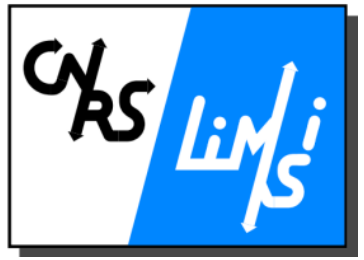
A logical model of Theory of Mind for virtual agents in the context of job interview simulation

Marwen BELKAID

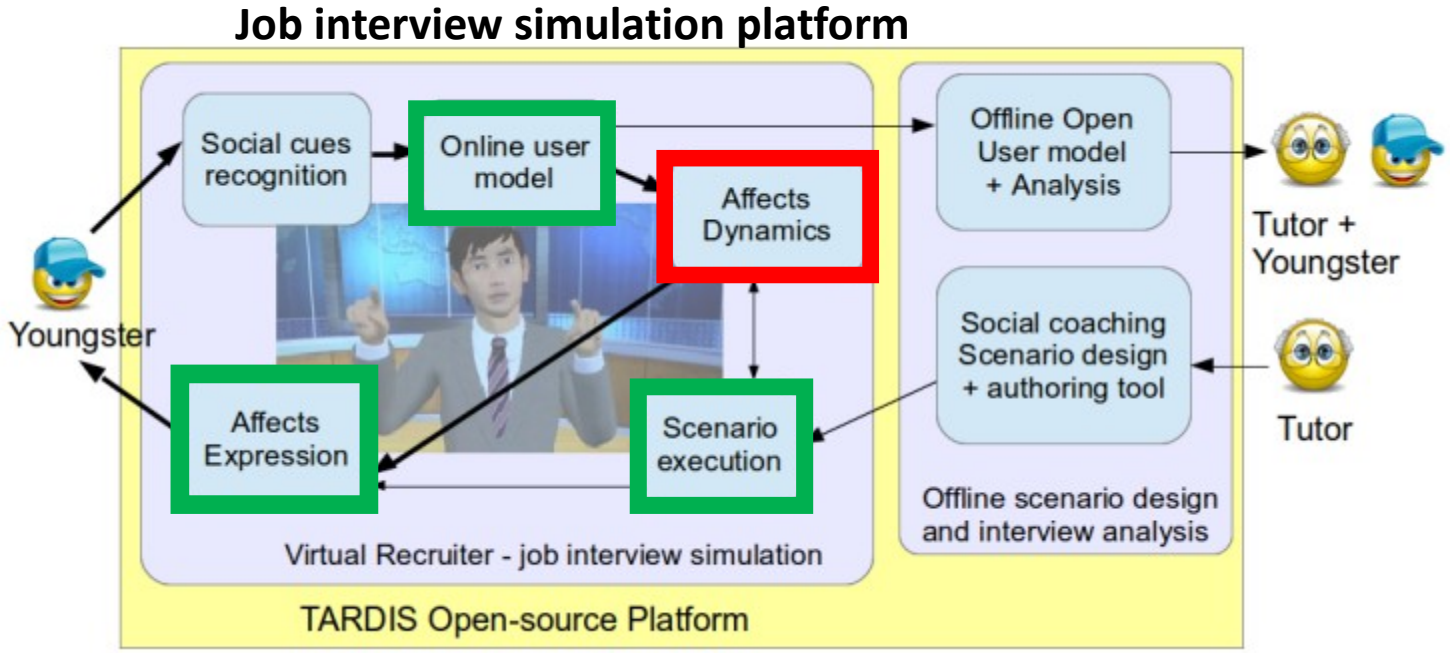
Nicolas SABOURET

(LIMSI-CNRS – Université Paris-Sud)

IDGEI 2014 – Haifa – Israel



The TARDIS project



- Affective Reasoner

  building the virtual recruiter's mental state

- Linked to :
  - *Social Cues Interpretation* : detected mental states
  - *Scenario* : context, situation of the dialogue
  - *Affect expression* : based on simulated mental states

Theory of Mind (ToM)

- What is ToM ?

  Representing and reasoning about another's mental state, beliefs, goals...

- Why is it important in job interviews
  - Identifying the interlocutor's personality, competencies, preferences, expectations...
  - TARDIS : virtual recruiter must show affects that reflect its interpretation of the applicant's attitude
- Our goal :
  - Reason about the interlocutor's beliefs, goals and affects
  - Reason about the interpretation of the interlocutor's of the interaction situation
    - Is he/she at ease ? Why ?
    - How can I influence this ?

Theory of Mind (ToM) - continued

- Two major theories
  - Theory-Theory : folk-psychology, commonsense reasoning
  - Simulation-Theory : mirroring interlocutor's mental state in our own model

  → toward hybrid models (neurobiology)
- A BDI-based model
  - TT : describe mental states as beliefs, goals, etc.
  - TT : design commonsense reasoning rules for affect appraisal (OCC : evaluation of the situation)
  - ST : apply the same inference engine to attributed mental states
  - Interactions : actions are speech acts

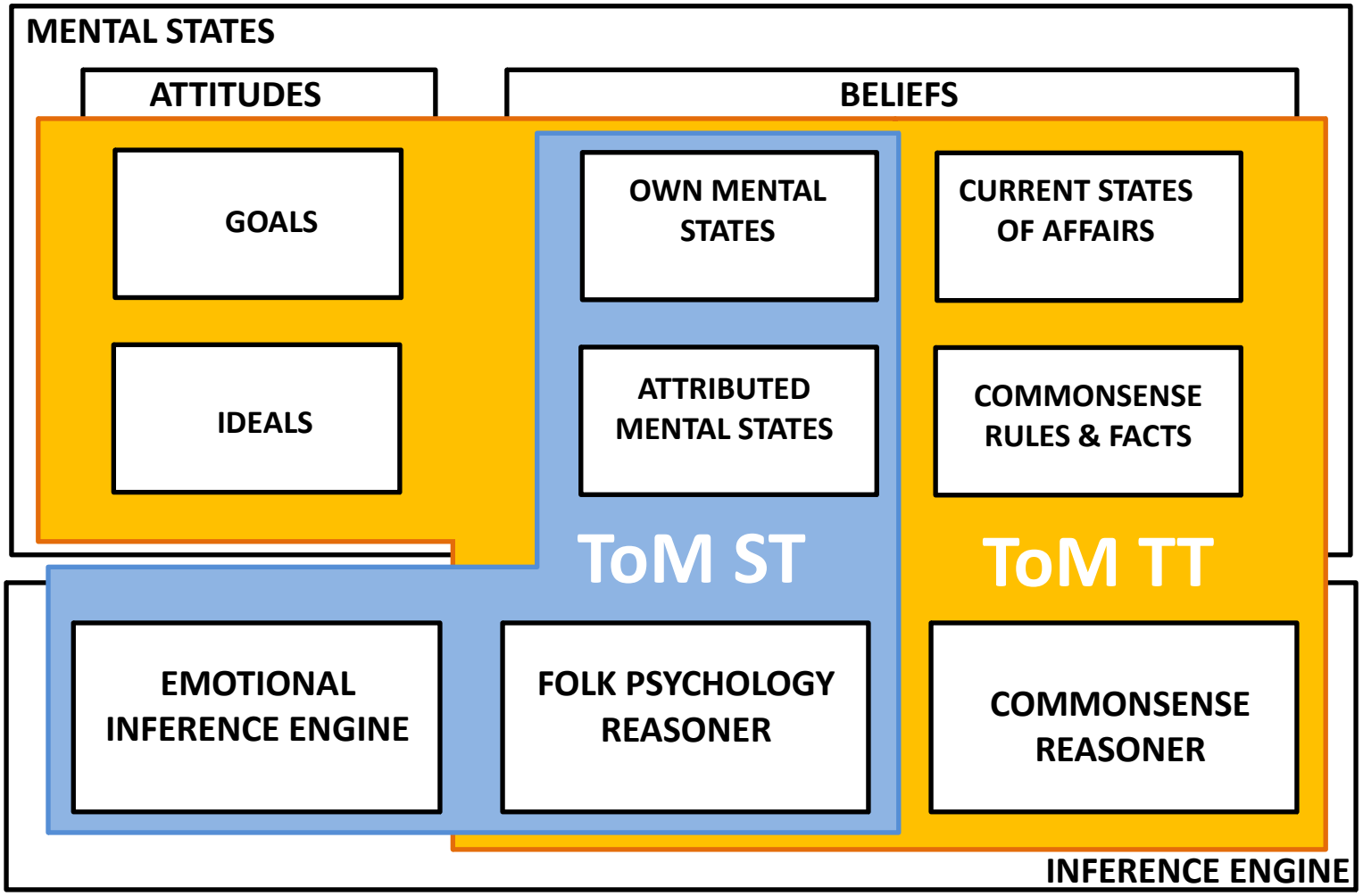
  → toward social influence

Overview

- Introduction
  - About ToM in TARDIS
- Theoretical model
  - Architecture
  - Logical model
- Implementation
- Evaluation
  - Methodology
  - First results
  - Analysis & Discussion

Theoretical model

General Architecture: hybrid TT & ST ToM



Logical model (1)

- Agent x Patent x Propositions/Actions → Event
  - <Candidate, − , sit_down>
  - <Recruter, Candidate, Assert(cv_is_good)>
- Speech acts : Assert, Request, Commit, Express
- Temporal modalities : Next, Future, Globally, Until
- BDI-like modalities : Beliefs, Attitudes & Goals
  - $Bel_{Recruiter}$ (cv_is_good)
  - $Bel_{Recruiter}$ (<Recruter , Candidate, Assert(cv_is_good)>

    → Future(Candidate_feels_at_ease))

Logical model (2)

- Graded beliefs & attitudes

$$Bel_a^l(\varphi) \overset{\text{def}}{\Longrightarrow} Bel_a^{1-l}(\neg\varphi)$$

- Desires and ideals

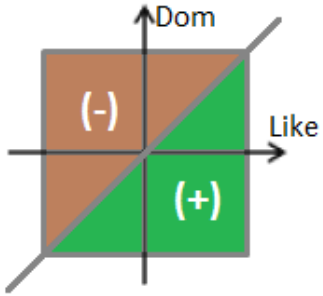
Ex :

$$Des_a^k(\varphi) \overset{\text{def}}{=} Att_a^k(F(\varphi))$$

$$Ideal_a^{k>0}(\varphi) \overset{\text{def}}{=} Att_a^{k>0}(G(\varphi)) = Des_a^{-k<0}(\neg\varphi)$$

Logical model (3)

- Social relations based on liking and dominance (Leary)



- Affect dynamics :

$$Bel_a^l(\gamma) \wedge Ideal_a^k(\neg\gamma) \wedge Bel_a^{l'}(Rsp_a(\gamma)) \overset{\text{def}}{\Longrightarrow} N(Shame_a^{i=f(l,l',k)}(\gamma))$$

$$Bel_a^d(\gamma) \wedge Bel_a^l(Att_b^{k>0}(\gamma)) \wedge Like_{a,b}^{k'>0} \overset{\text{def}}{\Longrightarrow} N(HappyFor_{a,b}^{i=f(l,k,k',d)}(\gamma))$$

- Impact of speech acts

  (what imply a request, ...)

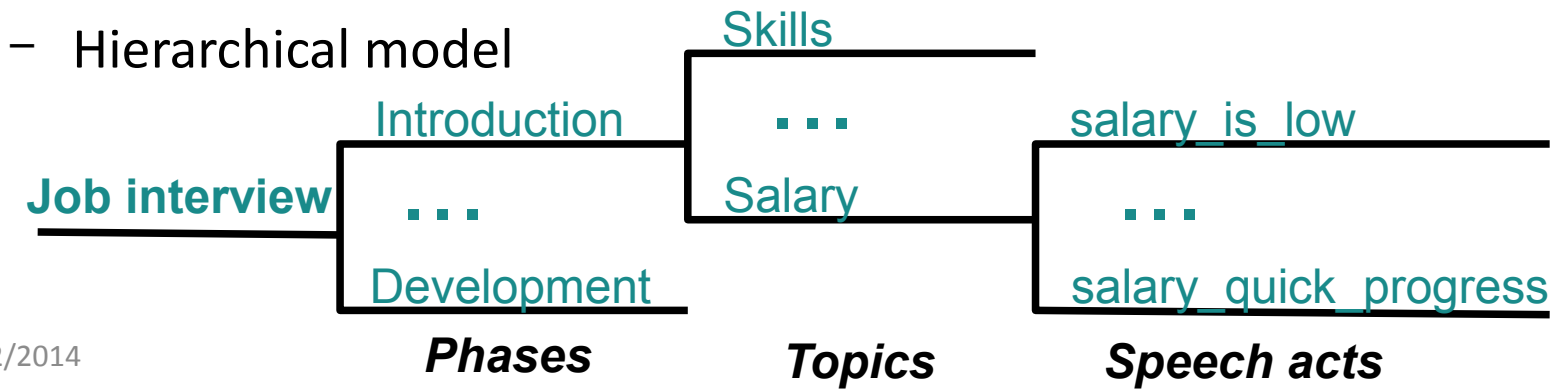
- Social influence (empathy, submission...)

$$Request_{b,a}(\varphi) \wedge Dom_{a,b}^{k<0} \overset{\text{def}}{\Longrightarrow} N(Int_a(\varphi)))$$

IMPLEMENTATION

General architecture

- SWI-prolog integrated in a C++ program
  - Reasoning model & inference engine in Prolog
  - Dynamics (BDI interpreter + time simulation) in C++
- Dynamics of graded modalities
  - Emotions
    - linear influence of attitude
    - logarithmic influence of degree of belief
- Job interview simulation (TARDIS)
  - Hierarchical model



EVALUATION

Method

- Simple text-based GUI
  - Slide bars to figure and to express affects
  - Not integrated in the TARDIS simulator
- Protocol
  - 30 participants
  - 3 agent's profiles: comprehensive, neutral, aggressive
  - Comparison with static rule-based system
  - Questionnaires: difficulty, credibility and pleasantness
- Results
  - Main result : impact of agent's profile on credibility
  - No impact of ToM vs rule-based using the simplified GUI

    → need for a real virtual agent and interview situation
  - Correlation between credibility and pleasantness

Conclusions & perspectives

- Conclusions

  Study how a Theory of Mind model could be used to enhance digital inclusion training tools

  - Theoretical model + hybrid TT&ST model
  - Application to job interview simulation can be generalised to other contexts

- Perspectives
  - Evaluation with a virtual agent in TARDIS
  - Definition of a valid experimental protocol

    How can user evaluate the quality of their interlocutor's ToM in an interaction context ?
  - Additional concepts in the model
    - Long term affects (moods, personality...)
    - Social manipulation (higher-level ToM)
    - Additional dimensions in social relations
    - Learning & adaptation

Thank you !