' beliefs
  - Mind reading
  - Theory of Mind, ToM

Challenge: social robots with theory of mind (Milliez et al., 2014)

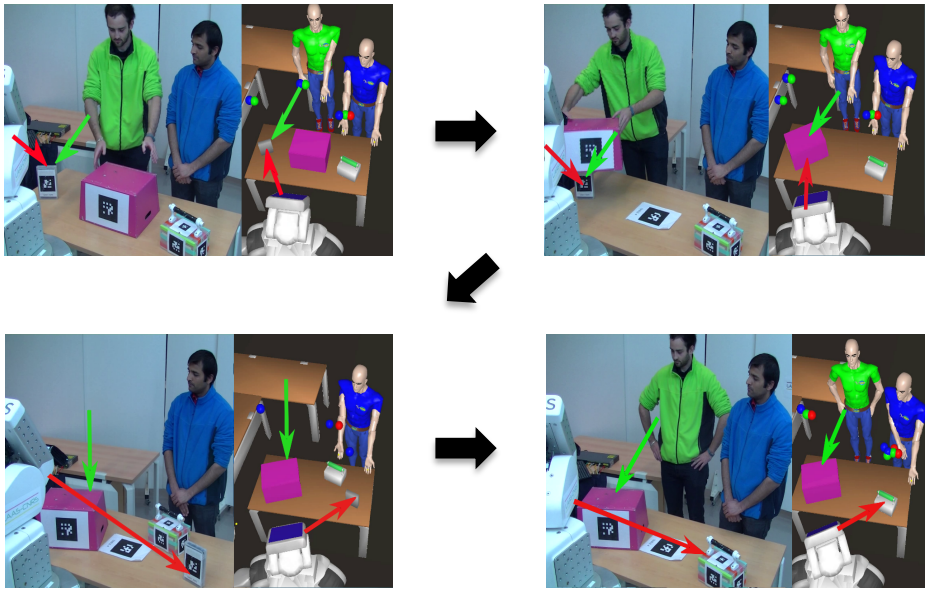
- At step 3 the beliefs of GREEN become false
- Coloured arrows = beliefs about the position of the book (robot = red)



Modeling theory of mind with epistemic logic

- Belief operator $\mathbf{B}_i$ = “agent i believes that”
- 3 possible epistemic attitudes w.r.t. a proposition φ :
  - $\mathbf{B}_i\varphi$ = “i believes that φ is true”
  - $\mathbf{B}\neg\varphi$ = “i believes that φ is false”
  - $\neg\mathbf{B}_i\varphi \wedge \neg\mathbf{B}_i\neg\varphi$ = “i has no opinion about φ”
- 6 possible epistemic situations w.r.t. a proposition φ :
  - $\mathbf{B}_i\varphi \wedge \varphi$ $\quad$ $\mathbf{B}_i\neg\varphi \wedge \varphi$ $\quad$ $\neg\mathbf{B}_i\varphi \wedge \neg\mathbf{B}_i\neg\varphi \wedge \varphi$
  - $\mathbf{B}_i\varphi \wedge \neg\varphi$ $\quad$ $\mathbf{B}_i\neg\varphi \wedge \neg\varphi$ $\quad$ $\neg\mathbf{B}_i\varphi \wedge \neg\mathbf{B}_i\neg\varphi \wedge \neg\varphi$
- Higher-order beliefs:
  - $\mathbf{B}_i(\varphi \wedge \mathbf{B}_i\varphi)$
  - $\mathbf{B}_i(\varphi \wedge \neg\mathbf{B}_i\varphi)$
  - $\mathbf{B}_i(\varphi \wedge \mathbf{B}_i\neg\varphi)$

Modeling theory of mind with epistemic logic



- R: robot (red)
- H: human (green)
- Under(L): object under box on the left
- Under(R): object under box on the right

<u>Initial situation</u>

$\mathbf{B}_R(\text{Under(L)} \wedge \mathbf{B}_H\text{Under(L)} \wedge \mathbf{B}_H\mathbf{B}_R\text{Under(L)})$

<u>At Step 4</u>

$\mathbf{B}_R(\text{Under(R)} \wedge \mathbf{B}_H\text{Under(L)} \wedge \mathbf{B}_H\mathbf{B}_R\text{Under(L)})$

INTEGRATION OF SYMBOLIC AND SUBSYMBOLIC APPROACHES

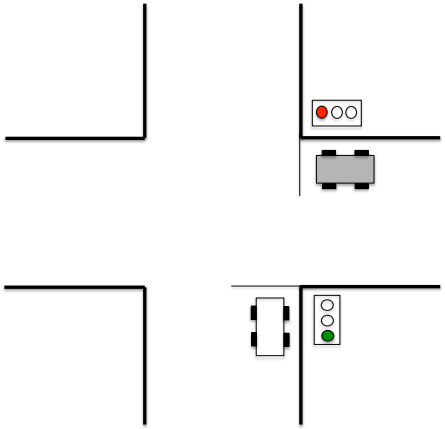
Two types of integration

- Merging reasoning and learning: deduction + induction
- Neuro-symbolic integration: integrating logic-based methods with ANN-based methods

Merging reasoning and learning

- Objective: developing AI systems with both learning and reasoning capabilities
- Examples of application
  - Artificial players for games under incomplete information (e.g., Libratus for poker)
  - Autonomous vehicles and robots
- Theories and tools
  - Multi-agent learning (e.g., Markov games, RL, fictitious play)
  - Temporal logics (e.g., LTL, CTL*) and epistemic logics
    - Automatic verification of stability properties via model checking
    - Automatic verification of convergence to equilibrium

Merging reasoning and learning



	S	C
S	0,0	0,1
C	-5,0	-15,-10

→

	S	C
S	0,0	0,1
C	-5,0	-15,-10

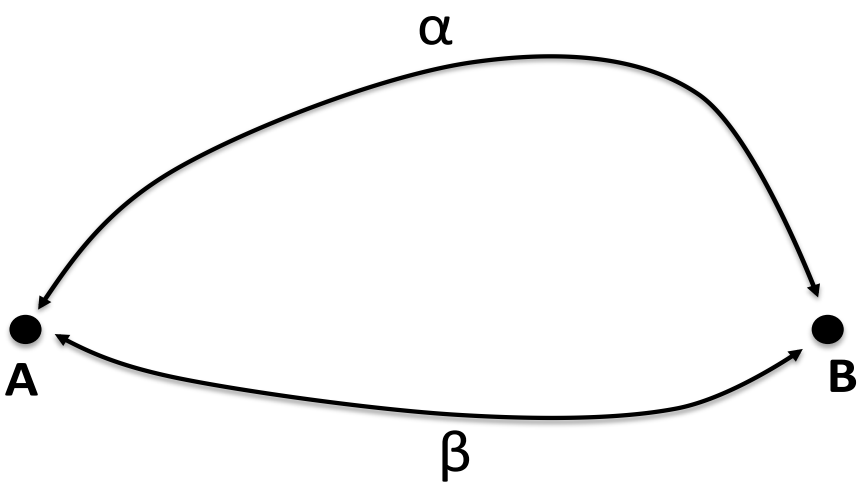
→

	S	C
S	0,0	0,1
C	-5,0	-15,-10

IDSDS solution of the game necessarily played under:

- Common knowledge of rationality
- Standard deductive capabilities of the players

Merging reasoning and learning



Routing scenario

	α	β
α	-1,-1	1,2
β	2,1	0,0

Routing game

- Nash equilibria: (α,β), (β,α) and ((0.25,0.75),(0.25,0.75))
- Convergence to equilibrium under fictitious play learning

Neuro-symbolic integration

- Knowledge extraction from an ANN
  - Formal verification
  - Integration of learning (via the ANN) and reasoning (via the extracted logical rules and information)
  - Explainable AI
- Symbol grounding problem: from perceptual raw data to abstract (logical) representations
  - Integration of object recognition and reasoning

Neuro-symbolic integration

Logic program

Atomic propositions

r = rain

w = wet

b = temperature below 0°C

i = ice

Set of rules

Π={r, b,

r → w,

w ∧ b → i}

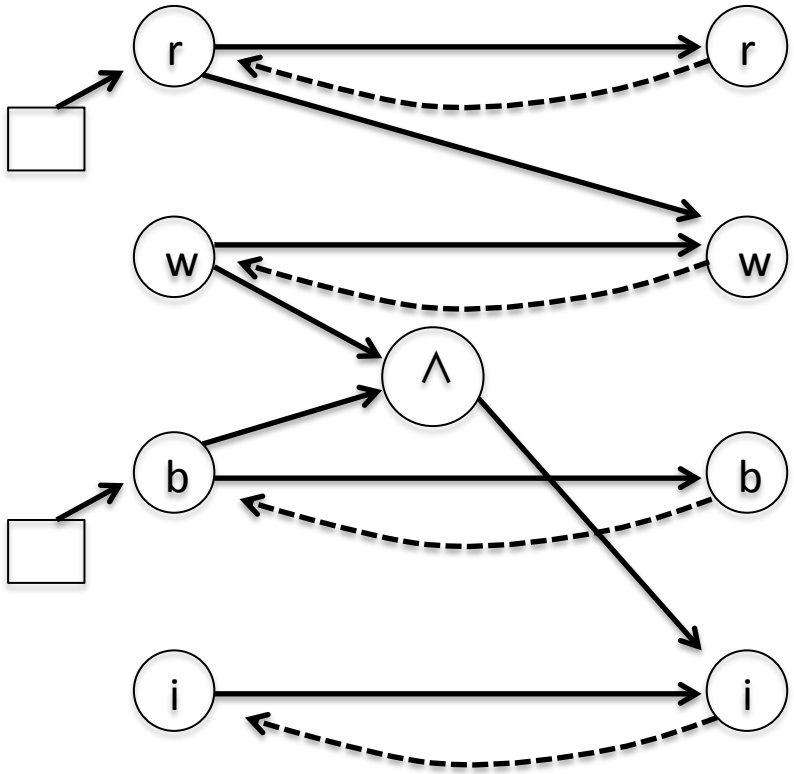
Inference process

$T_0(\Pi)=\{r,b\}$

$T_1(\Pi)=\{r,b,w\}$

$T_2(\Pi)=\{r,b,w,i\}$

Corresponding recurrent ANN



Merci pour votre attention !