Mutual Modelling in Robotics: Inspirations for the Next Steps

Séverin Lemaignan, Pierre Dillenbourg
Computer-Human Interaction in Learning and Instruction Laboratory (CHILI)
École Polytechnique Fédérale de Lausanne (EPFL)
CH-1015 Lausanne, Switzerland
firstname.lastname@epfl.ch

ABSTRACT

Mutual modelling, the reciprocal ability to establish a mental model of the other, plays a fundamental role in human interactions. This complex cognitive skill is however difficult to fully apprehend as it encompasses multiple neuronal, psychological and social mechanisms that are generally not easily turned into computational models suitable for robots.

This article presents several perspectives on mutual modelling from a range of disciplines, and reflects on how these perspectives can be beneficial to the advancement of social cognition in robotics. We gather here both basic tools (concepts, formalisms, models) and exemplary experimental settings and methods that are of relevance to robotics.

This contribution is expected to consolidate the corpus of knowledge readily available to human-robot interaction research, and to foster interest for this fundamentally cross-disciplinary field.

1. INTRODUCTION

Human social dynamics rely upon the ability to effectively attribute beliefs, goals and percepts to other people. This set of meta-representational abilities shapes what is called a theory of mind (ToM) or the ability to mentalize, and leads to mutual modelling: the reciprocal ability to establish a mental model of the other. This lays at the core of human interactions: normal human social interactions depend upon the recognition of other sensory perspectives, the understanding of other mental states, and the recognition of complex non-verbal cues of attention and emotional state. As such, adapting and transferring these cognitive skills to social robots is an important research objective.

Until now, however, the human-robot interaction (HRI) community has only scratched the surface: in [41], Scassellati gave an initial account of Leslie's and Baron-Cohen's respective models of the emergence of a theory of mind (we discuss them below) from the perspective of robotics, but reported implementation work was limited to simple perceptual precursors (like face detection or color saliencies detection). Since then, research in this field has been focused on applications relying on Flavell's *Level 1* [20] perspective-taking, *i.e.* perspective-taking that only requires perceptual abilities ("*I see (you do not see the book)*"), and actually mostly limited to visual perception (relevant work include Breazeal [10], Trafton [47] and Ros [38]).

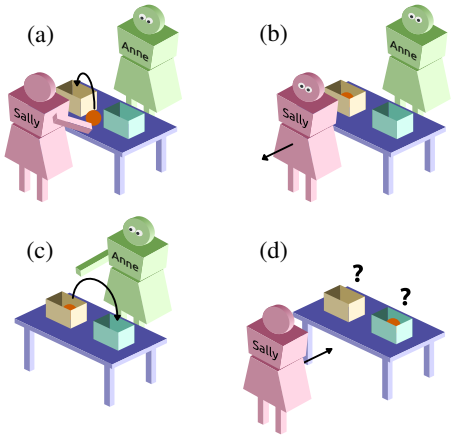


Figure 1: The false belief experiment: two puppets, "Anne" and "Sally" face each other, with two boxes between them. A child (the subject) observes. Sally puts a ball in the beige box and then leaves. While she is absent, Anne moves the ball to the blue box. Sally returns. The experimenter asks the child: *Where Sally would look for the ball?* Without a theory of mind, the child is not able to ascribe false beliefs to Sally, and therefore incorrectly answers *In the blue box.* Visual perspective taking only is sufficient to pass this task.

Based on perspective taking *Level 1* alone, Breazeal *et al.* [11] and Warnier *et al.* [54] successfully tackled the classical hallmark of theory of mind, the false belief experiment (also known as the "Sally and Anne" experiment, Figure 1, introduced by [55], original experimental setting by [6]). They demonstrated complete human-robot interaction scenarios where robots recognize and handle false belief situations in dyadic or triadic interactions, and exhibit helping behaviours that account for the missing/false beliefs of the human partners.

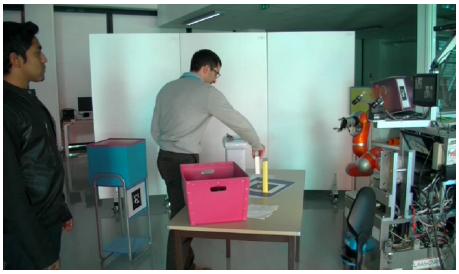
Those are significant achievements, also reassuring as to endow our robots with advanced socio-cognitive capabilities (Figure 2). However, one intuitively recognizes that mutual modelling goes indeed beyond computing what the human sees or does not see. Perceptions translate into subjective

HRI '15, March 02 - 05 2015, Portland, OR, USA

ACM 978-1-4503-2883-8/15/03 ...$15.00.
http://dx.doi.org/10.1145/2696454.2696493

(a) A robot interpreting situated multimodal dialogue [9]

(b) A robot changing mental perspective during social learning [11]



(c) False belief task, where the robot tracks which objects are visible to which human [54]

Figure 2: Examples of social robotics tasks benefiting from meta-cognition

representations: how can we access them? How to measure what the other know about oneself? Many levels of reciprocal modelling overlap, with endless "I know that you know that I know that...": how to represent and manipulate them? What about the breadth of the models we build? Local to a given task or broader, deeper? How to tell apart mimicry from cognitive modelling in a joint action? Those many questions underline the complexity of a cognitive mechanism whose study lays at the crossroad of several academic fields. Robotics, as the science of embodied artificial intelligence, may be a convergence point to validate our understanding of this socio-cognitive skill.

This article contributes to this endeavour by proposing a broad overview of mutual modelling from three different perspectives in three different academic domains. First, *developmental psychology* (and, in particular, developmental *patho*psychology) provides insights on human cognition and an extensive experimental framework that has a strong potential in robotics. *Psycho-linguistic* and *Collaborative Learning*, then, propose concepts and tools to apprehend the importance and the dynamics of knowledge sharing during social interactions. Finally, we give a look at the philosophy of the mind and the logical tools of *formal epistemology*: their use of modal logics to describe complex knowledge manipulation in groups appears as a pertinent direction for robotics to explore.

Conceptual bridges between these disciplines are numerous, and our aim is here to draw a picture of the field that can lay the bases for the further advancement of cognitive robotics in a social environment.

Note that the main focus of this contribution is *epistemic* mutual modelling: the modelling of (mostly declarative) knowledge of and by the others. As hinted at the beginning of this introduction, mutual modelling in general covers more: what are our feelings towards each other? What are our respective goals? What are our respective moods? etc. Those questions have also started to be explored in robotics. *Human-aware task planning* [1, 12], for instance, uses symbolic task planners to plan actions not only for the robot, but also for the interacting humans, hence trying to predict, *i.e.* model, what (rational) humans may do in the given situation. From a completely different perspective, Fink *et al.* [19] propose cognitive models of how the human perceive robot behaviours over time. Those two very different yet prototypical examples give an idea of the breadth of the "mutual modelling question".

2. MUTUAL MODELLING AND DEVELOPMENTAL PSYCHOLOGY

Connections vs. representations.

In [21], Flavell relates perspective taking *Level 1* to establishing *cognitive connections* (I see, I hear, I want, I like, I fear...), in contrast to perspective taking *Level 2* that relates to manipulating *representations*. This is exemplified by *appearance-reality* tasks, like the *elephant mask* experiment proposed in [21]: 3-years old children are not able to tell that an experimenter hidden behind a large elephant mask but who speaks normally *looks* like an elephant, *sounds* like the experimenter, and *really is* the experimenter. It appears that, while those children are able to explicitly manipulate cognitive connections (they know for instance that these are largely independent of each other and that they can evolve over time) and know as well that their own connections are independent of those of other people, they do not think that one concept can *seriously* (*i.e.* non playfully) hold several, possibly conflicting, representations.

This *connection-representation* account appears to be a significant component of a general theory of mind (one needs to recognize that the same object/concept may have different, serious, representations to then accept false beliefs for instance). Figure 3 illustrates this difference between cognitive connections and representations in an imaginary human-robot interaction scenario. The *visual* perspective of the baby and the mother are represented: a robot endowed with perspective-taking level 1 is able to compute that the baby looks at the plug and the mother looks at the baby. *Representation*-level perspective taking, on the other hand, would require the robot to represent what the socket means to the baby (an attractive affordance), and what the baby's behaviour represents to the mother (a potential danger).

Developmental pathopsychology.

The false belief experiment that we have mentioned above, was proposed by Baron-Cohen in the frame of his research on autistic spectrum disorders (he shows that autistic children seem to actually lack a theory of mind and suggests this as the primary cause of their social impairments), and Frith and Happé further note in [23] that this specific deficit of autism has led to a large amount of research which proved, in turn, highly beneficial to the study of the development of theory of mind in general. They reference in [23] eight such tasks (Table 1), identified during the study of social cogni-

Figure 3: *Visual* perspectives allow for a first level of mutual modelling. However, to correctly comprehend the scene (and for the robot to adequately react, *representation-level* perspective taking is required: what does the power socket means to the baby? What does the situation means to the mother?

tion by autistic children. Each of them is proposed in two versions: one does not require mentalizing, while the other does require it. One of these tasks, for example, required children to distinguish emotions, namely happy/sad faces on one hand (*situation-based* emotion), and surprised faces on the other (*belief-based* emotion) [8]. Another task, based on the *penny-hiding game*, contrasts the two conditions in terms of *object occlusion* vs. *information occlusion* [5] (we detail it hereafter). These tasks prototypically illustrate social metacognition: one need to represent and reflect on someone else representations (and not only perceptions), and they are not addressed by today's research on social robots.

Experimental protocols in research on autistic spectrum disorders are often striking by their apparent straightforwardness because of the careful choice of interaction modalities: since autistic children frequently exhibit impairments beyond social ones (such as motor or linguistic ones), the experiments must be designed such that they require only basic cognitive skills beyond the social abilities that are tested. The Sally and Anne task, for instance, requires the observing child to be able to visually follow the marble, to remember the true location of the marble, to understand simple questions ("Where will Sally look for her marble?" in Baron-Cohen's protocol [6]) and eventually to give an answer, either verbally or with a gesture – the two first points being actually explicitly checked through questions: "Where is the marble really?" (reality control question) and "Where was the marble in the beginning?" (memory control question).

Likewise, current social robots have limited cognitive skills (no fast yet fine motor skills, limited speech production and understanding, limited scene segmentation and object recognition capabilities, etc.) and such tasks that effectively test a single cognitive skill (in this case, mentalizing) in near isolation are of high relevance for experimental social robotics.

Frith and Happé's list (Table 1) is in that regard especially interesting in that it mirrors pairs of task (ones which do not require mentalizing with similar ones which do require mentalizing), thus providing control tasks. *Object occlusion* vs. *Information occlusion* is one example of a (pair of) task(s) which evidence representation-level per-

No mentalizing required	Mentalizing required
Ordering behavioural pictures	Ordering mentalistic pictures [7]
Understanding see	Understanding know [36]
Protoimperative pointing	Protodeclarative pointing [4]
Sabotage	Deception [44]
False photographs	False beliefs [29]
Recognizing happiness and sadness	Recognizing surprise [8]
Object occlusion	Information occlusion [5]
Literal expression	Metaphorical expression [24]

Table 1: Tasks requiring or not mentalizing to pass, listed by Frith and Happé in [23]

spective taking through *adaptive deception*: during a simple game, the experimenter adapts its strategy (deceptive/non-deceptive behaviour) to the representation skills of its child opponent. The experimental setting is derived from the penny-hiding game protocol originally proposed by Oswald and Ollendick [34] and replicated and extended by Baron-Cohen in [5], who describes it as a two-person game in which the subject is actively involved, either as a guesser or as a hider. The hider hides the penny in one hand or the other, and then invites a guess. The game is repeated several time before switching the roles. Baron-Cohen proposes a specific index to rate the level of the players based on the idea of *information occlusion*: minimally, the hider must ensure *object occlusion* (the penny must not become visible to the guesser), while good hiders, with representation-level perspective taking skills, develop strategies (like random hand switching or deictic hints at the wrong hand) to prevent the guesser to find the penny (*information occlusion*). One could imagine a similar protocol adapted to robotics: the robot would play the role of the experimenter, adapting online its behaviour to what it understands of the perspective taking capabilities of the children, and would consequently require *second-order*, *representation-level* perspective taking from the robot.

Higher-order Theory of Mind.

While a great deal of research concerns itself with *first-order* theory of mind, *higher-order* (and particularly, second-order) ToM are also studied. Verbrugge and Mol [53] describe the different levels in the following terms:

> To have a first-order ToM is to assume that someone's beliefs, thoughts and desires influence one's behavior. A first-order thought could be: *He does not know that his book is on the table.* In second-order ToM it is also recognized that to predict others' behavior, the desires and beliefs that they have of one's self and the predictions of oneself by others must be taken into account. So, for example, you can realize that what someone expects you to do will affect his behavior. For example, "(I know) he does not know that I know his book is on the table" would be part of my second-order ToM. To have a third-order ToM is to assume others to have a second-order ToM, etc.

Perner shows in [35] that 2^{nd}-order ToM is mastered around 8 years old, and Flobbe *et al.* propose in [22] a set of three tasks (a second-order false belief task, a strategic game and a sentence comprehension test) that require second-order mentalizing to succeed. The second-order false belief task that they propose (known as the *Chocolate bar task*) effectively evidence higher-order ToM:

> John and Mary are in the living room when their mother returns home with a chocolate bar that she bought. Mother gives the chocolate to John, who puts it into the drawer. After John has left the room, Mary hides the chocolate in the toy chest. But John accidentally sees Mary putting the chocolate into the toy chest. Crucially, Mary does not see John. When John returns to the living room, he wants to get his chocolate.

Flobbe then asks the subjects: *"Where is the chocolate now?"* (reality control question), *"Does John know that Mary has hidden the chocolate in the toy chest?"* (first-order ignorance question), *"Does Mary know that John saw her hide the chocolate?"* (linguistic control question), and *"Where does Mary think that John will look for the chocolate?"* (second-order false belief question). Besides, Flobbe asks the participants to justify their answer (*"Why does she think that?"*). In her study, 82% of a group of 40 children (M=9 year old) successfully passed the task.

While literature on higher-order of mutual modelling is generally scarce, *agreement* and *common belief* is another interesting social situation: Verbrugge [52, p. 664] reports after an experiment by Mant and Perner [32] where a child is disappointed by his father who changed the announced plan to go swimming. In one condition, the child and the father had previously mutually agreed, while in the other, no explicit agreement took place (to a child observer, it actually appears that the situation is **worse** if the child and the father did **not** previously explicitly agree). Children before ten do not distinguish between the two conditions, and Verbrugge's proposed explanation relies on the concept of *social commitment*, which implies the *common belief* between the two agents that the father *intends* to go swimming and the child is *interested* in going swimming.

Common belief ("we believe that we believe that we believe that... we agreed") is defined in epistemic logic (see section 4) as an infinite recursion ("∞-order" ToM), and Verbrugge suggests that this mutual modelling mechanism is therefore harder to master for children than 2^{nd}-order ToM for instance.

3. MUTUAL MODELLING IN PSYCHOLINGUISTICS AND COLLABORATIVE LEARNING

A support for shared understanding.

Computer Supported Collaborative Learning (CSCL) researches the cognitive mechanisms and practical techniques underpinning efficient learning in social situations. From its very beginning, CSCL research has been following Roschelle and Teasley's suggestion [39] that collaborative learning has something to do with the process of constructing and maintaining a *shared understanding* of the task at hand. Building a shared/mutual understanding refers to the upper class of collaborative learning situations, those in which students should build upon each other's understanding to refine their own understanding. What is expected to produce learning is not the mere fact that two students build the same understanding but the cognitive effort they have to engage to build this shared understanding [43].

The construction of a shared understanding has been investigated for several years in psycholinguistics, under the notion of *grounding*[1] (Clark, in [15]). However, the relevance of grounding mechanisms for explaining learning outcomes has been questioned in learning sciences. The monitoring and repair of misunderstanding explains for instance referential failures in short dialogue episodes but does hardly predict *conceptual change* (*i.e.* the acquisition, acceptation and integration of a new belief into one's mental model) over longer sessions [17]. The cumulative effect of grounding episodes can probably be better understood from a sociocultural perspective:

> Collaborative learning is associated with the increased cognitive-interactional effort involved in the transition from *learning to understand each other* to *learning to understand the meanings of the semiotic tools that constitute the mediators of interpersonal interaction* [3]

Along this line, several scholars suggest that CSCL research should go deeper in the understanding of how partners engage into shared meaning making [45] or *intersubjective* meaning making [46].

Paradoxically, while Clark's theory is somewhat too linguistic from a conceptual change viewpoint, it is criticized at the same time as being too cognitivist by some psycholinguists, *i.e.* as overestimating the amount of shared knowledge and mutual representations actually necessary to conduct a dialogue. The fundamental issue, as old as philosophy, is the degree of coupling between the different levels of dialogue, mostly between the lexical/syntactical level and the deeper semantic levels. In [37], Pickering and Garrod argue that the mutual understanding starts mostly with a *superficial alignment* at the level of the linguistic representations, due to priming mechanisms, and that this local alignment may – in some cases – lead to a *global alignment* of the semantic level (*deep grounding*). For these authors, the convergence in dialogue, and even the repair of some misunderstandings, is explained by this mimetic behavior more than by a monitoring of each other's knowledge: *"...interlocutors do not need to monitor and develop full common ground as a regular, constant part of routine conversation, as it would be unnecessary and far too costly. Establishment of full common ground is, we argue, a specialized and non-automatic process that is used primarily in times of difficulty (when radical misalignment becomes apparent)."* [37] This view is actually not incompatible with Clark's *grounding criterion* [14]: the degree of shared understanding that peers need to reach depends upon the task they perform. For instance, a dialogue between two surgeons might rely on superficial alignment if they talk about their friends but has to guarantee accurate common grounds when talking about

[1]Note that the meaning of *grounding* – ensuring a shared understanding of a situation during an interaction – that we employ in this article must be distinguished from its meaning in the context of *symbol grounding* as defined by Harnad [25].

which intervention will be conducted in which way on which patient.

Deep grounding or shared meaning making requires some cognitive load. For Clark, what is important is not the individual effort made by the receiver of a communicative act, but the overall *least collaborative effort* [15]. The cost of producing a perfect utterance may be higher than the cost of repairing the problems that may arise through misunderstandings. For instance, subjects are less careful about adapting their utterances to their partner when they know they can provide feedback on his/her understanding [42]. Dillenbourg *et al.* introduced the notion of *optimal collaborative effort* [16] to stress that misunderstanding should not be viewed as something to be avoided (if this was possible), but as an opportunity to engage into verbalization, explanation, negotiation, and so forth.

CSCL model of mutual modelling.

Dillenbourg proposes in [40] a model to represent mutual modelling situations. He uses the notation $\mathcal{M}(A, B, X)$ to denote "A knows that B knows X" (equivalent to the epistemic logic notation $\mathsf{K}_A\mathsf{K}_B X$ that we present in the next section). This notation does not mean that A has an explicit, monolithic representation of B: it must be understood as an abstraction referring to possibly complex socio-cognitive processes. Besides, he refer to the *degree of accuracy* of the model as $\mathcal{M}^\circ(A, B, X)$.

He parametrizes and assesses the mutual modelling *effort* through 3 variables:

1. Tasks vary a lot with respect to how much they require mutual understanding. The *grounding criterion* [15] $\mathcal{M}^\circ_{min}$refers to how important it is to mutually share a piece of information X to succeed the task T. It can be computed as the probability to succeed T despite the fact X is not grounded. $\mathcal{M}^\circ_{min}(A, B, X)$ can be estimated from the correlation between $\mathcal{M}^\circ(A, B, X)$ and the task performance.
2. Before any specific grounding action, there is usually a non-null probability that X is mutually understood by A and B (*e.g.* X is part of A's and B's cultures, it is manifest to co-present subjects or simply there is not much space for misunderstanding or disagreement about X). He notes the theoretical accuracy of initial grounds $\mathcal{M}^\circ_{t_0}(A, B, X)$.
3. The cost of grounding X refers to the physical and cognitive effort required to perform a grounding act α: a verbal repair (*e.g.* rephrasing), a deictic gesture, a physical move to adopt one partner's viewpoint, etc. This cost varies according to media features [13].

These notations lead to simple representations of mutual modelling during interactions, and Dillenbourg derives several questions out of this model. Adapted to a human-robot interaction situation, Figure 4 represents for instance a dyadic interaction (we use H to denote a human, while R stands for a robot). Δ_1 illustrates what Dillenbourg calls the *symmetry* question (*Is the accuracy of my model related or not to the accuracy of your model?*).

With triads (two humans H_1 and H_2 and a robot R), we may compute the accuracy of 6 models $\mathcal{M}^\circ(H_1, H_2, X)$, $\mathcal{M}^\circ(H_2, H_1, X)$, $\mathcal{M}^\circ(H_1, R, X)$, $\mathcal{M}^\circ(R, H_1, X)$, $\mathcal{M}^\circ(R, H_2, X)$ and $\mathcal{M}^\circ(H_2, R, X)$.

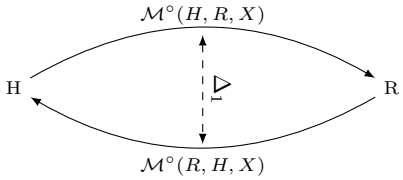


Figure 4: Mutual modelling in a dyadic interaction, $\Delta_1 = \Delta(\mathcal{M}^\circ(H, R, X), \mathcal{M}^\circ(R, H, X))$

This leads to two *triangle questions* relevant to HRI (Figure 5): Do H_1 and H_2 have the same accuracy when modelling the robot R? ($\Delta_2 = \Delta(\mathcal{M}(H_1, R, X), \mathcal{M}(H_2, R, X))$), and conversely, what may lead R to model more accurately H_1 or H_2? ($\Delta_3 = \Delta(\mathcal{M}(R, H_1, X), \mathcal{M}(R, H_2, X))$).

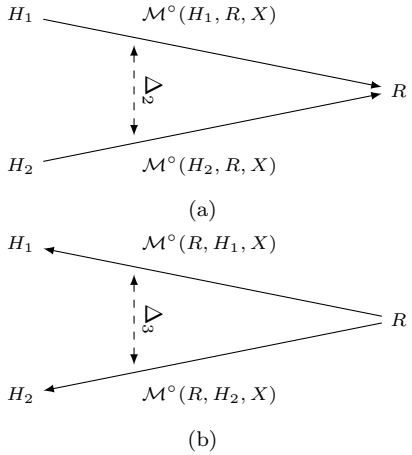


Figure 5: Mutual modelling in a triadic interaction

Finally, Dillenbourg also suggests a *rectangle question*: how self- versus other modelling compares (Δ_4 in Figure 6)? This gives an indication of meta-cognitive skills of the agents. We can also question if the modelling skills depend upon what aspects are being modeled (X or Y) which would explain vertical differences (Δ_5 in Figure 6).

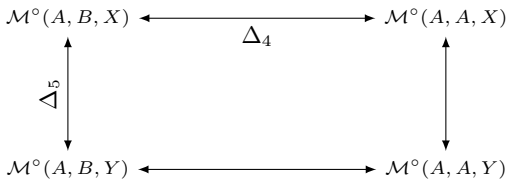


Figure 6: Meta-cognitive skills and domain-dependent modelling

This model, designed in the context of human collaboration, evidences questions that are relevant as well to human-robot interactions.

4. FORMAL EPISTEMOLOGY

The above model of mutual modelling is meant as a practical tool to reason on knowledge dynamics in group interactions and it does not look at being a formal model, whereas formal epistemology, a subfield of the philosophy of mind, focuses on this question.

Modal logics look at the formal representation of *possible worlds*, *i.e.* the *possibility* or *necessity* of certain assertions to hold, and is naturally suited to build mathematical representations of situations such as "*the robot knows [the baby may not know what a power socket is]*".

The *epistemic modal logic* in particular (see [26] for an overview and references) focuses on the formal representation of knowledge and beliefs of agents, with the operators $\mathsf{K}_i\varphi$ (epistemic operator: agent i knows φ) and $\mathsf{B}_i\varphi$ (doxatic operator: agent i believes φ). Every possible logical propositions belong then to possible *worlds* (noted w), that are *accessible* (*i.e.* compatible) or not to one's beliefs and knowledge.

Single-agent epistemic systems can naturally extend to multi-agent systems [18, chapt. 4]: if p stands for "the power socket is dangerous", $\mathsf{K}_{mother}p \wedge \mathsf{K}_{mother}\neg\mathsf{K}_{baby}p$ states that the mother knows that the socket is dangerous, and also knows that the baby is not aware of this. This provides a formal tool to represent mutual models (the *order* of mutual modelling as discussed in the context of developmental psychology is here directly related to the nesting depth of the epistemic operator).

This approach has led to applications to the representation of knowledge dynamics on concrete, albeit arguably toy, scenarios: van Ditmarsch presents for instance in [51] the formal description of possible Cluedo strategies based on what players know about other players' knowledge, and along the same line, Verbrugge and Mol analyse mutual modelling in a strategic game with imperfect information (derived from Mastermind) in [53].

Amongst the several *modal operators of knowledge* that have been developed, the *common-knowledge* operator CK is of particular interest. If we define the *shared-knowledge* operator EK as $\mathsf{EK}_J\varphi \leftrightarrow \bigwedge_{i\in J}\mathsf{K}_i\varphi$, *i.e.* φ is *shared knowledge* amongst the group J iff every agent in J knows φ, then $\mathsf{CK}_J\varphi \leftrightarrow \mathsf{EK}_J\varphi \wedge \mathsf{EK}_J\mathsf{EK}_J\varphi \wedge \mathsf{EK}_J\mathsf{EK}_J\mathsf{EK}_J\varphi \wedge ...$, *i.e.* φ is shared knowledge, and it is also shared knowledge that φ is shared knowledge, etc. (this presentation follows [27]). This illustrates how epistemic logic can represent non-trivial social knowledge situations.

Verbrugge further investigates the social aspect of epistemic logics in [52] and proposes a survey of epistemic logic applications to *social reasoning*. He underlines both the limits of epistemic logic for that purpose (common epistemic systems assume for instance $\mathsf{K}_i\varphi \rightarrow \mathsf{K}_i\mathsf{K}_i\varphi$, which reads "$i$ knows φ" implies "i knows that i knows φ", *i.e.* i can always introspect, a rather idealized model of human cognition) and the recent advancement towards modelling *human* social cognition, which implies for instance limited rationality. One of these attempts is formalized as a *doxastic epistemic logic* by van Ditmarsch and Labuschagne in [49], with an explicit focus on modelling *theory of mind* mechanisms. This model builds upon *dynamic epistemic logic* [50] (DEL, epistemic logics augmented with mechanisms for knowledge changes), and the modelling of agents' degrees of belief through a *preference* accessibility relation.

The mathematical objects build from these different modal logics are natural candidates for transposition into representational systems and controllers for robots. Historically in robotics, the main research perspective has been towards the *action logics*, and in particular the influential *situation calculus* (a propositional logic initially proposed by McCarthy, and fully axiomatized in the context of robotics by Levesque *et al.* in [30], which led to the GOLOG logic programming language [31]). Many other action logics have been proposed including modal logics like PDL (*Propositional Dynamic Logic*).

Recent efforts have focused on bridging action logics (that deal with *ontic* actions, *i.e.* actions which have tangible, physical consequences) with epistemic logics (that deal with *epistemic* actions, *i.e.* knowledge changes). Van Ditmarsch proposes in [48] for instance a solution to embed a practical subset of situation calculus into a dynamic epistemic logic, and Herzig provides in [27] a broader overview of the interplay between current action and epistemic logics.

From a practical perspective however, implementations of these logics into practical reasoners or programming languages remain rare. The development of *Description Logics* (DL) in the knowledge representation community, along with effective, practical tools (like reasoners) is a possible path forward, since DL semantics overlap to some extend with modal logics [2, chap. 4.2.2], and *Description Logics* have already been successfully used in robotics (see [28] for a review).

5. CONCLUSION: SCAFFOLDING FOR SOCIO-COGNITIVE ROBOTICS

Several lessons can be taken out of these three perspectives on mutual modelling. First, **concepts** and terminology stand out and social robotics would benefit from incorporating them. Second, **models** of mutual modelling exist that would make sense in robotics as well, along with investigation strategies and approaches that translate well to robotics. Finally, we believe that several **experimental settings** designed and successfully tested in other disciplines shape an interesting way forward for (experimental) social robotics.

Flavell's distinction between *cognitive connections* on one hand, and *mental representations* on the other hand helps in recognizing the limits of perceptual perspective taking as currently achieved in robotics. This *connection-representation* account, put into perspective with the semantic expressiveness of modal logics, leads to a first insight: we need to come up with an effective design for a meta-representational system (*i.e.* a system that *represents representations* as abstracted, manipulable entities) to integrate into homogeneous objects both cognitive connections (*i.e.* percepts) and representations (including suppositions, similar to the idea of *pre-supposition accommodation* [33]).

The depth of mutual modelling required for effective collaboration is another question that is relevant to HRI. Pickering and Garrod, with the idea of *superficial alignment* versus *global alignment* or *deep grounding*, come to the conclusion that *mimicking* behaviours is often a more efficient way to work together than establishing a full common ground, which is also expressed by Clark in terms of *least collaborative effort* and Dillenbourg in terms of *optimal collaborative effort*: misunderstandings should not be viewed as something to systematically avoid since the repair actions they may elicit can also be viewed as a way to engage the agents into new interactions (an idea not unrelated to the *entropy* concept in information theory). And Clark's concept of *grounding criterion* provides a practical tool to represent and manipulate this idea of degree of shared understanding required by the agents to perform a given task.

These considerations lead to a second insight: the level of *mutual grounding* (*i.e.* of mutual understanding) of a given situation of interaction may prove to be a valuable metric to measure the quality of human-robot interactions. Mutual modelling certainly plays a key role here, but lower-level, sub-cognitive strategies like one-off mimicking may be powerful complements and need to be investigated in parallel.

Then, the mere discussion of mutual modelling problems ("The robot knows that we both know that it knows that...") is also by itself a challenge. Formal models like the ones developed in epistemic logic, or practical tools like the ones stemming from CSCL are valuable objects to add to the toolset of HRI. Not only they help clarifying the problems, but they also evidence new challenges: the definition of common-knowledge as $\mathsf{CK}_J\varphi \leftrightarrow \mathsf{EK}_J\varphi \wedge \mathsf{EK}_J\mathsf{EK}_J\varphi \wedge \mathsf{EK}_J\mathsf{EK}_J\mathsf{EK}_J\varphi \wedge ...$ for instance points at the recursive nature of common knowledge. While "common-sense knowledge" is a commonly-used term in the field of knowledge representation and reasoning to describe "all those facts that everyone knows about", it is probably important to better distinguish in the future between *shared knowledge* and *common knowledge* when a robot interact with a human.

A third insight that follows relates to the adoption of modal logics in HRI: we believe that the current state of the technology in robotics (with the advancements in scene understanding, situation assessment, human activity recognition, and their subsequent symbolic grounding), combined with our current understanding of modal logics (and epistemic logics in particular), lay the foundations for us to now broadly embrace formal models to implement new, complex social mechanisms in robots.

Lastly, we hope that this article may suggest new ideas of experimental protocols and studies to conduct in HRI. We found experimental research in developmental psychology to be especially inspiring in that regard: the numerous protocols that have been designed over the years to evidence cognitive and social skills by children (and even more, cognitively impaired children) stand out as a source of inspiration for cognitive robotics, and this article hopefully shed some light on some of the less well-known studies and tasks that could become new milestones for the development of human-robot interaction.

Acknowledgments

This research was supported by the Swiss National Science Foundation through the National Centre of Competence in Research Robotics.

We would also like to thank Tony Belpaeme for his feedback on an early version of this article.

6. REFERENCES

[1] R. Alami, A. Clodic, V. Montreuil, E. A. Sisbot, and R. Chatila. Toward human-aware robot task planning. In *AAAI Spring Symposium: To Boldly Go Where No Human-Robot Team Has Gone Before*, pages 39–46, 2006.

[2] F. Baader, D. Calvanese, D. L. McGuinness, D. Nardi, and P. F. Patel-Schneider, editors. *The Description Logic Handbook: Theory, Implementation and Applications*. Cambridge University Press, New York, NY, USA, 2nd edition, 2003.

[3] M. Baker, T. Hansen, R. Joiner, and D. Traum. The role of grounding in collaborative learning tasks. *Collaborative learning: Cognitive and computational approaches*, pages 31–63, 1999.

[4] S. Baron-Cohen. Perceptual role taking and protodeclarative pointing in autism. *British Journal of Developmental Psychology*, 7(2):113–127, 1989.

[5] S. Baron-Cohen. Out of sight or out of mind? another look at deception in autism. *Journal of Child Psychology and Psychiatry*, 33(7):1141–1155, 1992.

[6] S. Baron-Cohen, A. Leslie, and U. Frith. Does the autistic child have a "theory of mind" ? *Cognition*, 1985.

[7] S. Baron-Cohen, A. M. Leslie, and U. Frith. Mechanical, behavioural and intentional understanding of picture stories in autistic children. *British Journal of developmental psychology*, 4(2):113–125, 1986.

[8] S. Baron-Cohen, A. Spitz, and P. Cross. Do children with autism recognise surprise? a research note. *Cognition & Emotion*, 7(6):507–516, 1993.

[9] T. Belpaeme, P. E. Baxter, R. Read, R. Wood, H. Cuayáhuitl, B. Kiefer, S. Racioppa, I. Kruijff-Korbayová, G. Athanasopoulos, V. Enescu, et al. Multimodal child-robot interaction: Building social bonds. *Journal of Human-Robot Interaction*, 1(2):33–53, 2012.

[10] C. Breazeal, M. Berlin, A. Brooks, J. Gray, and A. Thomaz. Using perspective taking to learn from ambiguous demonstrations. *Robotics and Autonomous Systems*, pages 385–393, 2006.

[11] C. Breazeal, J. Gray, and M. Berlin. An embodied cognition approach to mindreading skills for socially intelligent robots. *The International Journal of Robotics Research*, 28(5):656–680, 2009.

[12] M. Cirillo, L. Karlsson, and A. Saffiotti. Human-aware task planning: an application to mobile robots. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 1(2):15, 2010.

[13] H. H. Clark and S. E. Brennan. Grounding in communication. *Perspectives on socially shared cognition*, 13(1991):127–149, 1991.

[14] H. H. Clark and E. F. Schaefer. Contributing to discourse. *Cognitive science*, 13(2):259–294, 1989.

[15] H. H. Clark and D. Wilkes-Gibbs. Referring as a collaborative process. *Cognition*, 22(1):1–39, 1986.

[16] P. Dillenbourg, M. J. Baker, A. Blaye, and C. O'Malley. The evolution of research on collaborative learning. pages 189–211, 1995.

[17] P. Dillenbourg and D. Traum. Sharing solutions: Persistence and grounding in multimodal collaborative problem solving. *The Journal of the Learning Sciences*, 15(1):121–151, 2006.

[18] R. Fagin, J. Y. Halpern, Y. Moses, and M. Y. Vardi. *Reasoning About Knowledge*. MIT Press, Cambridge, MA, USA, 1995.

[19] J. Fink, S. Lemaignan, C. Braboszcz, and P. Dillenbourg. Dynamics of anthropomorphism in human-robot interaction. *Frontiers in Cognitive Science*, 2014. Submitted.

[20] J. H. Flavell. The development of knowledge about

visual perception. *Nebraska Symposium on Motivation*, 25:43–76, 1977.

[21] J. H. Flavell, F. L. Green, and E. R. Flavell. Developmental changes in young children's knowledge about the mind. *Cognitive Development*, 5(1):1–27, 1990.

[22] L. Flobbe, R. Verbrugge, P. Hendriks, and I. Krämer. Children's application of theory of mind in reasoning and language. *Journal of Logic, Language and Information*, 17(4):417–442, 2008.

[23] U. Frith and F. Happé. Autism: Beyond "theory of mind". *Cognition*, 50(1):115–132, 1994.

[24] F. G. Happé. Communicative competence and theory of mind in autism: A test of relevance theory. *Cognition*, 48(2):101–119, 1993.

[25] S. Harnad. The symbol grounding problem. *Physica D: Nonlinear Phenomena*, 42(1):335–346, 1990.

[26] V. Hendricks and J. Symons. Epistemic logic. In *Stanford Encyclopedia of Philosophy*. 2008.

[27] A. Herzig. Logics of knowledge and action: critical analysis and challenges. *Autonomous Agents and Multi-Agent Systems*, pages 1–35, 2014.

[28] S. Lemaignan. *Grounding the Interaction: Knowledge Management for Interactive Robots*, chapter Symbolic Knowledge Representation, pages 15–59. 2012.

[29] A. M. Leslie and L. Thaiss. Domain specificity in conceptual development: Neuropsychological evidence from autism. *Cognition*, 43(3):225–251, 1992.

[30] H. Levesque, F. Pirri, and R. Reiter. Foundations for the situation calculus. *Electronic Transactions on Artificial Intelligence*, 2:159–178, 1998.

[31] H. Levesque, R. Reiter, Y. Lesperance, F. Lin, and R. Scherl. GOLOG: A logic programming language for dynamic domains. *The Journal of Logic Programming*, 31(1-3):59–83, 1997.

[32] C. M. Mant and J. Perner. The child's understanding of commitment. *Developmental Psychology*, 24(3):343, 1988.

[33] N. Mavridis and D. Roy. Grounded situation models for robots: Where words and percepts meet. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2006.

[34] D. P. Oswald and T. H. Ollendick. Role taking and social competence in autism and mental retardation. *Journal of autism and developmental disorders*, 19(1):119–127, 1989.

[35] J. Perner. Higher-order beliefs and intentions in children's understanding of social interaction. *Developing theories of mind*, pages 271–294, 1988.

[36] J. Perner, U. Frith, A. M. Leslie, and S. R. Leekam. Exploration of the autistic child's theory of mind: Knowledge, belief, and communication. *Child development*, pages 689–700, 1989.

[37] M. J. Pickering and S. Garrod. Alignment as the basis for successful communication. *Research on Language and Computation*, 4(2-3):203–228, 2006.

[38] R. Ros, E. A. Sisbot, R. Alami, J. Steinwender, K. Hamann, and F. Warneken. Solving ambiguities with perspective taking. In *5th ACM/IEEE International Conference on Human-Robot Interaction*, 2010.

[39] J. Roschelle and S. D. Teasley. The construction of shared knowledge in collaborative problem solving. In *Computer supported collaborative learning*, pages 69–97. Springer, 1995.

[40] M. Sangin, N. Nova, G. Molinari, and P. Dillenbourg. Partner modeling is mutual. In *Proceedings of the 8th iternational conference on Computer Supported Collaborative Learning*, pages 625–632. International Society of the Learning Sciences, 2007.

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

[42] M. F. Schober. Spatial perspective-taking in conversation. *Cognition*, 47(1):1–24, 1993.

[43] D. L. Schwartz. The emergence of abstract representations in dyad problem solving. *The Journal of the Learning Sciences*, 4(3):321–354, 1995.

[44] B. Sodian and U. Frith. Deception and sabotage in autistic, retarded and normal children. *Journal of Child Psychology and Psychiatry*, 33(3):591–605, 1992.

[45] G. Stahl. Meaning making in cscl: Conditions and preconditions for cognitive processes by groups. In *Proceedings of the 8th iternational conference on Computer supported collaborative learning*, pages 652–661. International Society of the Learning Sciences, 2007.

[46] D. D. Suthers. Technology affordances for intersubjective meaning making: A research agenda for cscl. *International Journal of Computer-Supported Collaborative Learning*, 1(3):315–337, 2006.

[47] J. Trafton, N. Cassimatis, M. Bugajska, D. Brock, F. Mintz, and A. Schultz. Enabling effective human-robot interaction using perspective-taking in robots. *IEEE Transactions on Systems, Man and Cybernetics, Part A: Systems and Humans*, 35(4):460–470, 2005.

[48] H. van Ditmarsch, A. Herzig, and T. De Lima. From situation calculus to dynamic epistemic logic. *Journal of Logic and Computation*, 2010.

[49] H. van Ditmarsch and W. Labuschagne. My beliefs about your beliefs: a case study in theory of mind and epistemic logic. *Synthese*, 155(2):191–209, 2007.

[50] H. van Ditmarsch, W. van der Hoek, and B. P. Kooi. *Dynamic epistemic logic*, volume 337. Springer, 2007.

[51] H. P. van Ditmarsch. The description of game actions in cluedo. *Game theory and applications*, 8:1–28, 2002.

[52] R. Verbrugge. Logic and social cognition. *Journal of Philosophical Logic*, 38(6):649–680, 2009.

[53] R. Verbrugge and L. Mol. Learning to apply theory of mind. *Journal of Logic, Language and Information*, 17(4):489–511, 2008.

[54] 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 *Proceedings of the 21st IEEE International Symposium on Robot and Human Interactive Communication*, pages 948–954, 2012.

[55] H. Wimmer and J. Perner. Beliefs about beliefs: Representation and constraining function of wrong beliefs in young children's understanding of deception. *Cognition*, 13(1):103–128, 1983.