Artificial social reasoning: computational mechanisms for reasoning about others

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. With a view to supporting expressive, but tractable, collaborative interactions between humans and agents, we propose an approach for representing heterogeneous agent models, i.e., with potentially diverse mental abilities and holding stereotypical characteristics as members of a social reference group. We build a computationally grounded mechanism for progressing their beliefs about others' beliefs, supporting stereotypical as well as empathic reasoning. We comment on how this approach can be used to build finite-state games, restricting the analysis of possibly large-scale problems by focusing only on the set of plausible evolutions.

Keywords: agents, mental models, stereotypes

1 Introduction

In a multi-agent setting, equipping agents with an awareness of their social reality [5] will enable more seamless interdependent collective behaviour [8], where interdependency informally means that one agent's deliberation is dependent on what another agent does (or intends to do), and vice-versa. Agents can be thought of as following a social behaviour, depending on the particular context in which they are interacting, so one critical feature that needs to be ascribed to intelligent agency is the ability to represent and reason about the *common ground* between agents, including their beliefs about stereotypes [11].

In this paper, we propose an approach for representing both the beliefs and the *model* that one agent has of the environment and others, including their nested beliefs, so to allow for the synthesis of strategies to achieve goals. In doing so, we combine temporal and belief projection in an attempt to predict future decisions of others [10], focusing on *plausible* evolutions instead of just feasible ones. Of importance in our approach is that it supports two types of reasoning about others: *stereotypical reasoning*, which allows an agent to reason about another using simple social rules; and *empathic reasoning*, in which the agent casts itself into the mind of another agent and reasons about what it would do.

In a multi-agent setting, it is typical that group strategies are synthesized by an omniscient entity, and then dispatched to agents, which are merely executors with limited ability to reason about the reality in which they are immersed. In this paper, we devise a computationally grounded (and implementable) mechanism for representing belief and progression, which reflects the *local* perspective

an agent with respect to its own understanding of the world as well as of others (first-person view). This is contrast to considering the beliefs that an omniscient observer ascribes to each agent (third-person view). An agent can use its internal representation and inference mechanisms for itself, yet can use alternative representations and inference mechanisms for others. This can model *realistic* agents [1] (for example, with constrained resources) as well as *ideal* ones. Deliberation and action execution are both local – that is, the agent simulates other agents' deliberations to deliberate itself – thus empowering interdependence and awareness. Such capability is essential when modelling humans, whose adherence to a protocol is subject to their understanding of it, and supports our objective of enabling richer forms of collaboration between humans and agents.

The structure of this paper is as follows. Section 2 is the technical core of the paper, and presents a formal definition of an agent model, including support for an agent to hold an explicit representation of others'. Using this representation we define a notion of belief ascription that allows an agent i to cast itself into another agent j and reason as j would – i.e. reason *as* j, not just *about* j; such reasoning can also exploit a notion of stereotype. We then describe the deductive process involving one agent reasoning as another and how this can be done efficiently. In Section 3 we comment how this approach could be used to build finite-state games and indicate ways to achieve tractability in large-scale problems. Finally, Section 4 offers some closing comment.

1.1 Related work

In the context of multi-agent systems, considerable work has focused on the design of intelligent agents and the task of reasoning about their own knowledge and belief as well as that of others (e.g., [6, 12]). These approaches allow reasoning about nested beliefs (usually represented as a *flat* set), but do not generally consider the agents' mutual representation as part of an agent's state, and ignore the effects of the social context. Some work has considered, as we do, representations where agents maintain local (internal) models of other agents' beliefs [2], but the focus has been on rationality postulates, in contrast to our broader goal of tractable reasoning in a social context.

Studies of human-robot interactions and social robots, either virtual or concrete, identify the need for a human-oriented perception to represent and understand humans as well as other synthetic agents. Agents thus need the ability to attribute mental states –beliefs, desires, pretending, etc.– to oneself and others and to understand that others have mental states that are different from one's own (theory of mind) [13]. This applies to any human-robot interaction, from assistance to cooperation, to improve empathic interactions, e.g., [7] as well as objective and task-oriented sociable behaviours, e.g., [4]. However, this literature generally considers a finite set of fixed or probabilistic information about others –including their users– e.g., [14], and even when social behaviours are allowed to be emergent [4], the analysis is somehow limited to the agent alone, focusing on personal tendencies rather than projected mental states. For humans, the ability to take the perspective of another when reasoning about what to do in

interaction, is well studied in the psychology literature, and some recent work in human-robot interaction has sought to provide, as we do, rich and flexible mechanisms for making decisions that draw on a dynamic model of others' beliefs [9, 15]. Our work goes further both in the expressiveness of the internal agent model, and in the forms of supported reasoning.

2 Mental models and agent models

To allow one agent to reason about others in a social context, we provide agents with *agent models*. These models could describe a child, an elderly patient, a color-blind human, a highly moral (or prejudiced) agent, or a synthetic one. An agent model contains, among other components, a belief base and a set of rules for reasoning over the belief base. An agent is able to *assign* such models to others *and itself*, so when considering all possible eventualities, it is capable of determining its behaviour based on plausible estimates of others' behaviour. Our agent models can be used in almost opposing directions. On one hand they characterise both the reasoning capabilities of an individual agent, i.e., the logical system it uses, its limitations, its abilities and attitudes toward the others and, more generally, its description as a member of a reference group (*role* or *archetype*) [5]. On the other hand, they can model agents of which the role description (their mere function in the social context) is more characterising than their individual description and intimate understanding. 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 [5]. Manipulation of stereotypes enables shortcuts to be taken, both in human and computational reasoning mechanisms [11]. Departures from a stereotype which are essential for a specific model can then be made explicit.

Example 1. Imagine a superhero (1) and a police agent (2) facing a villain (3). Let us analyse the situation from the point of view of (1) – as if we were him (his perspective understood). Both (2) and (3) ignore that (1) is a superhero [S]: he is just *the average Joe* [J]. (3) knows that (2) is a police officer [P] (e.g., he is wearing a uniform), and all police officers are the same: (3) hates cops! However, (1) decided that (2) is actually a rookie [R] (he may have heard this on the police radio). There is also somebody else: a girl (4) has been taken hostage by the villain [C], and although the villain thinks she is just a girl [G], she is indeed (1)'s sidekick [K], who knows her moves! Note that all this is hardly expressible as mere *belief* formulae, as they convey resolutions, social/moral attitudes, etc; something that (1) knows by experience, as a veteran in the superhero business. Our aim is to capture these expressive concepts in a straightforward manner.

2.1 Mental models and agent models: A formal definition

We describe an agent (internal) logic L, starting with language $\mathcal{P}$, and a finite set of agent labels Ag. Let $\mathcal{L}$ be the language with the following grammar:

$$\varphi ::= \psi \mid \varphi \vee \varphi \mid \neg\varphi \mid Bel_i(\varphi)$$

where $i \in Ag$ and $\psi \in \mathcal{P}$. This language will be used by an agent to represent explicitly its own beliefs, as well as the beliefs of a fixed set of agents. By writing φ, we represent the fact that the agent in question believes that formula φ is true, whereas $Bel_i(\varphi)$ denotes the fact that the agent believes that agent i believes φ (i.e. we assume an implicit belief operator for an agent in front of formulae relevant to that agent). Φ is the set of wffs of $\mathcal{L}$. Note that *belief* refers to a syntactic object denoting a *fact* regarded as true in the world, with no assumed semantic properties. We can now go on to describe a *computational mechanism*.

Example 2. Consider Example 1. We can represent that superhero himself believes the girl is his sidekick, but that the villain believes she is a normal girl:

$$\texttt{girl=K} \wedge Bel_3(\texttt{girl=G}) \wedge Bel_3(\neg\texttt{girl=K})$$

Recall that this is represented from the viewpoint of the superhero himself, so we do not prefix that beliefs with Bel_1. Such a formalism can also represent a form of non-probabilistic uncertainty, in which believing neither a proposition nor its negation implies that we are unsure. For example, we can represent that the villain is unsure if the police officer is a rookie as:

$$\neg Bel_3(\texttt{pol=R}) \wedge \neg Bel_3(\neg\texttt{pol=R})$$

Definition 1 (Belief base). *Given the language $\mathcal{L}$, we define a belief base to be a subset of $\mathcal{L}$. We use kb as a variable to refer to a belief base, and $\boldsymbol{KB}$ to refer to the set of all belief bases.*

We place no further restrictions on the belief base: a belief base need not be consistent or closed under classical logical implication.
As a belief base is a set of beliefs that an agent holds about the world, including beliefs about others, we may want to reference our beliefs about a specific agent's beliefs. For a belief base kb, we use $kb|_i$ to represent this. Formally:

$$kb|_i = \{\varphi \mid Bel_i(\varphi) \in kb\}$$

Finally, in this paper, $kb_\downarrow$ denotes the set of formulas in kb *not* of the form $Bel_i(\varphi)$ (i.e., beliefs not about others).

Definition 2 (Mental model). *A* mental model *for agent i is a tuple $\mathfrak{M}^i = \langle KB^i, Ax^i, \mathsf{pr}_i\rangle$ where:*

- *$KB^i \subseteq \boldsymbol{KB}$ is the set of possible belief bases, denoting i's beliefs.*
- *Ax^i is a set of axioms that can be used to reason about the belief base KB^i. We do not restrict to a specific axiomatisation. On the contrary, we consider Ax^i as an* arbitrary *set of axioms, to allow modelling various forms of reasoning, adhering to different logics. Note, however, that the purpose of Ax^i is purely syntactic, and does not necessarily preserve any semantic property. Therefore, we can think of Ax^i as a set of rules $\Phi \Rightarrow \varphi$.*
- *$\mathsf{pr}_i : \boldsymbol{KB} \to KB^i$ is a surjective total function, called* projection*, which projects a belief base kb to another belief base kb' that contains only the relevant part of kb (namely, kb' holds the beliefs about i) – e.g., $\mathsf{pr}_i(kb) = kb|_i$.*

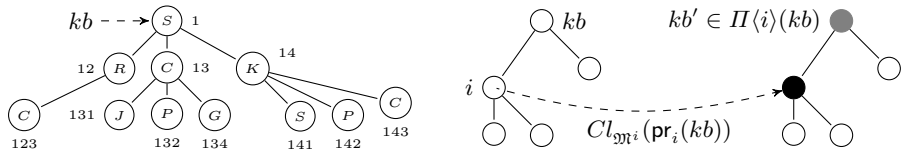

Fig. 1. (left) Representation of the set of ascribed mental states of Example 1. We stress that this induced "tree" is implicit: each node can be obtained through mental projections. (right) Application of Π. Mental states filled in black may have changed as an effect of the belief expansion; gray ones are affected for $kb|_j$ only, with $j \preccurlyeq i$.

Given $\mathfrak{M}^i$, a mental state is a tuple $\langle \mathfrak{M}^i, kb^i \rangle$. It is said to be *legal* iff $kb^i \in KB^i$. Mental models are therefore a belief base, a set of rules for inferring new propositions in that belief base (this will be formalised in detail later), a function for looking at beliefs about a specific individual, and a function for updating beliefs. As discussed at the beginning of this section, we imagine that one agent is able to assign such models to others and itself.

Definition 3 (Agent model). *An* agent model *is the tuple* $ag^i = \langle \mathfrak{M}^i, \mathfrak{A}^i \rangle$, *where* (i) $\mathfrak{M}^i$ *is a mental model and* (ii) $\mathfrak{A}^i = \langle Act, pre \rangle$ *is an action library, where* Act *is a finite set of action labels and* $pre : Act \times KB^i \to \{true, false\}$ *is an action plausibility function that, given an action and a belief base, determines whether the action is plausible;*

The latter will be discussed later. Note that, although this definition adds little to that of a mental model, it is possible to extend it by modelling the agent's ability to observe (how the agent acquires new beliefs through sensors), its mechanism for resolving inconsistencies, etc. For lack of space, these are omitted.

Definition 4 (Agent set). *Given* Ag, *consider the set* $A \subseteq (Ag)^m$, $m > 0$.

We will use these indices to refer to the representation that each agent has of others. For simplicity, we represent these indices as a tree, and will often make use of a tree terminology. As an example, Figure 1 depicts the set of ascribed mental states of Example 1. Given $i, j \in A$, we write $i \preceq j$ iff $j = i \cdot Ag$ (j is a child of i) and, similarly, $i \prec j$ iff $i \not\preceq j$ and $j = i \cdot A$ (j is not a child of, but a descendant of i). Finally, $i \preccurlyeq j$ denotes the fact that either $i \preceq j$ or $i \prec j$.

For example, agent 121 denotes the representation, according to agent 1, that agent 2 has of 1 itself. We regard nested agent labels as regular agent labels, i.e., we refer to the set A instead of Ag. When we need to distinguish, we call agents in Ag *concrete*, and others *virtual*. We assume that A is prefix-closed (i.e., if $i \in A$ then $j \preceq i$ is in A as well) and that $1 \in A$ is the index of the agent we are modelling, and thus $1 \preccurlyeq i$ for any $i \in A$, $i \neq 1$. Indexes i and j quantify over all agents in A (including 1). Also, we assume $1 \cdot Ag \subseteq A$.

2.2 Projections and stereotypes: reasoning *as* and *about* others

Consider two agents i and j, both in A, such that $i \preceq j$ (j is a child of i). Assume for now that they are assigned, respectively, mental models $\mathfrak{M}^i$ and $\mathfrak{M}^j$.

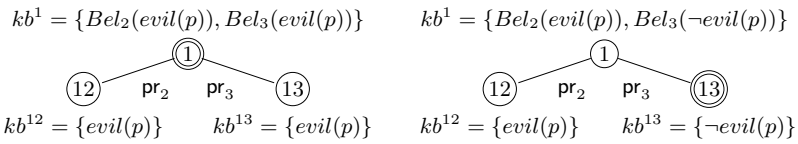

Fig. 2. Two possible evolutions of the situation of Example 3. There is also a third one, reaching a contradiction. Double circled nodes are those used for reasoning. This example anticipates one fundamental point: the reasoning happening at a given node affects the beliefs of children (e.g., left) as well as ancestors (e.g. kb^1, right).

Definition 5 (Ascribed mental state). *Given a mental state* $\mathfrak{S}^i = \langle \mathfrak{M}^i, kb^i \rangle$ *for* i*, the mental state* ascribed *to agent* j *by* i *is* $\mathfrak{S}^j = \langle \mathfrak{M}^j, \mathsf{pr}_j(kb^i) \rangle$*.*

In other words, we just apply the projection function of the target mental model. Intuitively, a mental state ascribed by agent i to j is composed by those (and only those) beliefs that (according to agent i) are possessed by j (together with the target mental model that i assigned to j). This technique allows i to cast itself into agent j and reason as j would (i.e., using $\mathfrak{M}^j$ in place of $\mathfrak{M}^i$). Note how projections can be also used to model different representations of the same phenomena (for example, even dictionaries). Finally, observe that the definition above does not consider the case in which j is not a direct child i, but we can easily take care of this by applying a chain of projections, in the trivial manner.

Definition 6 (Stereotype). *Given a mental model for an agent, a* stereotype *is a rule* $\Phi \Rightarrow \varphi$ *in* Ax *where* φ *(not* Φ*) contains some formula that is* non-local*; that is, the rules reasons about the beliefs of another agent; formally* $\{\varphi\}_\downarrow \neq \{\varphi\}$*. For example, if* φ *is of the form* $Bel_2(\psi)$ *then it is a stereotype* about *agent* 2*.*

Stereotypes allow an agent to reason *about* another agent instead of *as* that agent (that is, by using the projection function to compute the ascribed mental state and reason with it), often with different conclusions.

Example 3. An an example, consider reasoning about two people who are married. A stereotype of a married couple is that they often share similar political beliefs. If we (say, agent 1) believe that person 2 believes that a particular politician p is evil, and we have no belief about this for their spouse, we may model a stereotype rule as $\{Bel_2(evil(p))\} \Rightarrow Bel_3(evil(p))$ in Ax_1, taking advantage of our stereotype. However, it may be in fact that $Bel_3(\neg evil(p))$ is in our belief base, or that by casting ourselves into 3's mind (i.e. projecting our belief base through pr_3 and then reasoning with 3's mental model), we would reach the conclusion that 3 does not share 2's views. If we reason using the stereotype, we may get a different result than if we project what 3 believes. For example, according to the model we assign to 3 (namely, $\mathfrak{M}^{13}$), we may think that agent 3 has the axiom *schema* $\{veg(X)\} \Rightarrow \neg evil(X)$, and we believe that it believes that p is vegetarian. This is illustrated in Figure 2 ($veg(p)$ is omitted).

2.3 Expanding belief bases

Until now, we referred to the reasoning that each agent can perform by using its mental model, and in particular using Ax. The aim of this section is to formally describe how to update a mental state according to Ax: i.e., to add to kb (some) consequences of the beliefs already in kb. In doing so, we restrict the analysis to a single agent, and omit the agent index for readability.

Similar to expert systems, rules in Ax can be used to deduce additional belief, based on the beliefs that are already present in a belief base. A *derivation* of φ from kb by Ax is a finite sequence of deductive steps, each of which is either a formula of $\mathcal{L}$ that is in kb (already believed by the agent) or the result of the application of one rule in Ax. We denote this by writing $kb \vdash_{Ax} \varphi$. Given a deductive system Ax for $\mathcal{L}$ and a belief base kb, let $Cl_{Ax}(kb)$ denote the deductive *closure* of kb, i.e., the set of all consequences derivable from kb by Ax. Formally, $Cl_{Ax}(kb) = \{\varphi \in \mathcal{L} \mid kb \vdash_{Ax} \varphi\}$. Similarly, let $Cl^k_{Ax}(kb)$ denote the *bounded* closure of kb, in which the derivation of φ from kb by Ax is limited in length by k. This bounded version is particularly useful when modelling limited deductive resources: by bounding the length of derivations, we can restrict ourselves to *real* agents, as opposed to *ideal* ones, which are logically omniscient. Note that even when kb is finite, $Cl^k_{Ax}(kb)$ may be infinite if k is infinite.

Definition 7 (Belief expansion). *Given a mental state $\mathfrak{S} = \langle \mathfrak{M}, kb \rangle$, a* belief expansion *of kb wrt $\mathfrak{M}$ is a new belief base kb' that can be obtained by applying this deductive process. Intuitively, kb' is constructed by a derivation π that starts at kb and whose last step produces kb'. Formally, a derivation π can be seen as inducing a sequence of belief bases $\tau_\pi = kb_0, kb_1, \cdots, kb'$ such that $kb_0 = kb$, and $kb_{\ell \geq 1} \in Cl^1_{Ax}(kb_{\ell-1})$. We will denote this by writing $kb' \in Cl^k_{\mathfrak{M}}(kb)$. If $kb \subset kb'$ then the expansion is said to be* proper *(it generated at least one new belief formula). A belief base kb is* closed *wrt $\mathfrak{M}$ if there is no belief expansion of $kb' \in Cl^k_{\mathfrak{M}}(kb)$ such that $kb \subset kb'$. Due to the limitations imposed by KB, more than one closure may exist.*

2.4 Mental systems and successors

Definition 8 (Mental system). *A* mental system *(for agent 1) is a tuple $\Gamma = \{A, \mathcal{L}, \{ag^i\}_{i \in A}, \boldsymbol{k}\}$ where* (i) *A is a set of agent labels as before;* (ii) *$\mathcal{L}$ is the agent's language;* (iii) *$\{ag^i\}_{i \in A}$ is a set of agent models;* (iv) *$\boldsymbol{k}$ is a vector of non-negative integers, with $|\boldsymbol{k}| = |A|$ (which will be used to bound the belief expansion of each agent).*

We require that each KB^i is the product $\times_{i \preceq j}(KB'^j|_j) \times KB^i_{\downarrow}$, where KB'^j is the set $\{kb'^j \mid \mathsf{pr}_i(kb'^j) \in KB^j\}$. This ensures that it is always possible for an agent to build a legal belief base that is able to represent the beliefs of all children (modulo the projection function). This is a natural assumption, as the perspective of the agent representing ancestors (ultimately, agent 1), is always understood. Mental systems can then be designed *bottom-up*, and restructured in case a new mental model is added.

We now describe how a mental system is intended to evolve through belief expansions, as depicted in Figure 1 (right). We do so by defining the operator Π, which can be thought of as a *program specification* that is used to define under which condition a belief base kb' is a "legal extension" of another belief base kb. By "legal extension" we mean that kb' is obtained from kb by applying belief expansions in *some* ascribed mental state, keeping the beliefs of all other agents coherent with their current belief bases. Formally, Π takes a belief base kb and computes the set of belief bases $kb' \in \Pi(kb)$ such that *(i)* $\mathsf{pr}_i(kb') \in Cl_{\mathfrak{M}^i}(\mathsf{pr}_i(kb))$ for some $i \in A$, *(ii)* $\mathsf{pr}_j(kb') = \mathsf{pr}_j(kb)$ for any $j \neq i$ which is not an ancestor or descendant of i (resp., $j \not\preccurlyeq i$ and $i \not\preccurlyeq j$), *(iii)* $\mathsf{pr}_i(kb')_{\downarrow} = \mathsf{pr}_i(kb)_{\downarrow}$ for any ancestor.
Hence, a new belief base kb' is an *extension* of kb iff it can be obtained by a finite iteration of Π, i.e., iff $kb' \in \Pi^n(kb)$, and for some $0 < \ell < n$ we have that $\Pi(\Pi^\ell(kb))$ is proper. The coherency constraint is captured by imposing that, for any pair $i \preceq j$, $\mathsf{pr}_j(\Pi(kb)) = \mathsf{pr}_j(kb)$ implies $\mathsf{pr}_i(\Pi(kb))|_j = \mathsf{pr}_i(kb)|_j$. When this is the case, we say that kb' is a *successor* of kb.

2.5 A procedure for computing successors

The *definition* of Π suggests a *procedure* to compute successors, and hence an algorithm that implements it. The procedure defines a path that updates the tree (implicitly) induced by projecting a belief base. Each step is the result of a (local) belief expansion, a mental projection (parent-child) or inverse projection (child-parent). In the latter, this procedural definition makes sure that, if a node (say j) remains unchanged, then also its representation according to its parent (say i) remains the same (i.e. $\mathsf{pr}_i(kb)|_j$), thus preserving coherence.

Definition 9 (Mental expansion). *A mental expansion σ is a* path*, inside the tree of the mental system Γ, that represents the mental steps of agent* 1 *when it simulates an empathic belief expansion. This shows that the agent in question can direct its attention towards one virtual agent, visiting the corresponding node, unfolding and projecting mental states on demand, and identify a* representation *for the result of this simulated reasoning.*

We can use different bounds k to model the *attention*, or *focus*, we intend to grant to each agent. For example, bystanders in a crisis scenario can be safely ignored (yet modeled), the only relevant description being whether they are or may be interfering with the resolution team.
Due to space limitations we omit the formal definition of *mental expansion*, however, but point to the illustration in Figure 3. The following theorem establishes the correspondence between expansions and successors, and shows that we can always find a mental expansion that "simulates" a possible evolution of the mental system without computing a legal extension at each step (apply Π), but by just computing those mental states visited by the path.

Theorem 1. *Given a legal mental state $\langle \mathfrak{M}^{i_0}, kb_{i_0} \rangle$ there exists an expansion σ from kb_0 to kb_m iff kb_m is a successor of kb_0, with $kb_m \in \Pi\langle ind(\sigma)\rangle(kb_0)$.*

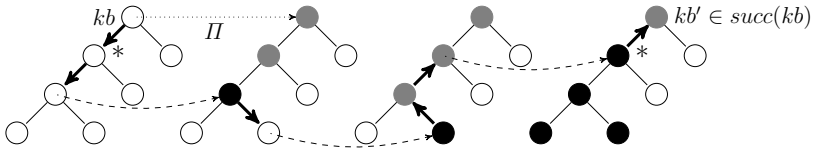

Fig. 3. An example of mental expansion. Colors have the same meaning as in Figure 1. Here, the last belief expansion (*) employed a stereotype about one child, but the same did not happen before, as the agent used the mental projection on that child. Finally, note that gray nodes are not unique in general, but only one is computed, if on σ.

Here, $ind(\sigma)$ denotes the sequence of agent labels of the expansion σ and $\Pi\langle ind(\sigma)\rangle$ a finite number of applications of Π: specifically, one in which the mental states that are expanded are those in $ind(\sigma)$.

3 About simulating plausible evolutions

In this section, we briefly comment on how to incorporate our approach into known settings for modelling and analysing multi-agent systems, or *games* [3], and use the agent models to foresee action deliberation, and thus physical evolutions. One fundamental advantage of using our framework is that we will preserve a first-person view. We imagine that the agent is capable of *simulating* the game "in its mind", by analysing all agent models together with an *approximation* of the environment, to foresee collective evolutions. This game is not *real*, but can be used to retrieve *plausible* strategies. An action α is *plausible* for an agent $ag = \langle \mathfrak{M}, \mathfrak{A}\rangle$ with belief base kb iff *(i)* $\alpha \in Act$ and *(ii)* $pre(\alpha, kb) = true$. Similarly, we can define the plausibility of a vector of actions, one for each concrete agent, by inspecting the ascribed mental state of each.
An environment is a finite state machine that evolves depending on the action chosen by all agents, typically synchronously. A possible evolution of the environment (a sequence of environment's states) is plausible iff it can be the result of a sequence of plausible action vectors. By expanding agent models (Defn. 3) to account for perceiving capabilities, and by looking at plausible evolutions, it is possible for the agent to retrieve the observations that other agents may have of the simulated environment, update their ascribed mental states accordingly, and repeat the process. By iterating this procedure, we can build a *finite-state* representation of the system, and restrict the analysis of possibly large-scale problems by focusing on plausible evolutions only. It is then possible to adopt existing verification and synthesis techniques (see, e.g., [3]) to verify properties of such games as well as synthesizing agent plans that are guaranteed to satisfy certain properties.

4 Conclusions and future work

In this paper, we proposed an approach for modelling the beliefs of one agent about the environment and other agents, as well as the mental model(s) it assigns

to itself and others. In future work, we plan to improve the notion of agent models via the abstraction of a finite set of relevant belief configurations based on [1, 3], and also to model notions that are not local to a specific agent, but to the social reality and practices [5]. We are also interested in studying dynamic assignment of agent models, to reflect the dynamics of reality. To this aim, we will use this approach to alternate between simulation and actual execution to obtain heuristics/plan fragments rather than complete strategies, as the significance of the simulated game decreases when the "noise" introduced by the model's inaccuracy increases.

Acknowledgements. This research was funded by Australian Research Council Discovery Grant DP130102825.

References

1. N. Alechina and B. Logan. A logic of situated resource-bounded agents. *J. of Logic, Language and Information*, 18(1):79–95, Jan. 2009.
2. G. Aucher. Internal models and private multi-agent belief revision. In *Proceedings of AAMAS 2008*, pages 721–727, 2008.
3. G. De Giacomo, P. Felli, F. Patrizi, and S. Sardiña. Two-player game structures for generalized planning and agent composition. In *AAAI*, 2010.
4. J. Dias and A. Paiva. Feeling and reasoning: A computational model for emotional characters. In *EPIA*, pages 127–140, 2005.
5. F. Dignum, G. J. Hofstede, and R. Prada. From autistic to social agents. In *AAMAS*, pages 1161–1164, 2014.
6. H. v. Ditmarsch, W. van der Hoek, and B. Kooi. *Dynamic Epistemic Logic*. Springer Publishing Company, Incorporated, 1st edition, 2007.
7. L. E. Hall, S. Woods, R. Aylett, and A. Paiva. Using theory of mind methods to investigate empathic engagement with synthetic characters. *I. J. Humanoid Robotics*, 3(3):351–370, 2006.
8. M. Johnson, J. M. Bradshaw, P. J. Feltovich, C. M. Jonker, M. B. van Riemsdijk, and M. Sierhuis. Coactive design: Designing support for interdependence in joint activity. *J. of Human-Robot Int.*, 3(1):43–69, 2014.
9. S. Lemaignan, R. Ros, L. Msenlechner, R. Alami, and M. Beetz. Oro, a knowledge management module for cognitive architectures in robotics. In *IROS 2010*, 2010.
10. A. Pearce, L. Sonenberg, and P. Nixon. Toward resilient human-robot interaction through situation projection for effective joint action. In *Robot-Human Teamwork in Dynamic Adverse Environment: AAAI Fall Symp.*, pages 44–48, 2011.
11. J. Pfau, Y. Kashima, and L. 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.
12. Y. M. Ronald Fagin, Joseph Y. Halpern and M. Y. Vardi. *Reasoning about Knowledge*. MIT Press, Cambridge, MA, 1995.
13. B. Scassellati. Theory of mind for a humanoid robot. *Auton. Robots*, 12(1):13–24, 2002.
14. R. Stocker, L. A. Dennis, C. Dixon, and M. Fisher. Verifying brahms human-robot teamwork models. In *JELIA*, pages 385–397, 2012.
15. M. Warnier, J. Guitton, S. Lemaignan, and R. Alami. When the robot puts itself in your shoes. managing and exploiting human and robot beliefs. In *Proc. of the 21th IEEE Int. Symp. in Robot and Human Interactive Communication*, 2012.