Cognitive Aspects in Epistemic Logic L-DINF*

Stefania Costantini[1,3], Andrea Formisano[2,3,*] and Valentina Pitoni[1]

[1]Università di L'Aquila (DISIM), Via Vetoio, 1, L'Aquila, 67100, Italy
[2]Università di Udine (DMIF), Via delle Scienze, 206, Udine, 33100, Italy
[3]INdAM — GNCS, Piazzale Aldo Moro, 5, Roma, 00185, Italy

Abstract

In this paper, we report about a line of work aimed to formally model via a logical framework —the Logic of "Inferable" *L-DINF*— (aspects of) the group dynamics of cooperative agents. We outline, in particular, the cognitive aspects built within our logic, that consist in features allowing a designer to model real-world situations encompassing joint intentions and plans with roles, preferences and costs concerning action execution, and involving aspects of a Theory of Mind, i.e., the ability to reason about beliefs of others.

Keywords

Epistemic Logic, Cognitive Aspect, Multi-Agent Systems, Cooperation and Roles assignment

1. Introduction

Agents and Multi-Agent Systems (MAS) have been widely adopted in Artificial Intelligence (AI) to model societies whose members are to some extent cooperative towards each other. Agents belonging to a MAS can be able to achieve better results via cooperation, because it is often the case that a group can fulfil objectives that are out of reach for a single agent. To this aim, it is useful for agents to be able to reason about what their group of agents can do, and what they can do within a group (including leaving/joining groups).

In this paper, we present a logical framework (the Logic of "Inferable" *L-DINF* [1]) that we have proposed and extended over the years, defined originally as an extension of a pre-existing epistemic logic by Lorini & Balbiani. In *L-DINF*, groups of cooperative agents can jointly perform actions and reach goals. We outline, in particular, the cognitive aspects built within such a logic. In *L-DINF*, our overall objective has been to devise an agent-oriented logical framework that allows a designer to formalize and formally verify MAS, modelling the capability to construct and execute joint plans within a group of agents. We have devoted all along a special attention to explainability, in the perspective of Trustworty AI and to cognitive aspect.

In the logic, we have considered actions' *cost*, and agents' *budget*, where an agent able to perform an action but not owning the needed budget can be supported by its group to cover the cost. The group takes into consideration the preferences that each agent can have for what concerns performing each action. We have also introduced agents' *roles* within a group, in terms

IJCAI 22 Workshop: Cognitive Aspects of Knowledge Representation, July 23-29, 2022, Messe Wien, Vienna, Austria
*Corresponding author.
stefania.costantini@univaq.it (S. Costantini); andrea.formisano@uniud.it (A. Formisano); valentina.pitoni@univaq.it (V. Pitoni)
0000-0002-5686-6124 (S. Costantini); 0000-0002-6755-9314 (A. Formisano); 0000-0002-4245-4073 (V. Pitoni)



CEUR Workshop Proceedings (CEUR-WS.org)

of the actions that each agent is allowed to perform in the context of the group. All these features are defined in a modular way, so as to be composed to model different group behaviours. For instance, normally an action can be performed only by an agent which is able and allowed to perform it; but, in exceptional circumstances, for lack of alternatives, also by an agent that is able, although not allowed, to perform it. Naturally, we devised a full syntax and semantics and the proof of strong completeness of our logic w.r.t. its canonical model.

Many kinds of logical frameworks can be found in the literature which try to emulate cognitive aspects of human beings, also from the cooperative point of view. However, what distinguishes our approach from related ones is that many key aspects are not specified in the logical theory defining an agent: rather, we introduce special functions in the definitions of model and of canonical model. For the practical realization of such functions, we envisage separate modules from which agent's logical theory "inputs" the results. Such modules might be specified even in some other logic or also, pragmatically, via pieces of code. Other conditions such as, e.g., feasibility of actions, are also defined modularly, as they concern aspects that should be verified contextually, according to agents' environmental conditions.

In recent work [2], we considered that there are classes of applications where agents can profit from the ability to represent group dynamics, and to understand the behaviour of others; i.e., agents should be able to assess or hypothesize what other agents (including human users) believe and intend to do. This accounts to be able to represent aspects of *"Theory of Mind"*, which is the set of social-cognitive skills involving: the ability to attribute mental states, including desires, beliefs, and knowledge, to oneself and to other agents; and, importantly, the ability to reason about the practical consequences of such mental states. Such ability is crucial for prediction and interpretation of other agents' behavioural responses (cf. Oxford Handbook of Philosophy and Cognitive Science [3], Chapter by Alvin I. Goldman). Thus, we have devised an extension of *L-DINF* able to represent aspects of Theory of Mind. Our motivation lies in our related research work in agent-oriented programming languages. In particular, our research group has defined the language DALI [4, 5, 6], which has been fully implemented [7] and is endowed with a fully logical semantics [8]. DALI has been applied in many applications, among which cognitive robotics [9] and eHealth [10, 11]: in such application fields, a socially and psychologically acceptable interaction with the user is required, whence our aim to develop a suitable logical formalization of (at least a basic version of) ToM to be in perspective incorporated into DALI semantics and implementation.

In [12] we have thoroughly discussed the relationship of logic *L-DINF* with related work, so we refer the reader to that paper for this point.

The paper is organized as follows. In Section 2 we introduce syntax and semantics of *L-DINF*. Sections 3 and 4 discuss significant examples of application of the new logic, concerning cognitive aspects. The first considers how to tune group's behaviour according to circumstances; the second concerns how to represent aspects of the Theory of Mind Finally, in Section 5 we conclude.

2. Logical Framework

L-DINF is a logic which consists of a static component and a dynamic one. The static component, called *L-INF*, is a logic of explicit beliefs and background knowledge. The dynamic component,

called *L-DINF*, extends the static one with dynamic operators capturing the consequences of the agents' inferential actions on their explicit beliefs as well as a dynamic operator capturing what an agent can conclude by performing some inferential action in its repertoire.

2.1. Syntax

In this section we provide and illustrate the syntax of the proposed logic. Let $Atm = \{p, q, \ldots\}$ be a countable set of atomic propositions. By $Prop$ we denote the set of all propositional formulas, i.e. the set of all Boolean formulas built out of the set of atomic propositions Atm. The set Atm_A represents the set of physical actions that agents can perform, including "active sensing" actions (e.g., "let's check whether it rains", "let's measure the temperature"). Let Agt be a set of n agents identified, for simplicity, by integer numbers: $Agt = \{1, 2, \ldots, n\}$. The language of *L-DINF*, denoted by $\mathcal{L}_{\textit{L-DINF}}$, is defined by the following grammar:

$$\begin{array}{lll} \varphi, \psi & ::= & p \mid \neg\varphi \mid \varphi \wedge \psi \mid \mathbf{B}_i\,\varphi \mid \mathbf{K}_i\,\varphi \mid do_i(\phi_A) \mid do_i^P(\phi_A) \mid can\_do_i(\phi_A) \mid \\ & & do_G(\phi_A) \mid do_G^P(\phi_A) \mid can\_do_G(\phi_A) \mid intend_i(\phi_A) \mid intend_G(\phi_A) \mid \\ & & exec_i(\alpha) \mid exec_G(\alpha) \mid [G{:}\alpha]\,\varphi \mid pref\_do_i(\phi_A, d) \mid pref\_do_G(i, \phi_A) \mid \\ \alpha & ::= & \vdash(\varphi,\psi) \mid \cap(\varphi,\psi) \mid \downarrow(\varphi, \psi) \mid \dashv(\varphi, \psi) \end{array}$$

where p ranges over Atm, $i \in Agt$, $G \subseteq Agt$, $\phi_A \in Atm_A$, and $d \in \mathbb{N}$. Other Boolean operators are defined from $\neg$ and $\wedge$ in the standard manner. Moreover, for simplicity, whenever $G = \{i\}$ we will write i as subscript in place of $\{i\}$. So, for instance, we often write $pref\_do_i(i, \phi_A)$ instead of $pref\_do_{\{i\}}(i, \phi_A)$ and similarly for other constructs.

The language of *inferential actions* of type α is denoted by $\mathcal{L}_{\mathsf{ACT}}$. The static part *L-INF* of *L-DINF*, includes only those formulas not having sub-formulas of type α, namely, no inferential operation is admitted.

Notice that, the framework we are introducing is propositional, but, to simplify the description, we will often write elements of Atm and of Atm_A as structured expressions, such as $in(doll, box)$, $in(doll, basket)$, or $put_A(doll, box)$.

Notice, moreover, that we do not consider in this paper the possibility that an agent has belief or knowledge involving nesting of modalities. Hence, in formulas of the forms $\mathbf{B}_i\,\varphi$ and $\mathbf{K}_i\,\varphi$, the subformula φ does not involve any modal operator $\mathbf{B}_j$ and $\mathbf{K}_j$ (for any i, j).

We consider the set of agents as partitioned in groups: each agent i always belongs to a single group $G \subseteq Agt$. Any agent i, at any time, can perform a (physical) action $join_A(i, j)$, for $j \in Agt$, in order to change its group and join j's group.

Before introducing the formal semantics let us briefly describe the intended informal meaning of basic formulas of *L-INF*. Expressions of the form $intend_i(\phi_A)$, where $\phi_A \in Atm_A$, indicate the intention of agent i to perform the action ϕ_A in the sense of the BDI agent model [13]. This intention can be part of an agent's knowledge base from the beginning, or it can be derived later. In this paper we do not cope with the formalization of BDI, for which the reader may refer, e.g., to [14]. So, we will treat intentions rather informally, assuming also that $intend_G(\phi_A)$ holds whenever all agents in group G intend to perform action ϕ_A.

The formula $do_i(\phi_A)$, where again we require that $\phi_A \in Atm_A$, indicates *actual execution* of action ϕ_A by agent i, automatically recorded by the new belief $do_i^P(\phi_A)$ (postfix "P" standing for "past" action). By precise choice, do and do^P (and similarly do_G and do_G^P) are not axiomatized.

In fact, they are realized by what has been called in [15] a *semantic attachment*, i.e., a procedure which connects an agent with its external environment in a way that is unknown at the logical level. The axiomatization concerns only the relationship between doing and being enabled to do.

The expressions $can\_do_i(\phi_A)$ and $pref\_do_i(\phi_A, d)$ (where it is required that $\phi_A \in Atm_A$) are closely related to $do_i(\phi_A)$. In fact, $can\_do_i(\phi_A)$ is to be seen as an enabling condition, indicating that agent i is enabled to execute action ϕ_A, while instead $pref\_do_i(\phi_A, d)$ indicates the level d of preference/willingness of agent i to perform that action. The expression $pref\_do_G(i, \phi_A)$ indicates that agent i exhibits the *maximum level* of preference on performing action ϕ_A among all members of its group G. Notice that, if a group of agents intends to perform an action ϕ_A, this will entail that the entire group intends to do ϕ_A, that will be enabled to be actually executed only if at least one agent $i \in G$ can do it, i.e., it can derive $can\_do_i(\phi_A)$.

Unlike explicit beliefs, i.e., facts and rules acquired via perceptions during an agent's operation and kept in the *working memory*, an agent's background knowledge is assumed to satisfy *omniscience* principles, such as closure under conjunction and known implication, and closure under logical consequence, and introspection. In fact, $\mathbf{K}_i$ is actually the well-known S5 modal operator often used to model/represent knowledge. The fact that background knowledge is closed under logical consequence is justified because we conceive it as a kind of stable reliable *knowledge base*, or *long-term memory*. We assume the background knowledge to include: facts (formulas) known by the agent from the beginning, and facts the agent has later decided to store in its long-term memory (by means of some decision mechanism not treated here) after having processed them in its working memory. We therefore assume background knowledge to be irrevocable, in the sense of being stable over time.

A formula of the form $[G : \alpha]\, \varphi$, with $G \subseteq Agt$, and where α must be an inferential action, states that "φ holds after action α has been performed by at least one of the agents in G, and all agents in G have common knowledge about this fact". We distinguish four types of inferential actions α which allow us to capture some of the dynamic properties of explicit beliefs and background knowledge: $\downarrow(\varphi, \psi)$, $\cap(\varphi,\psi)$, $\dashv(\varphi, \psi)$, and $\vdash(\varphi,\psi)$. These actions characterize the basic operations of forming explicit beliefs via inference:

- $\downarrow(\varphi, \psi)$: this inferential action infers ψ from φ, in case φ is believed and, according to agent's background knowledge, ψ is a logical consequence of φ. I.e., by performing this inferential action, an agent tries to retrieve from its background knowledge in long-term memory the information that φ implies ψ and, if it succeeds, it starts believing ψ.
- $\cap(\varphi,\psi)$: closes the explicit belief φ and the explicit belief ψ under conjunction. I.e., $\varphi \wedge \psi$ is deduced from the explicit beliefs φ and ψ.
- $\dashv(\varphi, \psi)$: this inferential action performs a simple form of "belief revision". It removes ψ from the working memory in case φ is believed and, according to agent's background knowledge, $\neg\psi$ is logical consequence of φ. Both ψ and φ are required to be ground atoms.
- $\vdash(\varphi,\psi)$: adds ψ to the working memory in case φ is believed and, according to agent's working memory, ψ is logical consequence of φ. Differently from $\downarrow(\varphi, \psi)$, this action operates on the working memory without retrieving anything from the background knowledge.

Formulas of the forms $exec_i(\alpha)$ and $exec_G(\alpha)$ express executability of inferential actions either by agent i, or by a group G of agents (which is a consequence of any of the group members

being able to execute the action). It has to be read as: "α is an inferential action that agent i (resp. an agent in G) can perform".

Remark 1. *In the mental actions $\vdash(\varphi,\psi)$ and $\downarrow(\varphi,\psi)$, the formula ψ which is inferred and asserted as a new belief can be $can\_do_i(\phi_A)$ or $do_i(\phi_A)$, which denote the possibility of execution or actual execution of physical action ϕ_A. We assume that when inferring $do_i(\phi_A)$ (from $can\_do_i(\phi_A)$ and possibly other conditions) then the action is actually executed, and the corresponding belief $do_i^P(\phi_A)$ is asserted, possibly augmented with a time-stamp. Actions are supposed to succeed by default; in case of failure, a corresponding failure event will be perceived by the agent. The do_i^P beliefs constitute a* history *of the agent's operation, so they might be useful for the agent to reason about its own past behaviour, and/or, importantly, they may be useful to provide* explanations *to human users.*

Remark 2. *Explainability in our approach can be directly obtained from proofs. Let us assume for simplicity that inferential actions can be represented in infix form as $\varphi_j\ OP\ \varphi_{j+1}$ for some j. Also, for agent i, $exec_i(\alpha)$ means that the mental action α is executable by the agent and it is indeed executed. If, for instance, the user wants an explanation of why the action ϕ_A has been performed, the system can exhibit the proof that has lead to ϕ_A, put in the explicit form:*
$\big(exec_i(\varphi_1 OP_1\ \varphi_2) \wedge \ldots \wedge exec_i(\varphi_{n-1} OP_n\ \varphi_n) \wedge exec_i(\varphi_n OP_n\ can\_do_i(\phi_A)) \wedge$
$intend_i(\phi_A)\big) \vdash do_i(\phi_A)$
where each OP_k is one of the (mental) actions discussed above. The proof can possibly be translated into natural language, and declined either top-down or bottom-up.

As said in the Introduction, we model agents which, to execute an action, may have to pay a cost, so they must have a consistent budget available. Moreover, agents are entitled to perform only those physical actions that they conclude they can do. Agents belonging to a group are assumed to be cooperative. An action can be executed by a group if at least one agent in the group is able to execute it, and the group has the necessary budget available, sharing the cost according to some policy. The cooperative nature of our agents manifests itself also in selecting, among the agents that are able to do some physical action, the one(s) which best prefer to perform that action. We do not have introduced costs and budget, feasibility of actions and willingness to perform them, *in the language* for two reasons: to keep the complexity of the logic reasonable, and to make such features customizable in a modular way. For instance, cost-sharing policies different from the one that we will show below might easily be introduced, even different ones for different resources.

2.2. Semantics

Definition 2.1 introduces the notion of *L-INF model*, which is then used to introduce semantics of the static fragment of the logic. Notice that many relevant aspects of an agent's behaviour are specified in the definition of *L-INF model*, including which mental and physical actions an agent can perform, which is the cost of an action and which is the budget that the agent has available, which is the preference degree of the agent to perform each action. This choice has the advantage of keeping the complexity of the logic under control, and of making these aspects modularly modifiable. In this paper, we introduce new function H that, for each agent i belonging to a

group, enables the agent to perform a certain set of actions, so, in this way, it specifies the *role* of i within the group. As before let Agt be the set of agents.

Definition 2.1. *A model is a tuple $M = (W, N, \mathcal{R}, E, B, C, A, H, P, V)$ where:*

- W *is a set of worlds (or situations);*
- $\mathcal{R} = \{R_i\}_{i \in Agt}$ *is a collection of equivalence relations on W: $R_i \subseteq W \times W$ for each $i \in Agt$;*
- $N : Agt \times W \longrightarrow 2^{2^W}$ *is a neighborhood function such that, for each $i \in Agt$, each $w, v \in W$, and each $X \subseteq W$ these conditions hold:*

  (C1) *if $X \in N(i, w)$ then $X \subseteq \{v \in W \mid wR_iv\}$,*

  (C2) *if wR_iv then $N(i, w) = N(i, v)$;*
- $E : Agt \times W \longrightarrow 2^{\mathcal{L}_{\mathsf{ACT}}}$ *is an executability function of mental actions such that, for each $i \in Agt$ and $w, v \in W$, it holds that:*

  (D1) *if wR_iv then $E(i, w) = E(i, v)$;*
- $B : Agt \times W \longrightarrow \mathbb{N}$ *is a budget function such that, for each $i \in Agt$ and $w, v \in W$, the following holds*

  (E1) *if wR_iv then $B(i, w) = B(i, v)$;*
- $C : Agt \times \mathcal{L}_{\mathsf{ACT}} \times W \longrightarrow \mathbb{N}$ *is a cost function such that, for each $i \in Agt$, $\alpha \in \mathcal{L}_{\mathsf{ACT}}$, and $w, v \in W$, it holds that:*

  (F1) *if wR_iv then $C(i, \alpha, w) = C(i, \alpha, v)$;*
- $A : Agt \times W \longrightarrow 2^{Atm_A}$ *is an executability function for physical actions such that, for each $i \in Agt$ and $w, v \in W$, it holds that:*

  (G1) *if wR_iv then $A(i, w) = A(i, v)$;*
- $H : Agt \times W \longrightarrow 2^{Atm_A}$ *is an enabling function for physical actions such that, for each $i \in Agt$ and $w, v \in W$, it holds that:*

  (G2) *if wR_iv then $H(i, w) = H(i, v)$;*
- $P : Agt \times W \times Atm_A \longrightarrow \mathbb{N}$ *is a preference function for physical actions ϕ_A such that, for each $i \in Agt$ and $w, v \in W$, it holds that:*

  (H1) *if wR_iv then $P(i, w, \phi_A) = P(i, v, \phi_A)$;*
- $V : W \longrightarrow 2^{Atm}$ *is a valuation function.*

To simplify the notation, let $R_i(w)$ denote the set $\{v \in W \mid wR_iv\}$, for $w \in W$. The set $R_i(w)$ identifies the situations that agent i considers possible at world w. It is the *epistemic state* of agent i at w. In cognitive terms, $R_i(w)$ can be conceived as the set of all situations that agent i *can retrieve* from its long-term memory and reason about.

While $R_i(w)$ concerns background knowledge, $N(i, w)$ is the set of all facts that agent i explicitly believes at world w, a fact being identified with a set of worlds. Hence, if $X \in N(i, w)$ then, the agent i has the fact X under the focus of its attention and believes it. We say that $N(i, w)$ is the explicit *belief set* of agent i at world w.

The executability of inferential actions is determined by the function E. For an agent i, $E(i, w)$ is the set of inferential actions that agent i can execute at world w. The value $B(i, w)$ is the budget the agent has available to perform inferential actions. The value $C(i, \alpha, w)$ is the cost to be paid by agent i to execute the inferential action α in the world w. $N(i, w)$ and $B(i, w)$ are updated whenever a mental action is performed, and this changes are described in [16]. The executability of physical actions is determined by the function A. For an agent i, $A(i, w)$ is the set of physical actions that agent i can execute at world w. $H(i, w)$ instead is the set of physical actions that agent i is enabled by its group to perform. Which means, H defines the *role* of an agent in its group, via the actions that it is allowed to execute.

Agent's preference on executability of physical actions is determined by the function P. For an agent i, and a physical action ϕ_A, $P(i, w, \phi_A)$ is an integer value d indicating the degree of willingness of agent i to execute ϕ_A at world w.

Constraint (**C1**) imposes that agent i can have explicit in its mind only facts which are compatible with its current epistemic state. Moreover, according to constraint (**C2**), if a world v is compatible with the epistemic state of agent i at world w, then agent i should have the same explicit beliefs at w and v. In other words, if two situations are equivalent as concerns background knowledge, then they cannot be distinguished through the explicit belief set. This aspect of the semantics can be extended in future work to allow agents make plausible assumptions. Analogous properties are imposed by constraints (**D1**), (**E1**), and (**F1**). Namely, (**D1**) imposes that agent i always knows which inferential actions it can perform and those it cannot. (**E1**) states that agent i always knows the available budget in a world (potentially needed to perform actions). (**F1**) determines that agent i always knows how much it costs to perform an inferential action. (**G1**) and (**H1**) determine that an agent i always knows which physical actions it can perform and those it cannot, and with which degree of willingness, where (**G2**) specifies that an agent also knows whether its group gives it the permission to execute a certain action or not, i.e., if that action pertains to its *role* in the group.

Truth values of *L-DINF* formulas are inductively defined as follows.

Given a model $M = (W, N, \mathcal{R}, E, B, C, A, H, P, V)$, $i \in Agt$, $G \subseteq Agt$, $w \in W$, and a formula $\varphi \in \mathcal{L}_{\textit{L-INF}}$, we introduce the following shorthand notation:

$$\|\varphi\|^M_{i,w} = \{v \in W : wR_iv \text{ and } M, v \models \varphi\}$$

whenever $M, v \models \varphi$ is well-defined (see below). Then, we set:

(t1) $M, w \models p$ iff $p \in V(w)$
(t2) $M, w \models \mathit{exec}_i(\alpha)$ iff $\alpha \in E(i, w)$
(t3) $M, w \models \mathit{exec}_G(\alpha)$ iff $\exists i \in G$ with $\alpha \in E(i, w)$
(t4) $M, w \models \mathit{can\_do}_i(\phi_A)$ iff $\phi_A \in A(i, w) \cap H(i, w)$
(t5) $M, w \models \mathit{can\_do}_G(\phi_A)$ iff $\exists i \in G$ with $\phi_A \in A(i, w) \cap H(i, w)$
(t6) $M, w \models \mathit{pref\_do}_i(\phi_A, d)$ iff $\phi_A \in A(i, w)$ and $P(i, w, \phi_A) = d$
(t7) $M, w \models \mathit{pref\_do}_G(i, \phi_A)$ iff $M, w \models \mathit{pref\_do}_i(\phi_A, d)$ for $d = \max\{P(j, w, \phi_A) \mid j \in G \wedge \phi_A \in A(j, w) \cap H(j, w)\}$
(t8) $M, w \models \neg\varphi$ iff $M, w \not\models \varphi$
(t9) $M, w \models \varphi \wedge \psi$ iff $M, w \models \varphi$ and $M, w \models \psi$

(t10) $M, w \models \mathbf{B}_i\, \varphi$ iff $||\varphi||^M_{i,w} \in N(i, w)$
(t11) $M, w \models \mathbf{K}_i\, \varphi$ iff $M, v \models \varphi$ for all $v \in R_i(w)$

As seen above, a physical action can be performed by a group of agents if at least one agent of the group can do it, and the level of preference for performing this action is set to the maximum among those of the agents enabled to do this action.

For any inferential action α performed by any agent i, we set:

$$M, w \models [G : \alpha]\varphi \text{ iff } M^{[G:\alpha]}, w \models \varphi$$

where $M^{[G:\alpha]} = \langle W, N^{[G:\alpha]}, \mathcal{R}, E, B^{[G:\alpha]}, C, A, H, P, V\rangle$, is the model representing the fact that the execution of an inferential action α affects the sets of beliefs of agent i and modifies the available budget. Such operation can add new beliefs by direct perception, by means of one inference step, or as a conjunction of previous beliefs. Hence, when introducing new beliefs (i.e., performing mental actions), the neighborhood must be extended accordingly. Notice that the execution of inferential actions only affects agents' beliefs, i.e., the contents of their working memory. Mental actions have no effect on agents' knowledge, which remains persistent.

For details about belief/neighborhood update, we refer the reader to [16] where an axiomatization of the logic framework described so far is also provided, together with results concerning soundness and completeness of the logic.

A key aspect in the definition of the logic is the following, which states under which conditions, and by which agent(s), an action may be performed.

$$enabled_w(G, \alpha) : \ \exists j \in G\, (\alpha \in E(j, w) \wedge \frac{C(j, \alpha, w)}{|G|} \leq \min_{h \in G} B(h, w)).$$

This condition states when an inferential action is enabled. In the above particular formulation (that is not fixed, but can be customized to the specific application domain) if at least an agent can perform it; and if the "payment" due by each agent, obtained by dividing the action's cost equally among all agents of the group, is within each agent's available budget. In case more than one agent in G can execute an action, we implicitly assume the agent j performing the action to be the one corresponding to the lowest possible cost. Namely, j is such that $C(j, \alpha, w) = \min_{h \in G} C(h, \alpha, w)$. This definition reflects a parsimony criterion reasonably adoptable by cooperative agents sharing a crucial resource such as, e.g., energy or money. Other choices might be viable, so variations of this logic can be easily defined simply by devising some other enabling condition and, possibly, introducing differences in neighborhood update. Notice that the definition of the enabling function basically specifies the "concrete responsibility" that agents take while concurring with their own resources to actions' execution. Also, in case of specification of various resources, different corresponding enabling functions might be defined.

Our contribution to modularity is that functions A, P and H, i.e., executability of physical actions, preference level of an agent about performing each action, and permission concerning which actions to actually perform, are not meant to be built-in. Rather, they can be defined via separate sub-theories, possibly defined using different logics, or, in a practical approach, even via pieces of code. This approach can be extended to function C, i.e., the cost of mental actions instead of being fixed may in principle vary, and be computed upon need.

3. Problem Specification and Inference: An Example, and Discussion

In this section, we propose an example to explain the usefulness of this kind of logic underlying cognitive aspects; in fact, to the best of our knowledge, no one in literature is using logic in this way. For the sake of simplicity of illustration and of brevity, the example is in "skeletal" form. Consider a group of four agents, who are the crew of an ambulance, including a driver, two nurses, and a medical doctor. The driver is the only one enabled to drive the ambulance. The nurses are enabled to perform a number of tasks, such as, e.g., administer a pain reliever, or clean, disinfect and bandage a wound, measure vital signs. It is however the task of a doctor to make a diagnosis, to prescribe medications, to order, perform, and interpret diagnostic tests, and to perform complex medical procedures. Let us identify the four agents by integer numbers and, accordingly, let $G = \{1, 2, 3, 4\}$ be their group.

Imagine that the hospital received notice of a car accident with an injured person. Then, it will inform the group of the fact that a patient needs help (how exactly is not treated here, because this depends on how the multi-agent system is implemented, but a message exchange will presumably suffice). The group will reason, and devise the intention/goal $\mathbf{K}_i(intend_G(rescue\_patient))$.

Among the physical actions that agents in G can perform, there are the following:

$diagnose\_patient$	$administer\_urgent\_treatment$
$measure\_vital\_signs$	$pneumothorax\_aspiration$
$local\_anesthesia$	$bandage\_wounds$
$drive\_to\_patient$	$drive\_to\_hospital.$

The group will now be required to perform a planning activity. Assume that, as a result of the planning phase, the knowledge base of each agent i contains the following rule, that specifies how to reach the intended goal in terms of actions to perform and sub-goals to achieve:

$$\mathbf{K}_i\big(intend_G(rescue\_patient) \rightarrow intend_G(drive\_to\_patient) \wedge intend_G(diagnose\_patient) \wedge intend_G(stabilize\_patient) \wedge intend_G(drive\_to\_hospital)\big)$$

Thanks to the mentioned axiomatization for *L-DINF* (specifically, by axiom 18 in [16], stating that $intend_G(\phi_A) \leftrightarrow \forall i \in G\, intend_i(\phi_A)$) each agent has the specialized rule (for $i \leq 4$):

$$\mathbf{K}_i\big(intend_G(rescue\_patient) \rightarrow intend_i(drive\_to\_patient) \wedge intend_i(diagnose\_patient) \wedge intend_i(stabilize\_patient) \wedge intend_i(drive\_to\_hospital)\big)$$

Therefore, the following is entailed for each agent:

$$\begin{array}{l}\mathbf{K}_i\big(intend_i(rescue\_patient) \rightarrow intend_i(drive\_to\_patient)\big)\\ \mathbf{K}_i\big(intend_i(rescue\_patient) \rightarrow intend_i(diagnose\_patient)\big)\\ \mathbf{K}_i\big(intend_i(rescue\_patient) \rightarrow intend_i(stabilize\_patient)\big)\\ \mathbf{K}_i\big(intend_i(rescue\_patient) \rightarrow intend_i(drive\_to\_hospital)\big)\end{array}$$

While driving to the patient and then going back to the hospital are actions, $intend_G(stabilize\_patient)$ is a goal.

Assume now that the knowledge base of each agent i contains also the following general rules, stating that the group is available to perform each of the necessary actions. Which agent will in particular perform each action ϕ_A? According to items (t4) and (t7) in the definition of truth values, for *L-DINF* formulas, this agent will be chosen as the one which best prefers to perform this action, among those that can do it. Formally, in the present situation, $pref\_do_G(i, \phi_A)$ identifies an agent i in the group with a maximum degree of preference on performing ϕ_A (any deterministic rule can be applied to select i in case more agents express the highest degree), and $can\_do_G(\phi_A)$ is true if there is some agent i in the group which is able and allowed to perform ϕ_A, i.e., $\phi_A \in A(i, w) \wedge \phi_A \in H(i, w)$.

$$\begin{aligned}
&\mathbf{K}_i\big(intend_G(drive_to_patient) \wedge can_do_G(drive_to_patient) \wedge \\
&\qquad pref_do_G(i, drive_to_patient) \rightarrow do_G(drive_to_patient)\big) \\
&\mathbf{K}_i\big(intend_G(diagnose_patient) \wedge can_do_G(diagnose_patient) \wedge \\
&\qquad pref_do_G(i, diagnose_patient) \rightarrow do_G(diagnose_patient)\big) \\
&\mathbf{K}_i\big(intend_G(drive_to_hospital) \wedge can_do_G(drive_to_hospital) \wedge \\
&\qquad pref_do_G(i, drive_to_hospital) \rightarrow do_G(drive_to_hospital)\big)
\end{aligned}$$

As before, such rules can be specialized to each single agent:

$$\begin{aligned}
&\mathbf{K}_i\big(intend_i(drive_to_patient) \wedge can_do_i(drive_to_patient) \wedge \\
&\qquad pref_do_i(i, drive_to_patient) \rightarrow do_G(drive_to_patient)\big) \\
&\mathbf{K}_i\big(intend_i(diagnose_patient) \wedge can_do_i(diagnose_patient) \wedge \\
&\qquad pref_do_i(i, diagnose_patient) \rightarrow do_i(diagnose_patient)\big) \\
&\mathbf{K}_i\big(intend_i(drive_to_hospital) \wedge can_do_i(drive_to_hospital) \wedge \\
&\qquad pref_do_i(i, drive_to_hospital) \rightarrow do_i(drive_to_hospital)\big)
\end{aligned}$$

So, for each action ϕ_A required by the plan, there will be some agent (let us assume for simplicity only one), for which $do_i(\phi_A)$ will be concluded. In our case, the agent driver j will conclude $do_j(drive\_to\_patient)$ and $do_j(drive\_to\_hospital)$; the agent doctor ℓ will conclude $do_\ell(stabilize\_patient)$. As previously stated, whenever an agent derives $do_i(\phi_A)$ for any physical action ϕ_A, the action is supposed to have been performed via some kind of *semantic attachment* which links the agent to the external environment.

Since $intend_G(stabilize\_patient)$ is not an action but a sub-goal, the group will have to devise a plan to achieve it. This will imply sensing actions and forms of reasoning not shown here. Assume that the diagnosis has been pneumothorax, and that the patient has also some wounds which are bleeding. Upon completion of the planning phase, the knowledge base of each agent i contains the following rule, that specifies how to reach the intended goal in terms of actions to be performed:

$$\begin{aligned}
\mathbf{K}_i\big(&intend_G(stabilize_patient) \rightarrow intend_G(measure_vital_signs) \wedge intend_G(local_anesthesia) \wedge \\
&intend_G(bandage_wounds) \wedge intend_G(pneumothorax_aspiration)\big)
\end{aligned}$$

As before, these rules will be instantiated and elaborated by the single agents, and there will be some agent who will finally perform each action. Specifically, the doctor will be the one to perform pneumothorax aspiration, and the nurses (according to their competences and their preferences) will measure vital signs, administer local anesthesia and bandage the wounds. The

new function H, in a sensitive domain such as healthcare, guarantees that each procedure is administered by one who is capable to (function A) but also enabled (function H), and so can take responsibility for the action.

An interesting point concerns derogation, i.e., for instance, life or death situations where, unfortunately, no-one who is enabled to perform some urgently needed action is available; in such situations perhaps, anyone who is capable to perform this action might perform it. For instance, a nurse, in absence of a doctor, might attempt urgent pneumothorax aspiration.

From such perspective, semantics could be modified as follows:

(t4') $M, w \models \mathit{able\_do}_i(\phi_A)$ iff $\phi_A \in A(i, w)$

(t4") $M, w \models \mathit{enabled\_do}_i(\phi_A)$ iff $\phi_A \in A(i, w) \cap H(i, w)$

(t4-new) $M, w \models \mathit{can\_do}_i(\phi_A)$ iff $(\phi_A \in A(i, w) \cap H(i, w)) \vee (\phi_A \in A(i, w) \wedge \nexists j \in G : \phi_A \in A(j, w) \cap H(j, w))$

(t5-new) $M, w \models \mathit{can\_do}_G(\phi_A)$ iff $\exists\, i \in G$ s.t. $M, w \models \mathit{can\_do}_i(\phi_A)$

Thanks to this example, the ductility of this approach and the importance that is given to the cognitive aspect of the agent is highlighted.

4. False-Beliefs: the "Sally-Anne" Task

In approach proposed in [2], each agent has a version of the belief set of all the other agents. In such a proposal, belief sets are shared when agents belong to the same group, say G. But, as soon as an agent, say j, leaves the group G, its belief set might evolve and become different from the version owned by the other agents $i \in G$. Sharing of belief sets, allows any agent i in a group G to perform inferences based on the beliefs of another agent $h \notin G$. Clearly, this is done w.r.t. the version of h's beliefs owned by i (namely, the one shared the last time the two joined the same group). This means, agent i might do some mental action *impersonating* h, by exploiting its version of h's belief set and possibly updating such set. This extension of the logic framework is obtained by introducing modalities of the form $\mathbf{B}_{i,j}$ (in place of the $\mathbf{B}_i$ described in Section 2.1). The rational is that the operator $\mathbf{B}_{i,j}$ is used to model the beliefs that agent i has about j's beliefs.

This extension opens to the possibility that agents in G infer false beliefs about h. This may happen because agents $i \in G$ are not aware of mental actions performed meanwhile by agents $h \notin G$. In this case, i's version of the belief set of h does not incorporate the last changes made by h in its working memory. Consequently, all $i \in G$ have an obsolete representation of h's beliefs. Note that all beliefs are shared among all members of a group G.

To demonstrate that our logic is able to model relevant basic aspects of the Theory of Mind, we formalized the "Sally-Anne" task, an example of situation where false beliefs arise. The Sally–Anne test was introduced as a psychological test in developmental psychology in order to measure a person's social cognitive ability to attribute false beliefs to others. The test was administered to children for the fist time in [17] and then in [18], to test their ability to develop a "Theory of Mind" concerning the expectation of how someone will act based on the awareness of that person's false beliefs. The test has been then recently adopted to evaluate the cognitive capabilities of intelligent (possibly robotic) agents, see, e.g., [19].

The specification of the Sally-Anne task is the following: a child (or an agent), say Rose, is told a story about two girls, Sally and Anne, who are in a room with a basket and a box. In the story, Sally puts a doll into the basket, then leaves the room, and, in her absence, Anne moves the doll to the box. The child is then asked: "where does Sally believe the doll to be?". To pass the test, the child should answer that Sally believes the doll to be in the basket. If asked for an explanation, she should answer that, since Sally did not see Anne moving the doll, she has the false belief that the doll is still in the basket.

Our formalization of the task is found below. Notice that an uppercase initial letter denotes a variable symbol. The use of variables has however to be intended as a shorthand notation, as indeed, the language is propositional.

We assume to have three agents, namely 1 (Sally), 2 (Anne), and 3 (observer). Initially, all agents belong to the same group and share the beliefs $\mathbf{B}_{i,j}(can\_do_1(put_A(doll, box)))$ and $\mathbf{B}_{i,j}(intend_1(put_A(doll, box)))$, for $i, j \in \{1, 2, 3\}$, and that each agent has in her background knowledge the rule that states that an agent which is able and willing to perform some physical action will indeed do it: $\mathbf{K}_i(can\_do_i(\Phi_A) \wedge intend_i(\Phi_A) \rightarrow do_i(\Phi_A))$.

It is easy to see that, by means of the mental actions, all the agents are able to derive $do_1(put_A(doll, box))$ and, consequently, $do_1^P(put_A(doll, box))$, which is the past event that records, in the agents' memory, that the action has indeed been performed. For each i, there will also be the general knowledge about the affect of the physical action of moving/putting an object into a place, i.e., that the object will indeed be in that place:
$\mathbf{K}_i(do_i^P(put_A(Obj, Place)) \wedge \neg do_i^P(move_A(Obj, Place, Place1)) \rightarrow in(Obj, Place))$
$\mathbf{K}_i(do_i^P(put_A(Obj, Place)) \wedge do_i^P(move_A(Obj, Place, Place1)) \rightarrow in(Obj, Place1))$

Therefore, for all agents (again by means of the mental actions that may exploit both knowledge and beliefs to infer new beliefs) we will easily have that $\mathbf{B}_{i,j}(in(doll, box))$ for $i, j = 1, 2, 3$.

Each agent i of the group also has the knowledge describing the effects of leaving a room. We represent the fact that someone leaves the room as the execution of an action $join_A$, for an agent to join a new group. In this case, the new group will be newly formed, and will contain the agent alone (in general, an agent would be able to join any existing group, but in this example there is no other group to join):
$\mathbf{K}_i(can\_do_i(join_A(i, j)) \wedge intend_i(join_A(i, j)) \rightarrow do_i(join_A(i, j)))$,
This piece of knowledge states that any agent who can and wishes to leave the room/-group will do so. Thus, if we have (for each i) $\mathbf{B}_{i,i}(can\_do_1(join_A(1, 1)))$ and $\mathbf{B}_{i,i}(intend_1(join_A(1, 1)))$, i.e., that Sally can and wants to leave the room, then all the agents conclude $do_1(join_A(1, 1))$ and, consequently, $do_1^P(join_A(1, 1))$. So, Sally is no longer in the group with agents 2 and 3. At this point therefore, agent 1 (Sally) is no longer able to observe what the others do. Let us assume, that agent 2 (Ann) acquires some other beliefs, possibly via interaction with the environment, of which Sally cannot be aware, being away from the group. Let us assume that the new beliefs are (for $i = 2, 3$): $\mathbf{B}_{i,i}(can\_do_2(move_A(doll, box, basket)))$ and $\mathbf{B}_{i,i}(intend_2(move_A(doll, box, basket)))$. From these beliefs and previous rules, via mental actions, agents 2 and 3 obtain $do_2(move_A(doll, box, basket))$, $do_2^P(move_A(doll, box, basket))$, and consequently, conclude that $\mathbf{B}_{i,j}(in(doll, basket))$ for $i, j = 2, 3$.

Now, assume that we ask agent 3 (the observer) what agent 1 (Sally) would answer if asked where the doll is now. This accounts, for agent 3, to prove $\mathbf{B}_{3,1}(in(doll, Place))$ for any constant replacing the variable $Place$ (namely, to find a suitable instantiation of the variable $Place$ to one

between box and $basket$). So, agent 3 puts herself "in the shoes" of agent 1. This involves, for agent 3, "simulating" what agent 1 would be able to infer at this stage. The semantics for belief update that we introduced for $\mathbf{B}_{i,j}$, applied here to $\mathbf{B}_{3,1}$, makes agent 3 able to reason in agent 1's neighborhood (i.e., within agent 1's beliefs). So, 3 concludes immediately $\mathbf{B}_{3,1}(in(doll, box))$. We have seen that agents 2 and 3 have instead concluded $in(doll, basket)$, inference that agent 3 cannot do when she puts herself "in the shoes" of agent 1, which, having left the group, does not have the necessary beliefs available, which were formed after she left. So, translating the result in "human" terms, agent 3 is able to answer that agent 1 believes the doll to be (still) in the box.

5. Conclusions

In this paper, we discussed cognitive aspects of an epistemic logic that we have proposed and developed to provide a way for a formal description of the cooperative activities of groups of agents. We have emphasized the importance of cognitive aspects, i.e.: having in the semantics particular functions that formalize the "way of thinking" of agents, and modelling the possibility for an agent to leave a group, thus losing the possibility, from then on, to be aware of the (changes in the) beliefs of its former group. We have shown how, in our logic, one can represent situations where the group's behaviour can be customized according to the context at hand, and we have discussed how to solve the well-known Sally-Anne task.

In past work, we have introduced the notion of canonical model of our logic, and we have performed the proof of strong completeness w.r.t. the proposed class of models (by means of a standard canonical-model argument). For lack of space we could not insert the proof in this paper, but we may note that it is very similar to that presented in [16]. The complexity of *L-DINF* is discussed in [16], which is the same as that of other similar logics.

In future work, we mean to extend the formalization of group dynamics, e.g., to better model the evolution of the beliefs of an agent before joining a group, during the stay, after leaving, and (possibly) after re-joining. We also intend to incorporate (at least aspects of) our logic into the DALI language: this involves devising a user-friendly and DALI-like syntax, enhancing the language semantics, and extending the implementation.

References

[1] S. Costantini, V. Pitoni, Towards a logic of "inferable" for self-aware transparent logical agents, in: C. Musto, D. Magazzeni, S. Ruggieri, G. Semeraro (Eds.), Proceedings of XAI.it@AIxIA 2020, volume 2742 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2020, pp. 68–79. URL: http://ceur-ws.org/Vol-2742/paper6.pdf.

[2] S. Costantini, A. Formisano, V. Pitoni, An epistemic logic for formalizing group dynamics of agents, Submitted to a journal (2022).

[3] A. I. Goldman, Theory of mind, in: E. Margolis, R. Samuels, S. P. Stich (Eds.), The Oxford Handbook of Philosophy of Cognitive Science, volume 1, Oxford University Press, 2012, pp. 402–424. doi:10.1093/oxfordhb/9780195309799.013.0017.

[4] S. Costantini, Towards active logic programming, in: Online Proceedings of the 2nd

Workshop on Component-based Software Development in Computational Logic, 1999. http://pages.di.unipi.it/brogi/ResearchActivity/COCL99/proceedings/index.html.

[5] S. Costantini, A. Tocchio, A logic programming language for multi-agent systems, in: S. Flesca, S. Greco, N. Leone, G. Ianni (Eds.), Proc. of JELIA-02, volume 2424 of *LNAI*, Springer, 2002, pp. 1–13. doi:10.1007/3-540-45757-7_1.

[6] S. Costantini, A. Tocchio, The DALI logic programming agent-oriented language, in: J. J. Alferes, J. A. Leite (Eds.), Proc. of JELIA-04, volume 3229 of *LNAI*, Springer, 2004, pp. 685–688. doi:10.1007/978-3-540-30227-8_57.

[7] G. De Gasperis, S. Costantini, G. Nazzicone, DALI multi agent systems framework, doi 10.5281/zenodo.11042, DALI GitHub Software Repository, 2014. DALI: github.com/AAAI-DISIM-UnivAQ/DALI.

[8] S. Costantini, A. Tocchio, About declarative semantics of logic-based agent languages, in: M. Baldoni, U. Endriss, A. Omicini, P. Torroni (Eds.), Declarative Agent Languages and Technologies III, volume 3904 of *LNCS*, Springer, 2005, pp. 106–123. doi:10.1007/11691792_7.

[9] S. Costantini, G. De Gasperis, G. Nazzicone, DALI for cognitive robotics: Principles and prototype implementation, in: Y. Lierler, W. Taha (Eds.), Practical Aspects of Declarative Languages - 19th Int. Symp. PADL 2017, Proceedings, volume 10137 of *LNCS*, Springer, 2017, pp. 152–162. doi:10.1007/978-3-319-51676-9_10.

[10] F. Aielli, D. Ancona, P. Caianiello, S. Costantini, G. De Gasperis, A. Di Marco, A. Ferrando, V. Mascardi, FRIENDLY & KIND with your health: Human-friendly knowledge-intensive dynamic systems for the e-health domain, in: J. B. et al. (Ed.), Highlights of Practical Applications of Scalable Multi-Agent Systems, volume 616 of *Communications in Computer and Information Science*, Springer, 2016, pp. 15–26. doi:10.1007/978-3-319-39387-2_2.

[11] S. Costantini, L. De Lauretis, C. Ferri, J. Giancola, F. Persia, A smart health assistant via DALI logical agents, in: S. Monica, F. Bergenti (Eds.), Proc. of CILC 2021, volume 3002 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2021, pp. 173–187.

[12] S. Costantini, A. Formisano, V. Pitoni, An epistemic logic for multi-agent systems with budget and costs, in: W. Faber, G. Friedrich, M. Gebser, M. Morak (Eds.), Proc. of JELIA-21, volume 12678 of *LNCS*, Springer, 2021, pp. 101–115. doi:10.1007/978-3-030-75775-5_8.

[13] A. S. Rao, M. P. Georgeff, Modeling rational agents within a BDI architecture, in: Proc. of the 2nd Int. Conf. on Principles of Knowledge Representation and Reasoning (KR'91), Morgan Kaufmann, 1991, pp. 473–484. ISBN:1-55860-165-1.

[14] H. Van Ditmarsch, J. Y. Halpern, W. Van Der Hoek, B. Kooi, Handbook of Epistemic Logic, College Publications, 2015. ISBN:1-84890-158-5.

[15] R. W. Weyhrauch, Prolegomena to a theory of mechanized formal reasoning, Artif. Intell. 13 (1980) 133–170. doi:10.1016/0004-3702(80)90015-6.

[16] S. Costantini, A. Formisano, V. Pitoni, An epistemic logic for modular development of multi-agent systems, in: N. Alechina, M. Baldoni, B. Logan (Eds.), Engineering Multi-Agent Systems 9th International Workshop, EMAS 2021, Revised Selected papers, volume 13190 of *LNCS*, Springer, 2022, pp. 72–91. doi:10.1007/978-3-030-97457-2_5.

[17] H. Wimmer, J. Perner, Beliefs about beliefs: Representation and constraining function of wrong beliefs in young children's understanding of deception, Cognition 13 (1983) 103–128. doi:10.1016/0010-0277.

[18] S. Baron-Cohen, A. M.Leslie, U. Frith, Does the autistic child have a "theory of mind"?, Cognition 1 (1985) 37–46. doi:10.1016/0010-0277(85)90022-8.

[19] L. Dissing, T. Bolander, Implementing theory of mind on a robot using dynamic epistemic logic, in: C. Bessiere (Ed.), Proc. of IJCAI 2020, ijcai.org, 2020, pp. 1615–1621. doi:10.24963/ijcai.2020/224.