Computing Social Behaviours Using Agent Models

Paolo Felli, Tim Miller, Christian Muise, Adrian R. Pearce, Liz Sonenberg
Department of Computing and Information Systems, University of Melbourne
{paolo.felli,tmiller,christian.muise,adrianrp,l.sonenberg}@unimelb.edu.au

Abstract

Agents can be thought of as following a social behaviour, depending on the context in which they are interacting. We devise a computationally grounded mechanism to represent and reason about others in social terms, reflecting the local perspective of an agent (first-person view), to support both stereotypical and empathetic reasoning. We use a hierarchy of agent models to discriminate which behaviours of others are plausible, and decide which behaviour for ourselves is socially acceptable, i.e. conforms to the social context. To this aim, we investigate the implications of considering agents capable of various degrees of theory of mind, and discuss a scenario showing how this affects behaviour.

1 Introduction

Performance in many complex systems depends on the coordinated activity of a team of individuals, often characterised by complex decision tasks in rapidly evolving environment, where explicit communication may be limited. Research in Team Performance suggests that human teams coordinate activities more effectively and achieve better overall task performance when team members manage to track each other's beliefs, intentions and task-related states [Klimoski and Mohammed, 1994; Mohammed *et al.*, 2010; Espevik *et al.*, 2011]. This intuition was captured by the notion of shared Mental Models and team Mental Models [Johnson-Laird, 1983; Cannon-Bowers *et al.*, 1993; Bolstad and Endsley, 1999]. The idea is that information is organised in structured patterns and processed in a rapid and flexible manner, to describe, explain and predict the system behaviour as well as the ramifications of potential decisions prior to action.

Moreover, various degrees of rich interactions depend not only on such ability to represent mental models (to attribute beliefs, desires, pretending, etc., to oneself and others), *but also* to understand that others have mental states that are different from one's own. This is often called *theory of mind* (ToM) [Premack and Woodruff, 1978; Scassellati, 2002; Isaac and Bridewell, 2014], and it is a widely studied phenomenon in many disciplines [Bergwerff *et al.*, 2014; Ficici and Pfeffer, 2008]. The ability to take the perspective of others is crucial, and studies of human-robot interaction and social robotics identify the need for a human-oriented perception [Lemaignan *et al.*, 2010; Warnier *et al.*, 2012].

Also in the context of agent systems, we have seen in recent years a growing demand of more realistic social behavior [Kaminka, 2013; Dignum *et al.*, 2014], so one critical feature becomes the ability to represent others in social terms. Giving agents an awareness of their social reality will enable more seamless interdependent collective behaviour [Dignum *et al.*, 2014; Johnson *et al.*, 2014], where interdependency informally means that one agent's deliberation is dependent on what another agent does (or intends to do), and vice-versa. We therefore need to investigate computational structures that allow agents to reason not just about themselves, but also about the so-called *social reality* [Dignum *et al.*, 2014].

There has been considerable work on the design of intelligent agents and reasoning about their own knowledge and belief as well as that of others – e.g., [Levesque, 1984; Lakemeyer, 1986; Fagin *et al.*, 1995; Wooldridge and Lomuscio, 2001; Ditmarsch *et al.*, 2007] – typically for devising some form of strategy, or plan, to achieve goals. In real-life scenarios, however, agents must deal with a high degree of uncertainty, and although humans routinely interact successfully in limited cue conditions, this process requires complex, often multimodal, exchange of information. A critical perspective has thus to be assumed, one that can discern what is *plausible* from what is not. Existing work, e.g. [Bulling and Jamroga, 2007; Andersen *et al.*, 2014], assumes that plausible traces are given as part of the model, rather than constructed.

The contribution of this paper is:

- a representation of the *model* that agents have of each other, and their nested beliefs. An agent can use its own model for itself, yet use different representations and inference mechanisms for others. It can simulate others to deliberate, empowering interdependence and awareness.
- a computational mechanism to use these representations to discriminate which behaviours of others are plausible, given the context, and decide which behaviour for ourselves is socially acceptable (conforms to the context).

Crucially, we preserve an agent's *local perspective* (first-person view), instead of considering an omniscient observer (third-person view). We support two types of reasoning about others: *stereotypical reasoning*, using simple social rules, and *empathetic reasoning*, in which the agent casts itself into the

mind of another agent and reasons as if it were them.

In Section 2 we present a running scenario. In Section 3, we introduce our notion of agent models, and in Section 4 we model our scenario in this framework. In Section 5 we formally define the two types of reasoning above, and in Section 6 we give our definition of acceptable behaviour, given a social context. In Section 7 we comment on future work.

2 The Wumpus Quest

In recent work (e.g., [Bergwerff *et al.*, 2014; Ficici and Pfeffer, 2008]) authors studied agents with the cognitive ability to use of the ToM, in (possibly iterated) one-step games. Here, we present a scenario attempting to bring together some strategic and social features, inspired by the Wumpus Hunt:

The lord of a castle is informed by a peasant that a Wumpus is dwelling in a dungeon nearby. It is known that the Wumpus can be killed by one hunter alone only if asleep; if awake, two hunters are required. The lord then tasks the peasant to go to fetch the White Knight, his loyal champion, and hunt down the beast together. The White Knight is known for being irreprehensible, trustworthy and brave; however, the peasant does not know any knight, and neither their looks. While looking for the White Knight, he runs into the Black Knight and, believing him the White Knight, tells him about the quest.

There is some additional information that needs to be taken into account: on one hand, the knight knows how a Wumpus can be killed by two hunters, but he is aware that a simple peasant may get scared by the thought of confronting an awake Wumpus. Also, the peasant can not hunt and is unable to see whether the Wumpus is awake (he can not approach unnoticed), but the knight can. Therefore it is not clear to him whether the peasant can be of any help to the quest. On the other hand, the knight is aware of the misunderstanding: he knows that the peasant attributes to him all the good qualities of the White Knight, so the peasant is confident that the knight won't put him in danger whenever possible.

The implicit and explicit information of this scenario does not allow a unique understanding of the context, and it is not clear how each agent can use such information.

While on the road, they agree on a protocol: they will enter the dungeon from two sides, and the Knight will use a whistle to signal whether the Wumpus is awake, then they will attack.

3 Agent models

To allow one agent to reason about others in a social context, we provide agents with *agent (mental) models*. An agent is able to *assign* such models to others *and itself* (from some fixed collection), so when considering all possible eventualities, it is capable of determining its behaviour based on plausible estimates of others' behaviour. These models can be used in orthogonal ways: they can describe either specific agents, or agents of which the role, in the present context, is more characterising than their intimate understanding. This is the case, for example, of a bank clerk or policeman, who can be modeled as members of a reference group (*role* or *archetype*) [Dignum *et al.*, 2014]. This latter representation is akin to the stereotypical reasoning of humans, who do not necessarily engage in deep cognitive thinking about others, but rely on habits and social practices [Brooks, 1991; Dignum *et al.*, 2014]. Manipulations of these models and stereotypes enable shortcuts to be taken [Pfau *et al.*, 2014].

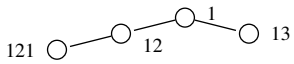

Figure 1: Example of set $\mathcal{A}$ for concrete labels $Ag = \{1, 2, 3\}$

Models are therefore *partial* and *task-specific*: a complete description of an individual, even when possible, would force us to consider numerous irrelevant details [Mccarthy, 1992]. Instead, models can be specialised to characterise different levels of approximation, depending on the context. Although we do not address a specific application domain for our framework, we nonetheless identify its applicability to human-agent teamwork, in which the social context can be considered to design agents capable of using, or simulating, human-like bounded reasoning. For example, in domains where plausible estimates of others come from training (e.g. firefighters), previous interactions (e.g. cognitive assistants for pilots), and social practices (e.g. sports events).

Consider a set of *concrete* agent labels $Ag = \{1, \ldots, n\}$. Let $\mathcal{A}$ be the set of *virtual* agent labels, obtained by concatenating labels in Ag, such that the first is unique. For instance, if $Ag = \{1, 2, 3\}$, then $\mathcal{A}$ can be a subset of $\{1, 11, 12, 13, 111, \ldots\}$, as informally represented in Figure 1. We will use these indices to refer to the representation that each agent has of others. E.g., agent 121 denotes the representation, according to agent 1, that 2 has of 1 itself. For simplicity, we use a tree representation, and will make use of a tree terminology. We will informally refer to the agent at the root as "reasoning agent".

Given $i, j \in \mathcal{A}$, we write $j \in ch(i)$ iff $j = i \cdot Ag$ (j is a child of i), and $i = pa(j)$ denotes that i is the parent of j. We regard concatenated labels as regular agent labels, i.e., we refer to $\mathcal{A}$ instead of Ag, and we assume that $\mathcal{A}$ is prefix-closed (i.e., if $j \in \mathcal{A}$ then any ancestor is in $\mathcal{A}$ as well). For simplicity, but without loss of generality, we exclude introspective agents, e.g. 11. Finally, $l(i)$ denotes the last *concrete* agent label of i, e.g., $l(123) = 3$, and $l(121) = 1$.

Intuitively, an agent model is a finite-state machine, whose states are called configurations, and transitions are labelled with observation symbols. This kind of modelling is in line with much of the literature on knowledge representation for multi-agent systems [Fagin *et al.*, 1995; Parikh and Ramanujam, 1985], in which the notions of computational *local* states are given prominence. All the information an agent has at its disposal (facts, observations, etc.) is captured in the state relating to the agent in question. Although we opted for an explicit representation, a rule-based one is also possible.

We denote the unique set of all possible configurations with $\mathbf{C}$. As we want an agent to be able to represent and reason about a group Ag, we assume a set of equivalence relations:

$$\sim_j : \mathbf{C} \times \mathbf{C} \quad \forall j \in Ag \qquad\qquad \sim_0 : \mathbf{C} \times \mathbf{C}$$

such that if $c \sim_j c'$ then these configurations hold the same

information about agent j, and if $c \sim_0 c'$ then they hold the same "local information", i.e. about self and the world. Hence, c is said to be *locally indistinguishable* from c' iff $c \sim_0 c'$. These arbitrary relations depend on the specific set of configurations (e.g. internal language) considered.

Definition 1. Formally, given $\mathbf{C}$, an *agent model* is a tuple $\mathcal{M}_i = \langle C_i, \mathcal{O}, \tau_i, \omega_i, pr_i, Act, pre_i \rangle$ where:

- $C_i \subseteq \mathbf{C}$ is a finite set of configurations;
- $\mathcal{O}$ is a finite alphabet of observations;
- $\tau_i \subseteq C_i \times C_i$ is a transition relation: $c' \in \tau_i(c)$ means that the agent can move from c to c' by reasoning;
- $\omega_i : C_i \times \mathcal{O} \to C_i$ is a transition function: $c' = \omega_i(c, o)$ models the act of registering the received observation symbol o. Therefore, it is such that $c \sim_0 c'$;
- $pr_i : \mathbf{C} \to C_i$ is a function, termed *projection* function, that will be discussed later (see Definition 2);
- Act is a finite set of action labels, and $pre_i : C_i \to 2^{Act}$ is a function such that α is *plausible* in c iff $\alpha \in pre_i(c)$.

The function pre_i is used to model action preconditions, but also to express plausibility with respect to goals and known plans, as in BDI agents [Rao and Georgeff, 1991].

Given $\mathcal{A}$ and a *library* M, i.e. a finite set of agent models, a *model assignment* is a function $\mathfrak{R} : \mathcal{A} \to M$ that assigns a model to each agent. We can imagine $\mathfrak{R}$ to capture the present situation, and we assume it to be fixed. Given $\mathfrak{R}$ and $\mathcal{A}$, we call $\Gamma = \langle \mathcal{M}_j \rangle_{j \in \mathcal{A}}$ the *context*, with $\mathcal{M}_j = \mathfrak{R}(j)$ for each j. A context captures the perspective of the reasoning agent. An agent model $\mathcal{M}_i$, together with a configuration $c \in C_i$, is called *mental state*, denoted $\mathcal{S}_i = \langle \mathcal{M}_i, c \rangle$.

Consider a propositional setting (that we will use in our scenario) describing an agent's internal logic with objective language $\mathcal{P}$. Let $\mathcal{L}$ be the language with grammar :

$$\varphi ::= \psi \mid Bel_j(\varphi) \mid \neg\varphi$$

where $j \in Ag$ and $\psi \in \mathcal{P}$. By writing φ, we represent the fact that the agent in question (say i) believes that formula φ is true, whereas $Bel_j(\varphi)$ denotes the fact that the agent believes that agent j believes φ. Here, *belief* refers to a syntactic object denoting a *fact* regarded as true, with no assumed semantic properties. We do not require a belief base to be consistent or closed under logical implication, as such automaticity may be overly optimistic representations of real believers.

Then $\mathbf{C}$ can be taken as the infinite set of all possible belief bases over $\mathcal{L}$ and, e.g., we can define $\sim_j$ to be such that $c \sim_j c'$ iff $\{\varphi \mid Bel_j(\varphi) \in c\} = \{\varphi \mid Bel_j(\varphi) \in c'\}$. Similarly, $c \sim_0 c'$ iff the set of formulas in c and c' not of the form $Bel_j(\varphi)$, for any $j \in ch(i)$ in $\mathcal{A}$, are the same.

For the agent model $\mathcal{M}_i$, C_i can be the set of allowed belief bases according to some syntactic restriction, e.g. size of belief bases (memory). Also, τ_i and ω_i can model a reasoning machinery Δ_i, i.e. a set of axioms and deductive rules that, together with $\mathcal{L}$, characterise a deductive system for i. This has some similarity with the notion of multilanguage systems [Giunchiglia and Serafini, 1994] and, in general, with resource-bounded agents [Alechina and Logan, 2009]. Similarly to the latter, these models can be used to capture agents with limited computational resources (humans), in which the cost of deliberation is considered [Isaac and Bridewell, 2014].

Definition 2. Consider two agents i, j such that $j \in ch(i)$. Given a mental state $\mathcal{S}_i = \langle \mathcal{M}_i, c \rangle$ for i, the mental state *ascribed* to agent j by i is $\mathcal{S}_j = \langle \mathcal{M}_j, pr_j(c) \rangle$.

We extend the definition to the case where j is not a child of i by trivially applying a chain of projections. This projection function allows us to *retrieve* the configuration that is currently considered by agent j (its current internal state, according to c). For this reason, we impose a constraint on each pr_j to agree with the notion of indistinguishability over $\mathbf{C}$:

$$pr_j(c) = pr_j(c') \text{ if } c \sim_j c'$$

for each $j \in \mathcal{A}$. In Section 5 we will make use of ascribed mental states to define the ability of reasoning *as* others.

Considering the propositional setting above, we may choose to have $pr_j(c) := \{\varphi \mid Bel_j(\varphi) \in c\}$ for each j, or we may define these functions to encode different *representations* of beliefs due to terminology or cognitive differences.

Definition 3. A context Γ is *complete* iff, for each $i \in \mathcal{A}$: (1) for each vector of configurations $\vec{c} \in \times_{j \in ch(i)} C_j$, one for each child j of i, there exists $c \in C_i$ such that $pr_j(c) = \vec{c}_j$; and (2) for each $c, c' \in C_i$ with $c \not\sim_0 c'$, there exists $c'' \in C_i$ holding the same information about children as c', and about i as c. Formally: $c'' \sim_0 c$ and $c'' \sim_j c'$ for any $j \in ch(i)$.

Γ is complete when any possible vector of configurations, one for each child, can be captured by a single configuration of the parent. From now on, we assume a complete Γ.

4 Modeling the Wumpus Quest

Let us consider again the scenario from Section 2. As customary in the Wumpus World, the dungeon is represented by a grid. The Wumpus occupies one cell, and each cell adjacent to this has a stench. When the Wumpus is killed, a scream can be heard throughout the dungeon. For simplicity, we do not consider pits and breezes. The set of observation symbols is $\mathcal{O} = \{s, a, bs, ba, d, sm\}$. The first two correspond to observing the state of the Wumpus (sleeping and awake), the second two to the signal from the knight (whether the Wumpus is awake). d is the observation of the Wumpus screaming, and sm of its smell. The set of actions is $Act = \{Mv, nil, At, Bs, Ba\}$, namely $move$, $wait$, $attack$, and signal that the Wumpus is asleep or awake.

Assuming the propositional setting as in Section 3, we model the agents' beliefs with a set of propositions $\{\texttt{bs}, \texttt{ba}, \texttt{W}, \texttt{WA}, \texttt{WS}, \texttt{dead}, \texttt{scared}, \texttt{dec}, \texttt{att}\}$, where `bs` and `ba` represent beliefs about the fact that the agent in question received observations bs and ba; `W` represents the belief that the position of the Wumpus is known; `WA`, `WS` and `dead` reprensent beliefs about the state of the Wumpus (awake, sleeping or dead, respectively); `scared`, `dec`, `att` are believed if the agent is scared, intends to deceive the other, or is ready to attack the Wumpus, respectively.

Assume that the reasoning agent is the Black Knight (agent 1). He assigns to himself the model $\mathcal{M}_{BK}$, to agent 12 the model $\mathcal{M}_P$ for the *peasants* reference group, and to agent 121 the model $\mathcal{M}_{WK}$. There models are depicted in in Figure 2, where transitions in each ω_i do not have labels, and loops of the form $\langle c, o, c \rangle \in \omega_i$ are omitted. The table lists the configurations in $\mathbf{C}$ used in the agent models (where $CB(\cdot)$ stands

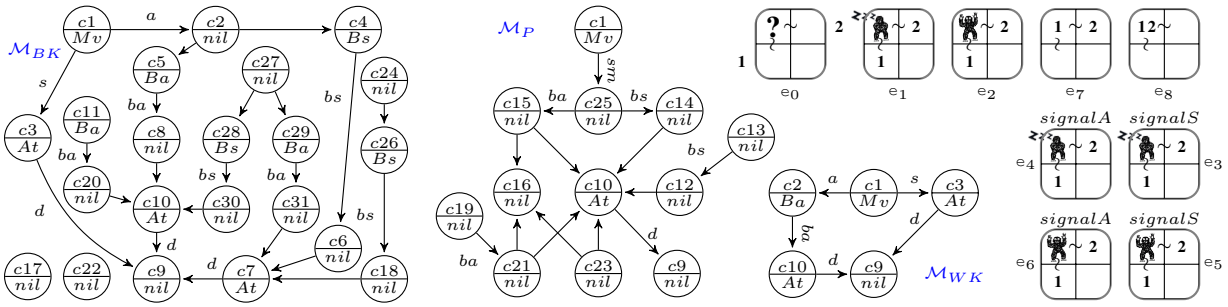

Figure 2: The agent models $\mathcal{M}_{BK}$, $\mathcal{M}_P$, $\mathcal{M}_{WK}$, and the set of environment states.

c1: **c2:** WA **c3:** WS **c4:** WA, dec **c5:** WA, dec **c6:** WA, bs **c7:** WA, $Bel_2(CB(\texttt{WS}))$, $CB(\texttt{att})$ **c8:** WA, ba **c9:** $CB(\texttt{dead})$ **c10:** $CB(\texttt{WA})$, $CB(\texttt{att})$ **c11:** WA, $Bel_2(Bel_1(\texttt{WA}))$, $Bel_2(\texttt{W})$ **c12:** W, $Bel_1(\texttt{WS})$, bs **c13:** W, $Bel_1(\texttt{WS})$ **c14:** W, bs **c15:** W, ba **c16:** scared **c17:** WA, $Bel_2(\texttt{scared})$, ba **c18:** WA, $Bel_2(Bel_1(\texttt{WS}))$, $Bel_2(\texttt{W})$, bs **c19:** W, $Bel_1(\texttt{WA})$ **c20:** WA, $Bel_2(Bel_1(\texttt{WA}))$, $Bel_2(\texttt{W})$, ba **c21:** W, $Bel_1(\texttt{WA})$, ba **c22:** WA, ba, $Bel_2(\texttt{W})$, $Bel_2(\texttt{ba})$, $Bel_2(Bel_1(CB(\texttt{att})))$, $Bel_2(Bel_1(CB(\texttt{WA})))$ **c23:** W, ba, $Bel_1(CB(\texttt{WA}))$, $Bel_1(CB(\texttt{att}))$ **c24:** WA, $Bel_2(Bel_1(\texttt{WS}))$, $Bel_2(\texttt{W})$ **c25:** W **c26:** WA, $Bel_2(Bel_1(\texttt{WS}))$, $Bel_2(\texttt{W})$, dec **c27:** WA, $Bel_2(\texttt{W})$ **c28:** WA, $Bel_2(\texttt{W})$, dec **c29:** WA, $Bel_2(\texttt{W})$, $\neg$dec **c30:** WA, bs, $Bel_2(\texttt{W})$, $Bel_2(\texttt{bs})$ **c31:** WA, ba, $Bel_2(\texttt{W})$, $Bel_2(\texttt{ba})$.

for a syntactic representation of common belief). Projections are defined accordingly. Each configuration corresponds to a state, and each state is labelled with the set of plausible actions (here only singletons). Consider for example $\mathcal{M}_{\mathcal{P}}$. In $c1$ the peasant has no relevant belief, and only action Mv is plausible. An agent adhering to this model will keep moving until the *smell* observation is received (transition $\langle c1, sm, c25\rangle$). When this happens, the peasant will wait for a signal. If ba is received, then the next configuration $c15$ contemplates two courses of thought: $\langle c15, c16\rangle$ and $\langle c15, c10\rangle$. In the first case the peasant will wait indefinitely (scared); otherwise it will attack. The rest of the configurations were omitted for brevity. Unreachable configurations will become reachable via empathetic reasoning, as we see in the next section. For instance, if the current mental state of agent 1 is $\mathcal{S}_1 = \langle \mathcal{M}_{BK}, c11\rangle$, then $\mathcal{S}_{12} = \langle \mathcal{M}_P, c19\rangle$ and $\mathcal{S}_{121} = \langle \mathcal{M}_{WK}, c2\rangle$.

To model any concrete example, we need to first define our notion of *environment*. A (nondeterministic) environment is:

$$\mathcal{E} = \langle E, e^0, \gamma, perc\rangle$$

where E is a finite set of environment states; $e^0 \in E$ is the initial state; $\gamma = E \times Act^{|Ag|} \times E$ is a transition relation; $perc : E \times Ag \to \mathcal{O}$ is an observation function that returns an observation o for each agent in Ag. As customary [Fagin *et al.*, 1995; Parikh and Ramanujam, 1985; Wooldridge and Lomuscio, 2001], $\mathcal{E}$ is used to represent the physical world.

Assume that the dungeon is as in the right side of Figure 2 (states $\texttt{e}_0$-$\texttt{e}_8$: we restrict its size for simplicity, and we consider only one position of the Wumpus). E.g., $\texttt{e}_0$ represents the situation when both hunters are outside; $\texttt{e}_3$ the one in which the Wumpus is asleep and the Knight signaled so. The transition relation is defined such that the effect of actions is reflected in the position of the hunters and in the communication channel. To model the fact that the state of the Wumpus is unknown, we consider two transitions from $\texttt{e}_0$ with actions $\vec{\alpha} = \langle Mv, Mv\rangle$, namely $\langle \texttt{e}_0, \vec{\alpha}, \texttt{e}_1\rangle$ and $\langle \texttt{e}_0, \vec{\alpha}, \texttt{e}_2\rangle$.

5 Reasoning as and about others

We now characterize two types of reasoning: (1) *Stereotypical*: when an agent represents and reasons about other agents and their beliefs; (2) *Empathetic*: the agent casts itself into the mind of another, reasoning as the other would. The former is similar to that of standard Epistemic Logic [Fagin *et al.*, 1995], in which every agent is homogenous, while the latter uses different perspectives and inference mechanisms for others. Recalling the notion of ascribed mental state (Definition 2) we can now state these concepts formally:

1. An agent i with mental state $\langle \mathcal{M}_i, c\rangle$ performs a *stereotypical* reasoning *about* agent $j \in ch(i)$ when it performs a transition $\langle c, c'\rangle \in \tau_i$ with $c \not\sim_j c'$. The transition captures the application of a *stereotype* about j.
2. An agent i performs an *empathetic* reasoning *as* agent j whenever it performs a transition $\langle pr_j(c), c'\rangle \in \tau_j$, i.e., progresses the ascribed mental state $\langle \mathcal{M}_j, pr_j(c)\rangle$.

These are depicted in Figure 3 (left), where mental states in black may have changed, while gray ones are updated with respect to one child. In the first (above) agent 1 performs a transition $c'_1 \in \tau(c_1)$ and, as a consequence, the (implicit) mental state ascribed to 12 may change. In the second, agent 1 computes $c_{12} = pr_{12}(c_1)$ then $c'_{12} \in \tau(c_{12})$ then c'_1, as we will see. Computing such c'_1 in case of empathetic reasoning is a nontrivial step: c'_1 needs to hold the updated information about 12 as well as preserving the rest of the prexisting information from c_1 that is about 1 or other children.

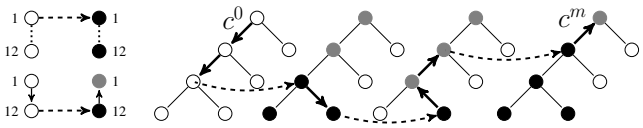

Figure 3: (left) Representation of stereotypical and empathetic steps. (right) Depiction of an expansion from c^0 to c^m.

With these two simple concepts at hand, we describe how an agent reasons by applying either or both these strategies. Of importance here is that we are describing the "state-space" in which an agent reasons, but not its *algorithmic* process. In other words, an agent that reasons with a context Γ needs its own heuristics for deciding when and how apply either form of reasoning, but we are agnostic with respect to this implementation. Instead, we define here the outcome of the reasoning through the notion of *expansion*.

A mental path for i is a sequence of agent labels $j^0 \cdots j^m$ such that $j^0 = j^m = i$ and, for every $0 < \ell < m$, either $j^{\ell+1} = j^\ell$, $j^{\ell+1} \in ch(j^\ell)$ or $j^{\ell+1} = pa(j^\ell)$. That is, a mental path is a path in the tree of $\mathcal{A}$ that starts and ends at the root. A path σ represents the mental steps of the reasoning agent when it direct its attention towards virtual agents, projecting the corresponding mental states, to identify a possible *representation* for the result of this simulated reasoning.

Finally, an *expansion* of $\langle \mathcal{M}_i, c_i \rangle$ is a sequence of mental states $\sigma = \langle \mathcal{M}_{j^0}, c^0 \rangle, \langle \mathcal{M}_{j^1}, c^1 \rangle, \cdots, \langle \mathcal{M}_{j^m}, c^m \rangle$ such that $c^0 = c_i$, $j^0 \cdots j^m$ is a mental path for i and, $\forall \ell < m$, either:

(a1) $j^{\ell+1} = j^\ell$ and $c^{\ell+1} \in \tau_{j^\ell}(c^\ell)$, i.e. a local transition;

(a2) $j^{\ell+1} \in ch(j^\ell)$ and $c^{\ell+1} = pr_{j^\ell}(c^\ell)$;

(a3) $j^{\ell+1} = pa(j^\ell)$ and $c^{\ell+1} \sim_{j^\ell} c$ for any c s.t. $pr_j(c) = pr_j(last_j(\sigma, \ell))$ and $j \in ch(j^{\ell+1})$. $last_j(\sigma, \ell)$ is the last configuration of agent j^ℓ in the prefix of σ of length ℓ.

Point (a2) represents a top-down projection from j^ℓ to $j^{\ell+1}$, and (a3) represents an inverse projection, that computes a configuration $c^{\ell+1}$ holding the preexisting information about the parent and other children, plus the updated information about child j^ℓ. Points (a1)-(a3) can be always computed, as we assumed Γ to be complete. For example, a possible expansion from $\langle \mathcal{M}_{BK}, c22 \rangle$ is $\sigma = c22, c23, c16, c17$, whose path is $1, 12, 12, 1$. Figure 3 (right) depicts an expansion (dashed edges are local transitions – point (a1) above).

Expansions allow to retrieve a new configuration (like c^m in Figure 3) through the thinking that happens along the path, but they do not consider observations. We now define a function that computes configurations resulting from mental expansions, by first considering received observations. Given Γ and a model $\mathcal{M}_i$, we compute the relation $next_i \subseteq C_i \times \mathcal{O}^{|\mathcal{A}|} \times C_i$ such that $\langle c^0, \vec{o}, c^m \rangle \in next_i$ iff:

- $c'_i \in \omega_i(c, \vec{o}_i)$ and, $\forall j \in \mathcal{A}$, $j \neq i$, $c'_j \in \omega_j(pr_j(c'_i), \vec{o}_j)$;
- c^m is the last configuration of an expansion from $\langle \mathcal{M}_i, c' \rangle$, with $pr_j(c') = c''_j$ and $c''_j \sim_0 c'_j$, $\forall j \in \mathcal{A}$.

The first item progresses each model separately; the second selects a new c' for i capturing all the mental states by only looking at their local information, then returns an expansion.

6 Social autism and ToM

In this section we illustrate different degrees of social behaviours, and how they affect computational aspects of the collective execution. We imagine that the agent can *simulate* executions "in its mind" to foresee plausible evolutions. Given Γ and $\mathcal{E}$, the set of *global states* for agent i is

$$G_i = E \times C_i$$

such that each global state $g = \langle e, c \rangle$ holds the (current) environment state and a configuration for agent i.

If we define an *action vector* as a tuple of size $|Ag|$ comprising an action for each concrete agent, then a sequence $\rho = g^0 \vec{\alpha}^0 g^1 \vec{\alpha}^1 \cdots$, alternating global states and action vectors, is termed a *run* for agent i on $\mathcal{E}$.

In what follows, to ease the notation, we assume that i is the reasoning agent, and we use j to quantify on $\mathcal{A}$.

Zero-order ToM: autistic agents

In the most basic setting, an agent is not socially intelligent: it does not consider models of other agents (although it may attribute them beliefs), thus it is not able to reason *as* them. The typical approach for group strategies is to synthesize a plan where the group collectively achieves some object. For this plan to be successful, each agent does not need to represent others: the synthesis of the plan is done externally, from a third-person view. Even if these agents may agree on a protocol, they are incapable of devising a new strategy autonomously when failure happens. $\mathcal{A} = \{i\}$ and $\Gamma = \langle \mathcal{M}_i \rangle$.

Given a mental state $\langle \mathcal{M}_i, c^0 \rangle$ for i, we say that $\rho = g^0 \vec{\alpha}^0 g^1 \vec{\alpha}^1 \cdots$ is a *feasible run* for agent i over $\mathcal{E}$ iff $g^0 = \langle e^0, c^0 \rangle$, and $g^{\ell+1} = \langle e^{\ell+1}, c^{\ell+1} \rangle$ is such that for each $\ell \geq 0$:

- $\vec{\alpha}_i^\ell$ is plausible in $\langle \mathcal{M}_i, c^\ell \rangle$;
- $e^{\ell+1} \in \gamma(e^\ell, \vec{\alpha}^\ell)$;
- $c^{\ell+1} \in next_i(c^\ell, \langle o \rangle)$, with $o = perc(e^{\ell+1}, i)$.

The agent only considers its own actions and received observations, ignoring the actions and observations of others.

Example 1. (Wumpus Quest). If we consider all feasible runs as possible, then a strategy for agent i guaranteed to kill the Wumpus does not exist. The knight does not take into consideration the fact that the other may get scared at the thought of facing an awake Wumpus: there are two distinct courses of thought $\langle c15, c16 \rangle, \langle c15, c10 \rangle$ in $\mathcal{M}_P$. Instead $\mathcal{M}_{BK}$ does not contemplate this, as the only path from $c31$ leads to the decision to attack: the knight modelled as a zero-order ToM agent is not able to foresee this.

First-order ToM: socially aware agents

The second category of social behaviours arises when the agent has its own representation of others, and can reason *as* them. A socially aware agent is aware of the fact that others have different mental states, but it only assigns a model to all concrete agents: $\mathcal{A}$ is the set $\{i\} \cup (i \cdot Ag)$.

We say that $\rho = g^0 \vec{\alpha}^0 g^1 \vec{\alpha}^1 \cdots$ is a *plausible run* of Γ over $\mathcal{E}$ for agent i iff $g_0 = \langle e^0, c^0 \rangle$, and $g^{\ell+1} = \langle e^{\ell+1}, c^{\ell+1} \rangle$ is s.t.:

- $\vec{\alpha}_i^\ell$ is plausible in $\langle \mathcal{M}_i, c^\ell \rangle$;
- $\vec{\alpha}_{l(j)}^\ell$ is plausible in $\langle \mathcal{M}_j, pr_j(c^\ell) \rangle$, for any $j \in ch(i)$;
- $e^{\ell+1} \in \gamma(e^\ell, \vec{\alpha}^\ell)$;
- $c^{\ell+1} \in next_i(c^\ell, \vec{o})$, with each $\vec{o}_j = perc(e^{\ell+1}, l(j))$;

Note how $l(j)$ is used to assign observations to virtual agents.

Example 2. (Wumpus Quest). A strategy now exists, and the runs it generates are plausible. Being able to assign $\mathcal{M}_P$ to 12, agent 1 has a strategy: if he sees the Wumpus sleeping he will kill it without help. Otherwise, he will signal that it is instead asleep, deceiving the other into cooperation. Indeed, $\mathcal{M}_P$ assumes that the peasant would not be reluctant to help

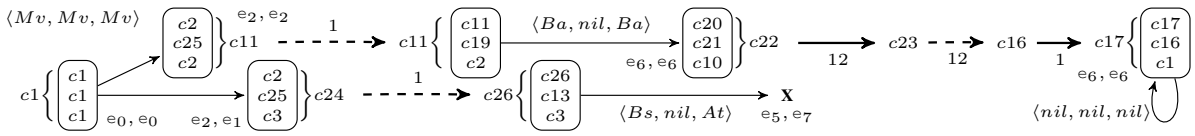

Figure 4: Two runs of the Wumpus Quest example.

in this case: the only course of thought ($\langle c14, c10\rangle$ in $\mathcal{M}_P$) leads to common belief about the state of the Wumpus.

Second-order ToM agents

Although socially aware agents may show social behaviours and attribute mental states to others, they lack an evolved theory of mind, i.e. to acknowledge that others are reasoning in the same way, and therefore have expectations and models of them: socially aware agents do not model virtual agents. For second-order ToM agents instead, $\mathcal{A}$ (and therefore Γ) includes two levels of the heirarchy: agent i, all agents $j \in i{\cdot}Ag$ and (some) $j' \in ch(j)$, with $j \in ch(i)$.

However, to keep track of what others regard as plausible, we need to understand what does it mean to behave in an *acceptable* way. *Our view is that the runs generated by the collective actions appear plausible to each concrete agent.*

This requires to (a) extend the size of action vectors to the size of $\mathcal{A}$, and (b) a new definition of global states, in which we keep track of the environment state that each concrete agent regards as both possible and acceptable:

$$G_i^+ = E_1 \times \cdots \times E_{|Ag|} \times C_i$$

A run ρ is an *acceptable run* for i of Γ over $\mathcal{E}$, iff $g_0 = \langle e_1^0, \ldots, e_n^0, c^0\rangle$, and $g^{\ell+1} = \langle e_1^{\ell+1}, \ldots, e_n^{\ell+1}, c^{\ell+1}\rangle$ is s.t.:

- each $\vec{\alpha}_j^\ell$ is a plausible action in c^ℓ, for any $j \in \mathcal{A}$;
- $e_i^{\ell+1} \in \gamma(e_i^\ell, act(\vec{\alpha}^\ell, i))$;
- $e_j^{\ell+1} \in \gamma(e_j^\ell, \langle \alpha_1, \ldots, \alpha_{|Ag|}\rangle)$ where
$$\alpha_i = \vec{\alpha}_{j\cdot i}^\ell \text{ and } \alpha_{l(j)} = \vec{\alpha}_j^\ell \text{ for each } j \in ch(i).$$
- $perc(e_i^{\ell+1}, l(j)) = perc(e_j^{\ell+1}, l(j))$ for each $j \in ch(i)$;
- $c^{\ell+1} \in next_i(c^\ell, \vec{o})$, such that $\vec{o}_i = perc(e_i^{\ell+1}, i)$, and $\vec{o}_j = perc(e_{j'}^{\ell+1}, l(j))$, with $j' = pa(j)$, for any $j \neq i$;

First, each action component is plausible to the ascribed mental state (first item). In the second item, $act(\vec{\alpha}, i)$ is the vector of actions computed from $\vec{\alpha}^\ell$ by only taking the plausible actions for i and each child $j \in ch(i)$ in $\mathcal{A}$, ordered by $l(j)$. E.g. if $i = 1$ and $\mathcal{A} = \langle 1, 12, 121\rangle$ then $act(\langle a_1, a_{12}, a_{121}\rangle, 1) = \langle a_1, a_{12}\rangle$ and $act(\langle a_1, a_{12}, a_{121}\rangle, 12) = \langle a_{121}, a_{12}\rangle$. This is the environment state that agent i considers as plausible (and, as $\mathcal{E}$ is nondeterministic, more than one can exist). Similarly, in the third item, a new environment component is computed for each agent $j \in ch(i)$, but the only actions forced to be plausible are those of j and i (indeed, we are not interested in assuming acceptable behaviours for agents other than i; in that case, the condition can be replaced with $e_j^{\ell+1} \in \gamma(e_j^\ell, act(\vec{\alpha}^\ell, j))$). In the fourth item, we compare the observations received if we take the environment state $e_i^{\ell+1}$ that the reasoning agent i considers as the next one, with the state $e_j^{\ell+1}$ that each other agent j is expecting to receive, i.e. would be observed if their prevision (about the action selected by agent i) was correct. That is, each $e_j^{\ell+1}$ *justifies* $perc(e_i^{\ell+1}, l(j))$ to agent j. Note that, in the last item, virtual agents observe the environment component of their parents. One such explanation is selected for each $g^{\ell+1}$ and, at each future step, a new one needs to be found. When one such run exists, then the observations received by all concrete agents, at each step, are *justifiable*, otherwise the behaviour of i violates Γ (in this case, it is "unmasked"). Therefore, acceptability depends on the observations that others receive. In the case of observable actions, this requires equality, but this is not true for private actions.

Example 3. (Wumpus Quest) The informal strategy in Example 2 does not always generate acceptable runs. Indeed, $\mathcal{M}_{121} = \mathcal{M}_{BK}$ tells ($c3$) that the White Knight would not ask for help if not needed. Hence, as the whistle is blown to signal that the Wumpus is sleeping when in reality it is awake ($e_1 = \mathrm{e}_5$), the peasant can detect a misalignment with his expectations (a next environment state $e_{12} = \mathrm{e}_7$ in which the scream was heard), hence will fail to find a justification for the signal: $perc(\mathrm{e}_5, 1) \neq perc(\mathrm{e}_7, 2)$; thus unmasking the Knight. This run is depicted in Figure 4 (run above; the other is one that leads the peasant to get scared, and the rest are omitted). If the knight shows enough empathetic attitude, he could realise, before acting, that his action may appear unjustifiable to 12. As we said, we are agnostic on his decision, but we are now capable of formalizing this notion.

Finally, note that acceptable runs capture the view stated earlier. Given one acceptable run ρ for i, the run over G *observed* by $j \in ch(i)$ is the sequence $\rho_j = g_j^0\ act(\vec{\alpha}^0, j) \cdots$, with $g_j^\ell = \langle e_{l(j)}^\ell, pr_j(c^\ell)\rangle$ for each $\ell \geq 0$.

Theorem 1. Given an acceptable run τ for i, any run of Γ over $\mathcal{E}$ observed by each agent $j \in ch(i)$ is plausible for j.

7 Conclusions and Future Work

Using agent models, we devised a computational mechanism to discriminate which behaviours of others are plausible, and decide which behaviour for ourselves is acceptable. In future work, we will define different notions of acceptability, and investigate how to build ATL (Alternating-time Temporal Logic [Alur *et al.*, 2002]) games to verify strategic abilities in this setting, and synthesise strategies that conform to Γ, i.e. induce acceptable runs, also in scenarios of deception.

Acknowledgements: This research is partially funded by Australian Research Council Discovery Grant DP130102825, *Foundations of Human-Agent Collaboration: Situation-Relevant Information Sharing*

References

[Alechina and Logan, 2009] Natasha Alechina and Brian Logan. A logic of situated resource-bounded agents. *J. of Logic, Language and Information*, 18(1):79–95, January 2009.

[Alur *et al.*, 2002] Rajeev Alur, Thomas A. Henzinger, and Orna Kupferman. Alternating-time temporal logic. *J. ACM*, 49(5):672–713, September 2002.

[Andersen *et al.*, 2014] Mikkel Birkegaard Andersen, Thomas Bolander, and Martin Holm Jensen. Don't plan for the unexpected: Planning based on plausibility models. *Logique et Analyse*, 1(1), 2014.

[Bergwerff *et al.*, 2014] Gerben Bergwerff, Ben Meijering, Jakub Szymanik, Rineke Verbrugge, and Stefan M Wierda. Computational and algorithmic models of strategies in turn-based games. *Proceedings of CogSci*, 2014.

[Bolstad and Endsley, 1999] Cheryl A. Bolstad and Mica R. Endsley. Shared mental models and shared displays: An empirical evaluation of team performance. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, volume 43, pages 213–217, 1999.

[Brooks, 1991] Rodney Brooks. Intelligence without representation. *Artificial Intelligence*, 47:139–159, 1991.

[Bulling and Jamroga, 2007] Nils Bulling and Wojciech Jamroga. Agents, beliefs, and plausible behavior in a temporal setting. In *AAMAS*, page 146, 2007.

[Cannon-Bowers *et al.*, 1993] Janis A. Cannon-Bowers, Eduardo Salas, and Sharolyn Converse. Shared mental models in expert team decision making. *Individual and group decision making: Current issues*, 221, 1993.

[Dignum *et al.*, 2014] Frank Dignum, Gert Jan Hofstede, and Rui Prada. From autistic to social agents. In *AAMAS*, pages 1161–1164, 2014.

[Ditmarsch *et al.*, 2007] Hans van Ditmarsch, Wiebe van der Hoek, and Barteld Kooi. *Dynamic Epistemic Logic*. Springer, Incorporated, 1st edition, 2007.

[Espevik *et al.*, 2011] Roar Espevik, Bjørn Helge Johnsen, and Jarle Eid. Outcomes of shared mental models of team members in cross training and high-intensity simulations. *J. of Cognitive Engineering and Decision Making*, 5(4):352–377, 2011.

[Fagin *et al.*, 1995] Ronald Fagin, Joseph Y. Halpern, Yoram Moses, and Moshe Y. Vardi. *Reasoning about Knowledge*. MIT Press, Cambridge, MA, 1995.

[Ficici and Pfeffer, 2008] Sevan G. Ficici and Avi Pfeffer. Modeling how humans reason about others with partial information. In *AAMAS*, pages 315–322, 2008.

[Giunchiglia and Serafini, 1994] Fausto Giunchiglia and Luciano Serafini. Multilanguage hierarchical logics, or: How we can do without modal logics. *Artificial Intelligence*, 65(1):29 – 70, 1994.

[Isaac and Bridewell, 2014] Alistair Isaac and Will Bridewell. Mindreading deception in dialog. *Cognitive Systems Research*, 28:12–19, 2014.

[Johnson *et al.*, 2014] Matthew Johnson, Jeffrey M. Bradshaw, Paul J. Feltovich, Catholijn M. Jonker, M. Birna van Riemsdijk, and Maarten Sierhuis. Coactive design: Designing support for interdependence in joint activity. *Journal of Human-Robot Int.*, 3(1):43–69, 2014.

[Johnson-Laird, 1983] Philip N Johnson-Laird. *Mental models: Towards a cognitive science of language, inference, and consciousness*. Harvard University Press, 1983.

[Kaminka, 2013] Gal A Kaminka. Curing robot autism: A challenge. In *AAMAS*, pages 801–804, 2013.

[Klimoski and Mohammed, 1994] Richard Klimoski and Susan Mohammed. Team mental model: Construct or metaphor? *Journal of management*, 20(2):403–437, 1994.

[Lakemeyer, 1986] Gerhard Lakemeyer. Steps towards a first-order logic of explicit and implicit belief. In *Proceedings of the 1986 Conference on Theoretical aspects of reasoning about knowledge*, pages 325–340, 1986.

[Lemaignan *et al.*, 2010] S. Lemaignan, R. Ros, L. Mösenlechner, R. Alami, and M. Beetz. Oro, a knowledge management module for cognitive architectures in robotics. In *IROS 2010*, 2010.

[Levesque, 1984] Hector J Levesque. A logic of implicit and explicit belief. In *AAAI*, pages 198–202, 1984.

[Mccarthy, 1992] John Mccarthy. Overcoming an unexpected obstacle. Technical report, Computer Science Department, Stanford University, 1992.

[Mohammed *et al.*, 2010] Susan Mohammed, Lori Ferzandi, and Katherine Hamilton. Metaphor no more: a 15-year review of the team mental model construct. *J. of Management*, 2010.

[Parikh and Ramanujam, 1985] R. Parikh and R. Ramanujam. Distributed processes and the logic of knowledge. In *Logic of Programs*, pages 256–268, 1985.

[Pfau *et al.*, 2014] Jens Pfau, Yoshihisa Kashima, and Liz Sonenberg. Towards agent-based models of cultural dynamics: A case of stereotypes. In *Perspectives on Culture and Agent-based Simulations*, pages 129–147. Springer, 2014.

[Premack and Woodruff, 1978] David Premack and Guy Woodruff. Does the chimpanzee have a theory of mind? *Behavioral and Brain Sciences*, 1:515–526, 12 1978.

[Rao and Georgeff, 1991] Anand S. Rao and Michael P. Georgeff. Modeling rational agents within a BDI-architecture. In *KR*, pages 473–484, 1991.

[Scassellati, 2002] Brian Scassellati. Theory of mind for a humanoid robot. *Autonomous Robots*, 12(1):13–24, 2002.

[Warnier *et al.*, 2012] Matthieu Warnier, Julien Guitton, Séverin Lemaignan, and Rachid Alami. When the robot puts itself in your shoes. Managing and exploiting human and robot beliefs. In *RO-MAN*, pages 948–954. IEEE, 2012.

[Wooldridge and Lomuscio, 2001] M. Wooldridge and A. Lomuscio. A computationally grounded logic of visibility, perception, and knowledge. *Logic Journal of the IGPL*, 9(2):273–288, 2001.