# A Simple Framework for Cognitive Planning

Jorge Luis Fernandez Davila, Dominique Longin, Emiliano Lorini, Frédéric Maris

# A Simple Framework for Cognitive Planning

**Jorge Luis Fernandez Davila,**[1] **Dominique Longin,**[2] **Emiliano Lorini,**[2] **Frédéric Maris**[1]
[1]IRIT, Toulouse University, France
[2]IRIT, CNRS, Toulouse University, France
Jorge.Fernandez@irit.fr, Dominique.Longin@irit.fr, Emiliano.Lorini@irit.fr, Frederic.Maris@irit.fr

**Abstract**

We present a novel approach to cognitive planning, i.e., an agent's planning aimed at changing the cognitive attitudes of another agent including her beliefs and intentions. We encode the cognitive planning problem in an epistemic logic with a semantics exploiting belief bases. We study a NP-fragment of the logic whose satisfiability problem is reduced to SAT. We provide complexity results for the cognitive planning problem. Moreover, we illustrate its potential for applications in human-machine interaction in which an artificial agent is expected to interact with a human agent through dialogue and to persuade the human to behave in a certain way.

## Introduction

In social sciences, influence is defined as "change in an individual's thoughts, feelings, attitudes, or behaviors that results from interaction with another individual or a group" (Rashotte 2009). It is conceived as tightly connected with persuasion. The latter is the intentional form of influence in which an agent (the persuader) tries to make someone (the persuadee) do or believe something by giving her a good reason (Cialdini 2001; Perloff 2003).

Models of persuasion in AI are mostly based on argumentation. (See (Prakken 2006) for a general introduction to the research in this area.) Some of these models are built on Walton & Krabbe's notion of persuasion dialogue in which one party seeks to persuade another party to adopt a belief or point-of-view she does not currently hold (Walton and Krabbe 1995). There exist models based on abstract argumentation (Bench-Capon 2003; Bonzon and Maudet 2011; Amgoud, Maudet, and Parsons 2000) as well probabilistic models where the persuader's uncertainty about what the persuadee knows or believes is represented (Hunter 2015). There exist also models based on possibility theory in which a piece of information is represented as an argument which can be more or less accepted depending on the trustworthiness of the agent who proposes it (Da Costa Pereira, Tettamanzi, and Villata 2011). Persuasion has also been formalized with the support of logical tools, e.g., by combining abstract argumentation with dynamic epistemic logic (DEL) (Proietti and Yuste-Ginel 2019) and epistemic logic with dynamic logic (Budzyńska and Kacprzak 2008).

Epistemic planning is a generalization of classical planning that has been increasingly studied in AI in the last years. The goal to be achieved is not necessarily a state of the world but some belief states of one or more agents. This requires a theory of mind by the planning agent (Goldman 2006). A typical goal in epistemic planning is to make a certain agent believe something. Such a belief of the persuadee may be a higher-order belief, i.e., a belief about another agents' beliefs. The initial proposal was to use a standard logic of knowledge or belief together with a representation of actions in terms of event models of DEL (Bolander and Andersen 2011; Löwe, Pacuit, and Witzel 2011). While the DEL framework is very expressive, it turned out that the existence of a solution becomes quickly undecidable even for very simple kinds of event models (Aucher and Bolander 2013; Bolander, Holm Jensen, and Schwarzentruber 2015; Lê Cong, Pinchinat, and Schwarzentruber 2018). Kominis and Geffner considered epistemic planning problems with very simple event models leading to a decidable fragment (Kominis and Geffner 2015). They distinguish three kinds of actions: physical actions modifying the world, public updates (DEL-like public announcements), and sensing actions by means of which an agent learns whether a formula is true. Other researchers investigated another source of complexity, namely that of standard epistemic logic. There, reasoning is strictly more complex than in classical logic: the satisfiability problem is at least in PSPACE (Halpern and Moses 1992). Based on earlier work by Levesque, Muise et al. studied epistemic planning in fragments of standard epistemic logic (Muise et al. 2015). They considered state descriptions in terms of conjunctions of epistemic literals: formulas that do not contain any conjunction or disjunction. Cooper et al. considered another fragment: boolean combinations of 'knowing-whether' operators followed by propositional variables (Cooper et al. 2016).

Our approach pushes the envelope of the above approaches to epistemic planning. Our main contribution is its generalization to cognitive planning: it is not only some belief state of a target agent that is to be achieved, but more generally a cognitive state. The latter could involve not only beliefs, but also intentions. Cognitive planning makes clear the distinction between *persuasion on beliefs* (i.e., inducing someone to believe that a certain fact is true) and *persuasion on intentions* (i.e., inducing someone to form a certain

intention) and elucidates the connection between these two notions. Specifically, since beliefs are the input of decision-making and provide reasons for deciding and for acting, the persuader can indirectly change the persuadee's intentions by changing her beliefs, through the execution of a sequence of speech acts. In other words, in cognitive planning, the persuader tries to modify the persuadee's beliefs *in order to* affect persuadee's intentions. Moreover, cognitive planning takes resource boundedness and limited rationality of the persuadee seriously. For this reason, it is particularly well-suited for human-machine interaction (HMI) applications in which an artificial agent is expected to interact with a human — who is by definition resource-bounded — through dialogue and to persuade her to behave in a certain way.

To meet the previous requirements of cognitive planning, we formalize it in a recently proposed variant of epistemic logic with operators for explicit and implicit belief (Lorini 2018, 2020). The logic allows us to represent, at the same time: (i) the limited reasoning of the human agent (the persuadee), whose explicit beliefs are not necessarily closed under deduction, and (ii) the unbounded inferential capability of the artificial agent (the persuader), which is capable of computing the logical consequences of its explicit beliefs and of finding an optimal persuasion plan.

The paper is organized as follows. We first introduce the language of explicit and implicit belief for the specification of the cognitive planning problem. Given that the satisfiability problem for the full language is PSPACE-hard, we study an interesting NP-fragment of it that can be used in the context of a real HMI application. Then, we present an extension of the language by the notion of belief base expansion which is necessary for representing the actions of the planning agent. The second part of the paper is devoted to cognitive planning. We first formulate the cognitive planning problem and study its complexity. Then, we instantiate it in a concrete example in which an artificial assistant has to help a human user to choose a sport to practice in her leisure time. To achieve its goal, the agent needs to have a model of the user's beliefs and desires. Thanks to this model, the agent will be able to plan a sequence of speech acts aimed at persuading the user that a certain sport is the ideal one for her and, consequently, at inducing the user to form the intention to practice it.

## A Language for Explicit and Implicit Belief

This section is devoted to present the language which will serve as a specification language for cognitive planning. We start by recalling the full language and the semantics presented in (Lorini 2018, 2020). This language distinguishes explicit belief (a fact in an agent's belief base), from implicit belief (a fact that is deducible from the agent's explicit beliefs). Then, we study its NP-fragment. Finally, we focus on the extension of the latter by belief base expansion operators.

### Full Language and Semantics

Assume a countably infinite set of atomic propositions $Atm$ and a finite set of agents $Agt = \{1, \ldots, n\}$. The language is defined in two steps. First, the language $\mathcal{L}_0(Atm, Agt)$ is defined by the following grammar in BNF:

$$\alpha \quad ::= \quad p \mid \neg\alpha \mid \alpha_1 \wedge \alpha_2 \mid \alpha_1 \vee \alpha_2 \mid \triangle_i \alpha,$$

where $p$ ranges over $Atm$ and $i$ ranges over $Agt$. $\mathcal{L}_0(Atm, Agt)$ is the language for representing agents' explicit beliefs. The formula $\triangle_i\alpha$ is read "$i$ explicitly believes that $\alpha$". Then, the language $\mathcal{L}(Atm, Agt)$ extends the language $\mathcal{L}_0(Atm, Agt)$ by modal operators of implicit belief and is defined by the following grammar:

$$\varphi \quad ::= \quad \alpha \mid \neg\varphi \mid \varphi_1 \wedge \varphi_2 \mid \varphi_1 \vee \varphi_2 \mid \Box_i \varphi \mid \Diamond_i \varphi,$$

where $\alpha$ ranges over $\mathcal{L}_0(Atm, Agt)$ and $i$ ranges over $Agt$. For notational convenience we write $\mathcal{L}_0$ instead of $\mathcal{L}_0(Atm, Agt)$ and $\mathcal{L}$ instead of $\mathcal{L}(Atm, Agt)$, when the context is unambiguous. The formula $\Box_i\varphi$ is read "$i$ implicitly believes that $\varphi$" and $\Diamond_i\varphi$ is read "$\varphi$ is compatible (or consistent) with $i$'s explicit beliefs". The other Boolean constructions $\top$, $\bot$, $\rightarrow$ and $\leftrightarrow$ are defined in the standard way. We introduce here $\vee$ and $\Diamond_i$ as primitive and do not define them from $\wedge$ and $\Box_i$ because at a later stage we will need them for translating formulas in negation normal form.

The interpretation of language $\mathcal{L}$ exploits the notion of belief base. While the notions of possible state (or world) and epistemic alternative are primitive in the standard Kripke semantics for epistemic logic, they are defined from the primitive concept of belief base in this semantics. In particular, a state is a composite object including a description of both the agents' belief bases and the environment.[1]

**Definition 1 (State)** *A state is a tuple $B = (B_1, \ldots, B_n, V)$ where: for every $i \in Agt$, $B_i \subseteq \mathcal{L}_0$ is agent $i$'s belief base; $V \subseteq Atm$ is the actual environment. The set of all states is noted* **S**.

Note that an agent's belief base $B_i$ can be infinite. The sublanguage $\mathcal{L}_0(Atm, Agt)$ is interpreted w.r.t. states, as follows:

**Definition 2 (Satisfaction)** *Let $B = (B_1, \ldots, B_n, V) \in$ **S**. Then:*

$$\begin{aligned}
B \models p &\iff p \in V,\\
B \models \neg\alpha &\iff B \not\models \alpha,\\
B \models \alpha_1 \wedge \alpha_2 &\iff B \models \alpha_1 \text{ and } B \models \alpha_2,\\
B \models \alpha_1 \vee \alpha_2 &\iff B \models \alpha_1 \text{ or } B \models \alpha_2,\\
B \models \triangle_i \alpha &\iff \alpha \in B_i.
\end{aligned}$$

Observe in particular the set-theoretic interpretation of the explicit belief operator: agent $i$ explicitly believes that $\alpha$ if and only if $\alpha$ is included in her belief base.

A multi-agent belief model (MAB) is defined to be a state supplemented with a set of states, called *context*. The latter includes all states that are compatible with the common ground (Stalnaker 2002), i.e., the body of information that the agents commonly believe to be the case.

[1] This is similar to the way states are modeled in the interpreted system semantics for multi-agent systems (Fagin et al. 1995; Lomuscio, Qu, and Raimondi 2017).

**Definition 3 (Multi-Agent Belief Model)** *A multi-agent belief model (MAB) is a pair* $(B, Cxt)$*, where* $B \in \mathbf{S}$ *and* $Cxt \subseteq \mathbf{S}$*. The class of all MABs is noted* $\mathbf{M}$*.*

Note that we do not impose that $B \in Cxt$. When $Cxt = \mathbf{S}$ then $(B, Cxt)$ is said to be *complete*, since $\mathbf{S}$ is conceivable as the complete (or universal) context which contains all possible states. We compute an agent's set of epistemic alternatives from the agent's belief base, as follows.

**Definition 4 (Epistemic alternatives)** *Let* $i \in Agt$*. Then* $\mathcal{R}_i$ *is the binary relation on the set* $\mathbf{S}$ *such that, for all* $B = (B_1, \ldots, B_n, V), B' = (B'_1, \ldots, B'_n, V') \in \mathbf{S}$*:*

$$B\mathcal{R}_i B' \text{ if and only if } \forall \alpha \in B_i : B' \models \alpha.$$

$B\mathcal{R}_i B'$ means that $B'$ is an epistemic alternative for agent $i$ at $B$. So $i$'s set of epistemic alternatives at $B$ includes exactly those states that satisfy all $i$'s explicit beliefs.

Definition 5 extends Definition 2 to the full language $\mathcal{L}$. Its formulas are interpreted with respect to MABs. We omit Boolean cases that are defined in the usual way.

**Definition 5 (Satisfaction)** *Let* $(B, Cxt) \in \mathbf{M}$*. Then:*

$$\begin{aligned}
(B, Cxt) \models \alpha &\Longleftrightarrow B \models \alpha,\\
(B, Cxt) \models \Box_i \varphi &\Longleftrightarrow \forall B' \in Cxt,\\
&\qquad \text{if } B\mathcal{R}_i B' \text{ then } (B', Cxt) \models \varphi,\\
(B, Cxt) \models \Diamond_i \varphi &\Longleftrightarrow \exists B' \in Cxt \text{ such that}\\
&\qquad B\mathcal{R}_i B' \text{ and } (B', Cxt) \models \varphi.
\end{aligned}$$

A formula $\varphi \in \mathcal{L}$ is valid in the class $\mathbf{M}$, noted $\models_{\mathbf{M}} \varphi$, if and only if $(B, Cxt) \models \varphi$ for every $(B, Cxt) \in \mathbf{M}$; it is satisfiable in $\mathbf{M}$ if and only if $\neg\varphi$ is not valid in $\mathbf{M}$.

**Theorem 1** *Checking satisfiability of* $\mathcal{L}(Atm, Agt)$ *formulas in the class* $\mathbf{M}$ *is a PSPACE-hard problem.*

This theorem is a consequence of the fact that our logic contains the basic modal logic K whose satisfiability problem is PSPACE-complete (Halpern and Moses 1992).

## NP-Complete Fragment

In this section, we study the following fragment of the language $\mathcal{L}$, called $\mathcal{L}_{\mathsf{Frag}}$:

$$\varphi \quad ::= \quad \alpha \mid \neg\varphi \mid \varphi_1 \wedge \varphi_2 \mid \varphi_1 \vee \varphi_2 \mid \Box_{\mathfrak{m}}\alpha \mid \Diamond_{\mathfrak{m}}\alpha,$$

where $\alpha$ ranges over $\mathcal{L}_0$ and $\mathfrak{m}$ is a special agent in $Agt$ called the 'machine'. In $\mathcal{L}_{\mathsf{Frag}}$, all agents have explicit beliefs but only agent $\mathfrak{m}$ has implicit beliefs, and moreover the latter are restricted to $\mathcal{L}_0$ formulas of type $\alpha$. So there are no nested implicit beliefs for agent $\mathfrak{m}$. Agent $\mathfrak{m}$ is assumed to be the unique artificial agent in the system which is endowed with unbounded reasoning and planning capabilities. The cognitive planning problem will be modeled from agent $\mathfrak{m}$'s perspective.

In the rest of this section, we are going to provide a polysize reduction of the satisfiability problem of $\mathcal{L}_{\mathsf{Frag}}$ to SAT. The reduction consists of three steps which are summarized in Figure 1. As a first step, we put $\mathcal{L}_{\mathsf{Frag}}$ formulas in negation

$$\mathcal{L}_{\mathsf{Frag}} \overset{nnf}{\dashrightarrow} \mathcal{L}^{\mathsf{NNF}}_{\mathsf{Frag}} \overset{tr_1}{\dashrightarrow} \mathcal{L}_{\mathsf{Mod}} \overset{tr_2}{\dashrightarrow} \mathcal{L}_{\mathsf{Prop}}$$

Figure 1: Summary of reduction process

normal form (NNF) via the following function $nnf$:

$$\begin{aligned}
nnf(p) &= p,\\
nnf(\triangle_i\alpha) &= \triangle_i\alpha,\\
nnf(\Box_{\mathfrak{m}}\alpha) &= \Box_{\mathfrak{m}}\, nnf(\alpha),\\
nnf(\Diamond_{\mathfrak{m}}\alpha) &= \Diamond_{\mathfrak{m}}\, nnf(\alpha),\\
nnf(\varphi \wedge \psi) &= nnf(\varphi) \wedge nnf(\psi),\\
nnf(\varphi \vee \psi) &= nnf(\varphi) \vee nnf(\psi),\\
nnf(\neg p) &= \neg p,\\
nnf(\neg\triangle_i\alpha) &= \neg\triangle_i\alpha,\\
nnf(\neg\neg\varphi) &= nnf(\varphi),\\
nnf\big(\neg(\varphi \wedge \psi)\big) &= nnf(\neg\varphi \vee \neg\psi),\\
nnf\big(\neg(\varphi \vee \psi)\big) &= nnf(\neg\varphi \wedge \neg\psi),\\
nnf(\neg\Box_{\mathfrak{m}}\alpha) &= \Diamond_{\mathfrak{m}}\, nnf(\neg\alpha),\\
nnf(\neg\Diamond_{\mathfrak{m}}\alpha) &= \Box_{\mathfrak{m}}\, nnf(\neg\alpha).
\end{aligned}$$

Let us define the NNF variant $\mathcal{L}_0^{\mathsf{NNF}}$ of the language $\mathcal{L}_0$ by the following grammar:

$$\beta \quad ::= \quad p \mid \neg p \mid \triangle_i\alpha \mid \neg\triangle_i\alpha \mid \beta_1 \wedge \beta_2 \mid \beta_1 \vee \beta_2,$$

where $p$ ranges over $Atm$, $i$ ranges over $Agt$ and $\alpha$ ranges over $\mathcal{L}_0$. Furthermore, let us define the language $\mathcal{L}^{\mathsf{NNF}}_{\mathsf{Frag}}$ by the following grammar. For $\beta$ ranging over $\mathcal{L}_0^{\mathsf{NNF}}$:

$$\varphi \quad ::= \quad \beta \mid \varphi_1 \wedge \varphi_2 \mid \varphi_1 \vee \varphi_2 \mid \Box_{\mathfrak{m}}\beta \mid \Diamond_{\mathfrak{m}}\beta.$$

**Proposition 1** *Let* $\varphi \in \mathcal{L}_{\mathsf{Frag}}$*. Then,* $\varphi \leftrightarrow nnf(\varphi)$ *is valid in the class* $\mathbf{M}$*, and* $nnf(\varphi) \in \mathcal{L}^{\mathsf{NNF}}_{\mathsf{Frag}}$*.*

Note that the size of $nnf(\varphi)$ is polynomial in the size of $\varphi$.

As a second step, we define the following modal language $\mathcal{L}_{\mathsf{Mod}}$ into which the language $\mathcal{L}^{\mathsf{NNF}}_{\mathsf{Frag}}$ will be translated:

$$\begin{aligned}
\omega \quad &::= \quad q \mid \neg\omega \mid \omega_1 \wedge \omega_2 \mid \omega_1 \vee \omega_2\ ,\\
\varphi \quad &::= \quad q \mid \neg\varphi \mid \varphi_1 \wedge \varphi_2 \mid \varphi_1 \vee \varphi_2 \mid \blacksquare\omega \mid \blacklozenge\omega\ ,
\end{aligned}$$

where $q$ ranges over the following set of atomic formulas:

$$Atm^+ = Atm \cup \{p_{\triangle_i\alpha} : i \in Agt \text{ and } \alpha \in \mathcal{L}_0(Atm, Agt)\}.$$

So $p_{\triangle_i\alpha}$ is nothing but a special propositional variable.

We interpret the language $\mathcal{L}_{\mathsf{Mod}}$ w.r.t. a pair $(M, w)$, called pointed Kripke model, where $M = (W, \Rightarrow, \pi)$, $W$ is a non-empty set of worlds, $\Rightarrow \subseteq W \times W$ and $\pi : Atm^+ \longrightarrow 2^W$. (Boolean cases are again omitted as they are defined in the usual way.)

**Definition 6** *The semantic interpretation for formulas in* $\mathcal{L}_{\mathsf{Mod}}$ *w.r.t. a pointed Kripke model* $(M, w)$ *is as follows:*

$$\begin{aligned}
(M, w) \models q &\Longleftrightarrow w \in \pi(q);\\
(M, w) \models \blacksquare\omega &\Longleftrightarrow \forall v \in W, \text{ if } w \Rightarrow v \text{ then } (M, v) \models \omega;\\
(M, w) \models \blacklozenge\omega &\Longleftrightarrow \exists v \in W \text{ s.t. } w \Rightarrow v \text{ and } (M, v) \models \omega.
\end{aligned}$$

The class of pointed Kripke models is noted $\mathbf{K}$. Satisfiability and validity of formulas in $\mathcal{L}_{\mathsf{Mod}}$ relative to the class $\mathbf{K}$ is defined in the usual way.

Let $tr_1 : \mathcal{L}^{\mathsf{NNF}}_{\mathsf{Frag}} \longrightarrow \mathcal{L}_{\mathsf{Mod}}$ be a translation such that:

$$\begin{aligned}
tr_1(p) &= p,\\
tr_1(\neg p) &= \neg p,\\
tr_1(\varphi_1 \wedge \varphi_2) &= tr_1(\varphi_1) \wedge tr_1(\varphi_2),\\
tr_1(\varphi_1 \vee \varphi_2) &= tr_1(\varphi_1) \vee tr_1(\varphi_2),\\
tr_1(\triangle_i \alpha) &= \begin{cases} p_{\triangle_{\mathfrak{m}}\alpha} \wedge \blacksquare tr_0(\alpha), & \text{if } i = \mathfrak{m},\\ p_{\triangle_i \alpha}, & \text{otherwise,}\end{cases}\\
tr_1(\neg\triangle_i \alpha) &= \neg p_{\triangle_i \alpha},\\
tr_1(\Box_{\mathfrak{m}} \beta) &= \blacksquare tr_0(\beta),\\
tr_1(\Diamond_{\mathfrak{m}} \beta) &= \blacklozenge tr_0(\beta);
\end{aligned}$$

with $tr_0 : \mathcal{L}_0 \longrightarrow \mathcal{L}_{\mathsf{Mod}}$ such that:

$$\begin{aligned}
tr_0(p) &= p,\\
tr_0(\neg\alpha) &= \neg tr_0(\alpha),\\
tr_0(\alpha_1 \wedge \alpha_2) &= tr_0(\alpha_1) \wedge tr_0(\alpha_2),\\
tr_0(\alpha_1 \vee \alpha_2) &= tr_0(\alpha_1) \vee tr_0(\alpha_2),\\
tr_0(\triangle_i \alpha) &= p_{\triangle_i \alpha}.
\end{aligned}$$

As the following theorem indicates, the polynomial translation $tr_1$ guarantees the transfer of satisfiability from model class $\mathbf{M}$ to model class $\mathbf{K}$.

**Theorem 2** *Let $\varphi \in \mathcal{L}^{\mathsf{NNF}}_{\mathsf{Frag}}$. Then, $\varphi$ is satisfiable in the class $\mathbf{M}$ if and only if $tr_1(\varphi)$ is satisfiable in the class $\mathbf{K}$.*

SKETCH OF PROOF. The proof relies on the fact that the belief base semantics for the language $\mathcal{L}_{\mathsf{Frag}}$ is equivalent to a "weaker" semantics exploiting pointed structures of the form $(X, s)$ where $X = (S, \mathcal{B}, (\Rightarrow_i)_{i\in Agt}, \tau)$, $S$ is a non-empty set of states, $s \in S$ is the actual state, $\mathcal{B} : Agt \times S \longrightarrow 2^{\mathcal{L}_0}$ is a belief base function, $\tau : Atm \longrightarrow 2^S$ is valuation function, $\Rightarrow_i \subseteq S \times S$ is agent $i$'s epistemic accessibility relation and with respect to which $\mathcal{L}$-formulas are interpreted as follows (boolean cases are omitted for simplicity): (i) $(X, s) \models p$ iff $s \in \tau(p)$, (ii) $(X, s) \models \triangle_i \alpha$ iff $\alpha \in \mathcal{B}(i, s)$, (iii) $(X, s) \models \Box_i \varphi$ iff $\forall s' \in S$, if $s \Rightarrow_i s'$ then $(X, s') \models \varphi$. In particular, for every $\varphi \in \mathcal{L}_{\mathsf{Frag}}$, we have that $\varphi$ is satisfiable in $\mathbf{M}$ iff $\varphi$ is satisfiable in the subclass of pointed structures $(X, s)$ such that $\Rightarrow_{\mathfrak{m}}(s) \subseteq \bigcap_{\alpha\in\mathcal{B}(\mathfrak{m},s)} ||\alpha||_{(X,s)}$ with $||\alpha||_{(X,s)} = \{s' \in X : (X, s') \models \alpha\}$. ■

As a last step, we provide a polysize reduction of $\mathcal{L}_{\mathsf{Mod}}$-satisfiability to SAT, where the underlying propositional logic language $\mathcal{L}_{\mathsf{Prop}}$ is built from the following set of atomic propositions:

$$\begin{aligned}
Atm^{++} = &\{q_x : q \in Atm^+ \text{ and } x \in \mathbb{N}\} \cup\\
&\{r_{x,y} : x, y \in \mathbb{N}\}.
\end{aligned}$$

The set $Atm^{++}$ includes two types of atomic propositions: one of the form $q_x$ denoting the fact that $q$ is true at world $x$ and the other of the form $r_{x,y}$ denoting the fact that world $x$ is related to world $y$.

Let $tr_2 : \mathcal{L}_{\mathsf{Mod}} \times \mathbb{N} \times \mathbb{N} \longrightarrow \mathcal{L}_{\mathsf{Prop}}$ be the following translation function:

$$\begin{aligned}
tr_2(q, x, y) &= q_x,\\
tr_2(\neg\varphi, x, y) &= \neg tr_2(\varphi, x, y),\\
tr_2(\varphi_1 \wedge \varphi_2, x, y) &= tr_2(\varphi_1, x, y) \wedge tr_2(\varphi_2, x, y),\\
tr_2(\varphi_1 \vee \varphi_2, x, y) &= tr_2(\varphi_1, x, y) \vee tr_2(\varphi_2, x, y),\\
tr_2(\blacksquare\omega, x, y) &= \bigwedge_{0\le z\le y} \big(r_{x,z} \rightarrow tr_2(\omega, z, y)\big),\\
tr_2(\blacklozenge\omega, x, y) &= \bigvee_{0\le z\le y} \big(r_{x,z} \wedge tr_2(\omega, z, y)\big).
\end{aligned}$$

Translation $tr_2$ is similar to the translation of modal logic S5 into propositional logic given in (Caridroit et al. 2017) and, more generally, to the standard translation of modal logic into FOL in which accessibility relations are encoded by special predicates. The size of an $\mathcal{L}_{\mathsf{Mod}}$ formula, $size(\varphi)$, is defined by:

$$\begin{aligned}
size(p) &= 1,\\
size(\varphi_1 \wedge \varphi_2) &= size(\varphi_1) + size(\varphi_2) + 1,\\
size(\varphi_1 \vee \varphi_2) &= size(\varphi_1) + size(\varphi_2) + 1,\\
size(\neg\varphi) &= size(\varphi) + 1,\\
size(\blacksquare\omega) &= size(\blacklozenge\omega) = size(\omega) + 1.
\end{aligned}$$

Note that the size of $tr_2\big(\varphi, 0, size(\varphi)\big)$ is polynomial in the size of $\varphi$.

**Theorem 3** *Let $\varphi \in \mathcal{L}_{\mathsf{Mod}}$. Then, $\varphi$ is satisfiable in the class $\mathbf{K}$ if and only if $tr_2\big(\varphi, 0, size(\varphi)\big)$ is satisfiable in propositional logic.*

SKETCH OF PROOF. The theorem is proved in the same way as the standard translation of modal logic to FOL plus a straightforward adaptation of (Ladner 1977, Lemma 6.1) about polysize-model property for S5 to our case. ■

The size of $tr_2\big(\varphi, 0, size(\varphi)\big)$ being polynomial in the size of $\varphi$, thanks to Proposition 1, Theorem 2 and Theorem 3 we state the following complexity result.

**Theorem 4** *Checking satisfiability of formulas in $\mathcal{L}_{\mathsf{Frag}}$ in the class $\mathbf{M}$ is an NP-complete problem.*

## Dynamic Extension

In this section, we extend the language $\mathcal{L}_{\mathsf{Frag}}$ by belief expansion operations. Such an extension will allow us to represent the actions of the planning agent in the cognitive planning problem. Specifically, we introduce the following language $\mathcal{L}^{+}_{\mathsf{Frag}}$:

$$\varphi ::= \alpha \mid \neg\varphi \mid \varphi_1 \wedge \varphi_2 \mid \varphi_1 \vee \varphi_2 \mid \Box_{\mathfrak{m}}\alpha \mid \Diamond_{\mathfrak{m}}\alpha \mid [+_i\alpha]\varphi,$$

where $\alpha$ ranges over $\mathcal{L}_0$ and $i$ ranges over $Agt$. The formula $[+_i\alpha]\varphi$ is read "$\varphi$ holds after agent $i$ has privately expanded her belief base with $\alpha$". Events of type $+_i\alpha$ are generically called informative actions.

Our extension has the following semantics relative to a MAB:

**Definition 7 (Satisfaction relation, cont.)** *Let* $B = (B_1, \ldots, B_n, V) \in \mathbf{S}$ *and let* $(B, Cxt) \in \mathbf{M}$. *Then:*

$$(B, Cxt) \models [+_i\alpha]\varphi \iff (B^{+_i\alpha}, Cxt) \models \varphi$$

*with* $V^{+_i\alpha} = V$, $B_i^{+_i\alpha} = B_i \cup \{\alpha\}$ *and* $B_j^{+_i\alpha} = B_j$ *for all* $j \neq i$.

Intuitively speaking, the private expansion of $i$'s belief base by $\alpha$ simply consists in agent $i$ adding the information that $\alpha$ to her belief base, while all other agents keep their belief bases unchanged.

The following equivalences are valid in the class $\in \mathbf{M}$:

$$[+_i\alpha]\alpha' \leftrightarrow \begin{cases} \top, & \text{if } \alpha' = \triangle_i\alpha, \\ \alpha', & \text{otherwise;} \end{cases}$$

$$[+_i\alpha]\neg\varphi \leftrightarrow \neg[+_i\alpha]\varphi;$$

$$[+_i\alpha](\varphi_1 \wedge \varphi_2) \leftrightarrow [+_i\alpha]\varphi_1 \wedge [+_i\alpha]\varphi_2;$$

$$[+_i\alpha](\varphi_1 \vee \varphi_2) \leftrightarrow [+_i\alpha]\varphi_1 \vee [+_i\alpha]\varphi_2;$$

$$[+_i\alpha]\Box_{\mathfrak{m}}\alpha' \leftrightarrow \begin{cases} \Box_{\mathfrak{m}}(\alpha \to \alpha'), & \text{if } i = \mathfrak{m}, \\ \Box_{\mathfrak{m}}\alpha', & \text{otherwise;} \end{cases}$$

$$[+_i\alpha]\Diamond_{\mathfrak{m}}\alpha' \leftrightarrow \begin{cases} \Diamond_{\mathfrak{m}}(\alpha \wedge \alpha'), & \text{if } i = \mathfrak{m}, \\ \Diamond_{\mathfrak{m}}\alpha', & \text{otherwise.} \end{cases}$$

Thanks to these equivalences we can define the following reduction $red$ transforming every $\mathcal{L}^+_{\mathsf{Frag}}$ formula $\varphi$ into an equivalent $\mathcal{L}_{\mathsf{Frag}}$ formula $red(\varphi)$:

$$\begin{aligned}
red(p) &= p, \\
red(\triangle_i\alpha) &= \triangle_i\alpha, \\
red(\neg\varphi) &= \neg red(\varphi), \\
red(\varphi_1 \wedge \varphi_2) &= red(\varphi_1) \wedge red(\varphi_2), \\
red(\varphi_1 \vee \varphi_2) &= red(\varphi_1) \vee red(\varphi_2), \\
red(\Box_{\mathfrak{m}}\varphi) &= \Box_{\mathfrak{m}} red(\varphi), \\
red(\Diamond_{\mathfrak{m}}\varphi) &= \Diamond_{\mathfrak{m}} red(\varphi), \\
red([+_i\alpha]\alpha') &= \begin{cases} \top, & \text{if } \alpha' = \triangle_i\alpha, \\ red(\alpha'), & \text{otherwise;} \end{cases} \\
red([+_i\alpha]\neg\varphi) &= red(\neg[+_i\alpha]\varphi), \\
red\big([+_i\alpha](\varphi_1 \wedge \varphi_2)\big) &= red([+_i\alpha]\varphi_1 \wedge [+_i\alpha]\varphi_2), \\
red\big([+_i\alpha](\varphi_1 \vee \varphi_2)\big) &= red([+_i\alpha]\varphi_1 \vee [+_i\alpha]\varphi_2), \\
red([+_i\alpha]\Box_{\mathfrak{m}}\alpha') &= \begin{cases} red\big(\Box_{\mathfrak{m}}(\alpha \to \alpha')\big), & \text{if } i = \mathfrak{m}, \\ red(\Box_{\mathfrak{m}}\alpha') & \text{otherwise;} \end{cases} \\
red([+_i\alpha]\Diamond_{\mathfrak{m}}\alpha') &= \begin{cases} red\big(\Diamond_{\mathfrak{m}}(\alpha \wedge \alpha')\big) & \text{if } i = \mathfrak{m}, \\ red(\Diamond_{\mathfrak{m}}\alpha') & \text{otherwise;} \end{cases} \\
red([+_i\alpha_1][+_j\alpha_2]\varphi) &= red\big([+_i\alpha_1] red([+_j\alpha_2]\varphi)\big).
\end{aligned}$$

**Proposition 2** *Let* $\varphi \in \mathcal{L}^+_{\mathsf{Frag}}$. *Then,* $\varphi \leftrightarrow red(\varphi)$ *is valid in the class* $\mathbf{M}$, *and* $red(\varphi) \in \mathcal{L}_{\mathsf{Frag}}$.

The following theorem is a consequence of Theorem 4, Proposition 2 and the fact that the size of $red(\varphi)$ is polynomial in the size of $\varphi$.

**Theorem 5** *Checking satisfiability of formulas in* $\mathcal{L}^+_{\mathsf{Frag}}$ *in the class* $\mathbf{M}$ *is an NP-complete problem.*

Before concluding this section, we define the concept of logical consequence for the language $\mathcal{L}^+_{\mathsf{Frag}}$ which will be used in the formulation of the cognitive planning problem at a later stage. Let $\Sigma$ be a finite subset of $\mathcal{L}_0$ and let $\varphi \in \mathcal{L}^+_{\mathsf{Frag}}$. We say that $\varphi$ is a logical consequence of $\Sigma$ in the class $\mathbf{M}$, noted $\Sigma \models_{\mathbf{M}} \varphi$, if and only if, for every $(B, Cxt) \in \mathbf{M}$ such that $Cxt \subseteq \mathbf{S}(\Sigma)$ we have $(B, Cxt) \models \varphi$, with $\mathbf{S}(\Sigma) = \{B \in \mathbf{S} : \forall \alpha \in \Sigma, B \models \alpha\}$. We say that $\varphi$ is $\Sigma$-satisfiable in the class $\mathbf{M}$ if and only if, $\neg\varphi$ is not a logical consequence of $\Sigma$ in $\mathbf{M}$. Clearly, $\varphi$ is valid if and only if $\varphi$ is a logical consequence of $\emptyset$, and $\varphi$ is satisfiable if and only if $\varphi$ is $\emptyset$-satisfiable.

As the following deduction theorem indicates, the logical consequence problem with a finite set of premises can be reduced to the satisfiability problem.

**Theorem 6** *Let* $\varphi \in \mathcal{L}^+_{\mathsf{Frag}}$ *and let* $\Sigma \subset \mathcal{L}_0$ *be finite. Then,* $\Sigma \models_{\mathbf{M}} \varphi$ *if and only if* $\models_{\mathbf{M}} \bigwedge_{\alpha\in\Sigma} \Box_{\mathfrak{m}}\alpha \to \varphi$.

## Cognitive Planning

This section is devoted to the formal specification of cognitive planning by way of the language $\mathcal{L}^+_{\mathsf{Frag}}$ and to its illustration with an example of HMI.

### Problem

The planning problem in the context of the logic $\mathcal{L}^+_{\mathsf{Frag}}$ consists in finding a sequence of informative actions for agent $\mathfrak{m}$ of type $+_{\mathfrak{m}}\alpha$ which guarantees that it believes that its goal $\alpha_G$ is satisfied. As we emphasized above, agent $\mathfrak{m}$ is assumed to be an artificial agent which interacts with other resource-bounded human agents.

Let $Act_{\mathfrak{m}} = \{+_{\mathfrak{m}}\alpha : \alpha \in \mathcal{L}_0\}$ be agent $\mathfrak{m}$'s set of informative actions and let elements of $Act_{\mathfrak{m}}$ be noted $\epsilon, \epsilon', \ldots$ Agent $\mathfrak{m}$'s informative actions have executability preconditions that are specified by the following function: $\mathcal{P} : Act_{\mathfrak{m}} \longrightarrow \mathcal{L}_{\mathsf{Frag}}$. So, we can define the following operator of successful occurrence of an informative action:

$$\langle\!\langle\epsilon\rangle\!\rangle\varphi \stackrel{\text{def}}{=} \mathcal{P}(\epsilon) \wedge [\epsilon]\varphi$$

with $\epsilon \in Act_{\mathfrak{m}}$. The formula $\langle\!\langle\epsilon\rangle\!\rangle\varphi$ has to be read "agent $\mathfrak{m}$'s informative action $\epsilon$ can take place and $\varphi$ holds after its occurrence".

Informative actions of type 'speech act' are of interest here. In particular, we consider speech acts of type 'to inform', where $\mathfrak{m}$ is assumed to be the speaker and $j \in Agt$ such that $j \neq \mathfrak{m}$ is assumed to be the hearer. We identify the speech act "agent $\mathfrak{m}$ informs agent $j$ that $\alpha$" with the perlocutionary effect (Searle 1969, Sect. 6.2) of the speaker learning that the hearer has learnt that $\alpha$:[2]

$$inform(\mathfrak{m},j,\alpha) \stackrel{\text{def}}{=} +_{\mathfrak{m}} \triangle_j\alpha.$$

[2]We implicitly assume that, by default, $\mathfrak{m}$ believes that $j$ trusts its judgment, so that $j$ will believe what it says.

The previous abbreviation and, more generally, the idea of describing actions of a plan with agent $\mathfrak{m}$'s private belief expansion operations is justified by the fact that we model cognitive planning from the perspective of the planning agent $\mathfrak{m}$. Therefore, we only need to represent the effects of actions on agent $\mathfrak{m}$'s beliefs.

**Definition 8 ($\mathcal{L}^+_{\mathsf{Frag}}$-planning problem)** *A $\mathcal{L}^+_{\mathsf{Frag}}$-planning problem is a tuple $\langle \Sigma, Op, \alpha_G \rangle$ where:*

- $\Sigma \subset \mathcal{L}_0$ *is a finite set of agent $\mathfrak{m}$'s available information,*
- $Op \subset Act_{\mathfrak{m}}$ *is a finite set of agent $\mathfrak{m}$'s operators,*
- $\alpha_G \in \mathcal{L}_0$ *is agent $\mathfrak{m}$'s goal.*

A solution plan to a $\mathcal{L}^+_{\mathsf{Frag}}$-planning problem $\langle \Sigma, Op, \alpha_G \rangle$ is a sequence of operators $\epsilon_1, \ldots, \epsilon_k$ from $Op$ for some $k$ such that $\Sigma \models_{\mathbf{M}} \langle\!\langle \epsilon_1 \rangle\!\rangle \ldots \langle\!\langle \epsilon_k \rangle\!\rangle \Box_{\mathfrak{m}} \alpha_G$.

**Theorem 7** *Checking plan existence for a $\mathcal{L}^+_{\mathsf{Frag}}$-planning problem is in $\mathsf{NP}^{\mathsf{NP}} = \Sigma^{\mathsf{P}}_2$.*

SKETCH OF PROOF. A $\mathcal{L}^+_{\mathsf{Frag}}$-planning problem $\langle \Sigma, Op, \alpha_G \rangle$ has a solution plan if and only if it has a poly-size solution plan $\epsilon_1, \ldots, \epsilon_k$ with $k \leq |Op|$. Indeed, it is easily seen that if an operator has been executed in a plan, another future occurrence of the same operator will not change the planning state due to the monotonicity of private belief expansion:

$$\big(\big(\ldots\big(B_i^{+_i\alpha}\big)^{+_i\alpha_1\ldots}\big)^{+_i\alpha_h}\big)^{+_i\alpha} = \big(\ldots\big(B_i^{+_i\alpha}\big)^{+_i\alpha_1\ldots}\big)^{+_i\alpha_h}.$$

Consider a poly-time non-deterministic Turing machine with an NP-oracle ($\Sigma^{\mathsf{P}}_2$-Turing machine). It begins with an empty plan and branches over all poly-size plans of length $k \leq |Op|$ choosing non deterministically operators to add to the plan. It accepts if $\Sigma \models_{\mathbf{M}} \langle\!\langle \epsilon_1 \rangle\!\rangle \ldots \langle\!\langle \epsilon_k \rangle\!\rangle \Box_{\mathfrak{m}} \alpha_G$ i.e., using Theorem 6, if $\neg \big( \big( \bigwedge_{\alpha \in \Sigma} \Box_{\mathfrak{m}} \alpha \big) \rightarrow \langle\!\langle \epsilon_1 \rangle\!\rangle \ldots \langle\!\langle \epsilon_k \rangle\!\rangle \Box_{\mathfrak{m}} \alpha_G \big)$ is unsatisfiable in the class $\mathbf{M}$. Thanks to Theorem 5, unsatisfiability of this $\mathcal{L}^+_{\mathsf{Frag}}$ formula can be checked by the NP-oracle. When $k = |Op|$ and the formula is satisfiable, the Turing machine rejects. ■

## Example: Artificial Assistant

We consider a HMI scenario in which agent $\mathfrak{m}$ is the artificial assistant of the human agent $\mathfrak{h}$. Agent $\mathfrak{h}$ has to choose a sport to practice since her doctor recommended her to do a regular physical activity to be in good health. Agent $\mathfrak{m}$'s aim is to help agent $\mathfrak{h}$ to make the right choice, given her actual beliefs and desires. The finite set of sport activities from which $\mathfrak{h}$ can choose is noted $Opt$. Elements of $Opt$ are noted $o, o', \ldots$ Each option in $Opt$ is identified with a finite set of variables $Var$. Each variable $x$ in $Var$ takes a value from its corresponding finite set of values $Val_x$.

In this example, we suppose that $Opt$ is composed of the following eight elements: swimming (sw), running (ru), horse riding (hr), tennis (te), soccer (so), yoga (yo), diving (di) and squash (sq). Moreover, there are exactly six variables in $Var$ which are used to classify the available options: environment (**env**), location (**loc**), sociality (**soc**), cost (**cost**), dangerousness (**dan**) and intensity (**intens**). The set of values for the variables are:

$$\begin{aligned}
&Val_{\mathbf{env}} = \{land, water\},\\
&Val_{\mathbf{loc}} = \{indoor, outdoor, mixed\},\\
&Val_{\mathbf{soc}} = \{single, team, mixed\},\\
&Val_{\mathbf{cost}} = \{low, med, high\},\\
&Val_{\mathbf{dan}} = \{low, med, high\},\\
&Val_{\mathbf{intens}} = \{low, med, high\}.
\end{aligned}$$

The set of assignments for variable $x$ is defined as follows:

$$Assign_x = \{x \mapsto v : v \in Val_x\}.$$

The set of variable assignments is

$$Assign = \bigcup_{x \in Var} Assign_x.$$

Elements of $Assign$ are noted $a, a', \ldots$

We assume that the content of an atomic desire is a variable assignment or its negation. That is, agent $\mathfrak{h}$'s atomic desire can be any element from the following set:

$$Des_0 = Assign \cup \{\sim a : a \in Assign\}.$$

Elements of $Des_0$ are noted $d, d', \ldots$. For example, the fact that $\mathfrak{h}$ has $\mathbf{loc} \mapsto indoor$ as a desire means that $\mathfrak{h}$ would like to practice an indoor activity, while if $\mathfrak{h}$'s desire is $\sim \mathbf{cost} \mapsto high$, then $\mathfrak{h}$ would like to practice an activity whose cost is not high. Agent $\mathfrak{h}$'s desires are either atomic desires or conditional desires. That is, $\mathfrak{h}$'s desire can be any element from the following set:

$$Des = Des_0 \cup \big\{[d_1, \ldots, d_k] \rightsquigarrow d : d_1, \ldots, d_k, d \in Des_0\big\}.$$

Elements of $Des$ are noted $\gamma, \gamma', \ldots$ For example, if agent $\mathfrak{h}$ has $[\mathbf{cost} \mapsto high] \rightsquigarrow \mathbf{dan} \mapsto low$ as a desire, then she would like to practice a sport whose dangerousness level is low, if its cost is high. We define $2^{Des*} = 2^{Des} \setminus \emptyset$.

Let us assume that the set $Atm$ includes four types of atomic formulas, for every $x \mapsto v \in Assign$, $o, o' \in Opt$ and $\Gamma \in 2^{Des*}$: (i) $\mathsf{val}(o, x \mapsto v)$ standing for "option $o$ has value $v$ for variable $x$", (ii) $\mathsf{ideal}(\mathfrak{h}, o)$ standing for "$o$ is an ideal option for agent $\mathfrak{h}$", (iii) $\mathsf{justif}(\mathfrak{h}, o)$ standing for "agent $\mathfrak{h}$ has a justification for choosing option $o$", and (iv) $\mathsf{des}(\mathfrak{h}, \Gamma)$ standing for "$\Gamma$ is agent $\mathfrak{h}$'s set of desires".

The following function $f_{comp}$ specifies, for every option $o \in Opt$ and possible desire $\gamma \in Des$, the condition guaranteeing that $o$ satisfies (or, complies with) $\gamma$:

$$\begin{aligned}
&f_{comp}(o, a) = \mathsf{val}(o, a),\\
&f_{comp}(o, \sim a) = \neg \mathsf{val}(o, a),\\
&f_{comp}\big(o, [d_1, \ldots, d_k] \rightsquigarrow d\big) = \neg f_{comp}(o, d_1) \vee \ldots \vee\\
&\qquad\qquad \neg f_{comp}(o, d_k) \vee f_{comp}(o, d).
\end{aligned}$$

The following function $f^{\mathfrak{h}}_{comp}$ specifies, for every option $o \in Opt$ and possible desire $\gamma \in Des$, the condition guaranteeing that agent $\mathfrak{h}$ believes that $o$ satisfies $\gamma$:

$$\begin{aligned}
&f^{\mathfrak{h}}_{comp}(o, a) = \triangle_{\mathfrak{h}} f_{comp}(o, a),\\
&f^{\mathfrak{h}}_{comp}(o, \sim a) = \triangle_{\mathfrak{h}} f_{comp}(o, \sim a),\\
&f^{\mathfrak{h}}_{comp}\big(o, [d_1, \ldots, d_k] \rightsquigarrow d\big) = \triangle_{\mathfrak{h}} \neg f_{comp}(o, d_1) \vee \ldots \vee\\
&\qquad\qquad \triangle_{\mathfrak{h}} \neg f_{comp}(o, d_k) \vee \triangle_{\mathfrak{h}} f_{comp}(o, d).
\end{aligned}$$

The previous formulation of $f^{\mathfrak{h}}_{comp}\big(o, [d_1, \ldots, d_k] \rightsquigarrow d\big)$ presupposes an understanding of conditional (goal) sentences by agent $\mathfrak{h}$. In particular, agent $\mathfrak{m}$ does not need to provide information to agent $\mathfrak{h}$ about the antecedent of the conditional, if the consequent is true.

We assume that the artificial agent $\mathfrak{m}$ has the following pieces of information in its belief base:

$$\alpha_1 \stackrel{\text{def}}{=} \bigwedge_{\substack{o\in Opt\\ x\in Var\\ v,v'\in Val_x: v\neq v'}} \big(\mathsf{val}(o, x\mapsto v) \rightarrow \neg\mathsf{val}(o, x\mapsto v')\big),$$

$$\alpha_2 \stackrel{\text{def}}{=} \bigwedge_{\substack{o\in Opt\\ x\in Var\\ v,v'\in Val_x: v\neq v'}} \begin{pmatrix}\triangle_{\mathfrak{h}}\mathsf{val}(o, x\mapsto v) \rightarrow\\ \triangle_{\mathfrak{h}}\neg\mathsf{val}(o, x\mapsto v')\end{pmatrix},$$

$$\alpha_3 \stackrel{\text{def}}{=} \bigwedge_{\Gamma,\Gamma'\in 2^{Des*}:\Gamma\neq\Gamma'} \big(\mathsf{des}(\mathfrak{h}, \Gamma) \rightarrow \neg\mathsf{des}(\mathfrak{h}, \Gamma')\big),$$

$$\alpha_4 \stackrel{\text{def}}{=} \bigvee_{\Gamma\in 2^{Des*}} \mathsf{des}(\mathfrak{h}, \Gamma),$$

$$\alpha_5 \stackrel{\text{def}}{=} \bigwedge_{o\in Opt} \begin{pmatrix}\mathsf{ideal}(\mathfrak{h}, o) \leftrightarrow \bigvee_{\Gamma\in 2^{Des*}} \big(\mathsf{des}(\mathfrak{h}, \Gamma)\wedge\\ \bigwedge_{\gamma\in\Gamma} f_{comp}(o, \gamma)\big)\end{pmatrix},$$

$$\alpha_6 \stackrel{\text{def}}{=} \bigwedge_{o\in Opt} \begin{pmatrix}\mathsf{justif}(\mathfrak{h}, o) \leftrightarrow \bigvee_{\Gamma\in 2^{Des*}} \big(\mathsf{des}(\mathfrak{h}, \Gamma)\wedge\\ \bigwedge_{\gamma\in\Gamma} f^{\mathfrak{h}}_{comp}(o, \gamma)\big)\end{pmatrix}.$$

Formula $\alpha_1$ captures the fact that a sport cannot have two different values for a given variable. Formula $\alpha_2$ is its subjective version for agent $\mathfrak{h}$. Formulas $\alpha_3$ and $\alpha_4$ capture together the fact that agent $\mathfrak{h}$ has exactly one non-empty set of desires. According to formula $\alpha_5$, an option $o$ is ideal for agent $\mathfrak{h}$ if and only if it satisfies all agent $\mathfrak{h}$'s desires. Finally, according to formula $\alpha_6$, agent $\mathfrak{h}$ has a reasonable justification for choosing option $o$ if and only if she has all necessary information to conclude that option $o$ satisfies all her desires.

We also assume that agent $\mathfrak{m}$ has in its belief base a complete representation of Table 1, which specifies the variable assignments for all options:

$$\alpha_7^{o,x} \stackrel{\text{def}}{=} \mathsf{val}(o, x\mapsto v_{o,x}).$$

| | **env** | **loc** | **soc** | **cost** | **dan** | **intens** |
|---|---|---|---|---|---|---|
| sw | *water* | *mixed* | *single* | *med* | *low* | *high* |
| ru | *land* | *outdoor* | *single* | *low* | *med* | *high* |
| hr | *land* | *outdoor* | *single* | *high* | *high* | *low* |
| te | *land* | *mixed* | *mixed* | *high* | *med* | *med* |
| so | *land* | *mixed* | *team* | *med* | *med* | *med* |
| yo | *land* | *mixed* | *single* | *med* | *low* | *low* |
| di | *water* | *mixed* | *single* | *high* | *high* | *low* |
| sq | *land* | *indoor* | *mixed* | *high* | *med* | *med* |

Table 1: Variable assignments. For every option $o \in Opt$ and variable $x \in Var$, we denote by $v_{o,x}$ the corresponding entry in the table. For instance, we have $v_{\text{sw},\mathbf{env}} = water$.

In order to help agent $\mathfrak{h}$ to select an activity, agent $\mathfrak{m}$ also needs information about $\mathfrak{h}$'s set of actual desires. The latter is captured by the following formula:

$$\alpha_8 \stackrel{\text{def}}{=} \mathsf{des}(\mathfrak{h}, \Gamma_{\mathfrak{h}}), \text{ with}$$

$$\begin{aligned}\Gamma_{\mathfrak{h}} =&\{\mathbf{env}\mapsto land, \mathbf{intens}\mapsto med, \sim\mathbf{loc}\mapsto indoor,\\ &[\mathbf{cost}\mapsto high]\rightsquigarrow \mathbf{soc}\mapsto mixed\}.\end{aligned}$$

This means that, according to agent $\mathfrak{m}$, agent $\mathfrak{h}$ would like to practice a land activity, with medium intensity, which is not exclusively indoor, and which can be practiced both in single and team mode, if its cost is high.

Let us now turn to the cognitive planning problem. We suppose agent $\mathfrak{m}$'s set of operators $Op$ is:

$$\begin{aligned}Op =&\big\{inform\big(\mathfrak{m},\mathfrak{h},\mathsf{val}(o, a)\big) : o \in Opt \text{ and } a \in Assign\big\} \cup\\ &\big\{inform\big(\mathfrak{m},\mathfrak{h},\mathsf{ideal}(\mathfrak{h}, o)\big) : o \in Opt\big\}.\end{aligned}$$

In other words, agent $\mathfrak{m}$ can only inform agent $\mathfrak{h}$ about an option's value for a certain variable or about the ideality of an option for her. We suppose the following executability precondition for every $o \in Opt$ and $a \in Assign$:

$$\begin{aligned}\mathcal{P}\Big(inform\big(\mathfrak{m},\mathfrak{h},\mathsf{val}(o, a)\big)\Big) =& \Box_{\mathfrak{m}}\Big(\mathsf{val}(o, a) \wedge \bigwedge_{v\in Val_{\mathbf{dan}}}\\ &\big(\mathsf{val}(o, \mathbf{dan}\mapsto v) \rightarrow\\ &\triangle_{\mathfrak{h}}\mathsf{val}(o, \mathbf{dan}\mapsto v)\big)\Big)\\ &\text{if } a \notin Assign_{\mathbf{dan}},\\ \mathcal{P}\Big(inform\big(\mathfrak{m},\mathfrak{h},\mathsf{val}(o, a)\big)\Big) =& \Box_{\mathfrak{m}}\mathsf{val}(o, a)\\ &\text{if } a \in Assign_{\mathbf{dan}},\\ \mathcal{P}\Big(inform\big(\mathfrak{m},\mathfrak{h},\mathsf{ideal}(\mathfrak{h}, o)\big)\Big) =& \Box_{\mathfrak{m}}\big(\mathsf{ideal}(\mathfrak{h}, o) \wedge \mathsf{justif}(\mathfrak{h}, o)\big).\end{aligned}$$

According to the first definition, agent $\mathfrak{m}$ can inform agent $\mathfrak{h}$ about an option's value for a certain variable, if and only if this information is believed by $\mathfrak{m}$ and $\mathfrak{m}$ believes that $\mathfrak{h}$ has been already informed about the dangerousness level of the option. Indeed, we assume that, before being presented with an option's features, agent $\mathfrak{h}$ must be informed about its the dangerousness level and agent $\mathfrak{m}$ complies with this rule. The second definition simply stipulates that $\mathfrak{m}$ can inform $\mathfrak{h}$ about the dangerousness level of an option if and only if it believes what it says. Finally, according to the third definition, $\mathfrak{m}$ can inform $\mathfrak{h}$ about the ideality of an option only if it believes that $\mathfrak{h}$ has a reasonable justification for choosing it. Indeed, we assume $\mathfrak{m}$ will inform $\mathfrak{h}$ about the ideality of an option only after having explained why the option is ideal for her. The three definitions presuppose that agent $\mathfrak{m}$ cannot spread fake news (i.e., something that it does not implicitly believe).

We moreover suppose that, for agent $\mathfrak{h}$ to have a potential intention to choose option $o$, denoted by $\mathsf{potIntend}(\mathfrak{h}, o)$, she must have a justified belief that $o$ is an ideal option for her:[3]

$$\mathsf{potIntend}(\mathfrak{h}, o) \stackrel{\text{def}}{=} \triangle_{\mathfrak{h}}\mathsf{ideal}(\mathfrak{h}, o) \wedge \mathsf{justif}(\mathfrak{h}, o).$$

[3]Our account of potential intention is reminiscent of the JTB ('justified true belief') account to knowledge (Goldman 1979).

This abbreviation together with the abbreviation $\alpha_6$ given above relate intention with belief and desire, in line with existing theories of intention (Audi 1973; Davidson 1980).

It turns out that the sequence of speech acts $\epsilon_1, \epsilon_2, \epsilon_3, \epsilon_4, \epsilon_5, \epsilon_6$ with

$$\begin{aligned}
\epsilon_1 &\stackrel{\text{def}}{=} \mathit{inform}\big(\mathfrak{m},\mathfrak{h},\mathsf{val}(\mathrm{te}, \mathbf{dan} \mapsto \mathit{med})\big),\\
\epsilon_2 &\stackrel{\text{def}}{=} \mathit{inform}\big(\mathfrak{m},\mathfrak{h},\mathsf{val}(\mathrm{te}, \mathbf{env} \mapsto \mathit{land})\big),\\
\epsilon_3 &\stackrel{\text{def}}{=} \mathit{inform}\big(\mathfrak{m},\mathfrak{h},\mathsf{val}(\mathrm{te}, \mathbf{intens} \mapsto \mathit{med})\big),\\
\epsilon_4 &\stackrel{\text{def}}{=} \mathit{inform}\big(\mathfrak{m},\mathfrak{h},\mathsf{val}(\mathrm{te}, \mathbf{loc} \mapsto \mathit{mixed})\big),\\
\epsilon_5 &\stackrel{\text{def}}{=} \mathit{inform}\big(\mathfrak{m},\mathfrak{h},\mathsf{val}(\mathrm{te}, \mathbf{soc} \mapsto \mathit{mixed})\big),\\
\epsilon_6 &\stackrel{\text{def}}{=} \mathit{inform}\big(\mathfrak{m},\mathfrak{h},\mathsf{ideal}(\mathfrak{h}, \mathrm{te})\big).
\end{aligned}$$

provides a solution for the planning problem $\langle \Sigma, Op, \alpha_G \rangle$, where

$$\begin{aligned}
\Sigma = &\{\alpha_1, \alpha_2, \alpha_3, \alpha_4, \alpha_5, \alpha_6, \alpha_8\} \cup\\
&\{\alpha_7^{o,x} : o \in \mathit{Opt} \text{ and } x \in \mathit{Var}\},
\end{aligned}$$

$Op$ has the previous specifications and agent $\mathfrak{h}$'s persuasive goal $\alpha_G$ is defined as follows:

$$\alpha_G \stackrel{\text{def}}{=} \bigvee_{o \in Opt} \mathsf{potIntend}(\mathfrak{h}, o).$$

This means that, by performing the sequence of operators $\epsilon_1, \epsilon_2, \epsilon_3, \epsilon_4, \epsilon_5, \epsilon_6$, agent $\mathfrak{m}$ will induce agent $\mathfrak{h}$ to form a potential intention to choose an activity. In other words, agent $\mathfrak{m}$ will provide an effective recommendation to agent $\mathfrak{h}$.

We conclude this section with a general observation about the formulation of the planning problem for our example. Let $\langle \Sigma, Op, \alpha_G \rangle$ be the planning problem we want to solve. Let $\mathfrak{m}$'s set of operators $Op(o)$ for option $o \in Opt$ relative to $\langle \Sigma, Op, \alpha_G \rangle$ be defined as follows:

$$\begin{aligned}
Op_o^{\langle \Sigma, Op, \alpha_G \rangle} = &\big\{ \mathit{inform}\big(\mathfrak{m},\mathfrak{h},\mathsf{val}(o, a)\big) : \mathsf{val}(o, a) \in \Sigma \big\} \cup\\
&\big\{ \mathit{inform}\big(\mathfrak{m},\mathfrak{h},\mathsf{ideal}(\mathfrak{h}, o)\big) \big\}.
\end{aligned}$$

It is easy to verify that the planning problem $\langle \Sigma, Op, \alpha_G \rangle$ has a solution if and only if there exists $o \in Opt$ such that the planning problem $\langle \Sigma, Op_o^{\langle \Sigma, Op, \alpha_G \rangle}, \alpha_G \rangle$ has a solution. Therefore, in order to solve the planning problem $\langle \Sigma, Op, \alpha_G \rangle$, we simply need to linearly order the options in $Opt$ and solve the planning problems $\langle \Sigma, Op_o^{\langle \Sigma, Op, \alpha_G \rangle}, \alpha_G \rangle$ in sequence one after the other according to the ordering.

## Conclusion

We have presented a simple logic-based framework for cognitive planning which can be used to endow an artificial agent with the capability of influencing a human agent's beliefs and intentions. We have studied both complexity of satisfiability for the logic and complexity of the cognitive planning problem. Our approach relies on SAT, given the NP-completeness of the satisfiability problem for the epistemic language we consider. We believe that this aspect opens up interesting perspectives of application in HCI based on the exploitation of existing SAT techniques. We are currently implementing a cognitive planning algorithm using a SAT-solver as well as the HMI scenario we presented in the paper.

Directions of future work are manifold. The dynamic extension $\mathcal{L}^+_{\mathsf{Frag}}$ considers belief expansion only. We plan to extend our approach by belief revision operators. This will allow the artificial planning agent to revise its belief base, in the light of new information received from the human agent that contradicts its explicit beliefs. We also plan to enrich our analysis of speech acts by taking 'yes-no' questions into consideration. To this aim, we plan to extend the language $\mathcal{L}^+_{\mathsf{Frag}}$ by program constructions of propositional dynamic logic (PDL) including non-deterministic choice ($\cup$). Last but not least, we plan to combine our cognitive planning approach with inductive logic programming (ILP) (Muggleton and de Raedt 1994), in order to construct an agent's prior information $\Sigma$, as used in the formulation of the cognitive planning problem, through inductive methods. This will allow the persuader to predict the persuadee's beliefs, like in models of theory of mind based on neural networks (Rabinowitz et al. 2018).

## Acknowledgments

This work is supported by the ANR project CoPains ("Cognitive Planning in Persuasive Multimodal Communication"). Support from the ANR-3IA Artificial and Natural Intelligence Toulouse Institute is also acknowledged.

## References

Amgoud, L.; Maudet, N.; and Parsons, S. 2000. Modelling dialogues using argumentation. In *Proceedings of the Fourth International Conference on MultiAgent Systems*, 31–38. IEEE.

Aucher, G.; and Bolander, T. 2013. Undecidability in epistemic planning. In *Proceedings of the 23rd International Joint Conference on Artificial Intelligence (IJCAI 2013)*, 27–33. AAAI Press.

Audi, R. 1973. Intending. *The Journal of Philosophy* 70(13): 387–403.

Bench-Capon, T. J. M. 2003. Persuasion in practical argument using value-based argumentation frameworks. *Journal of Logic and Computation* 13(3): 429–448.

Bolander, T.; and Andersen, M. B. 2011. Epistemic planning for single- and multi-agent systems. *Journal of Applied Non-Classical Logics* 21(1): 9–34.

Bolander, T.; Holm Jensen, M.; and Schwarzentruber, F. 2015. Complexity Results in Epistemic Planning. In *Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence (IJCAI 2015)*, 2791–2797. AAAI Press.

Bonzon, E.; and Maudet, N. 2011. On the Outcomes of Multiparty Persuasion. In *Proceedings of the 8th International Conference on Argumentation in Multi-Agent Systems (ArgMAS 2011)*, 86–101. Springer-Verlag.

Budzyńska, K.; and Kacprzak, M. 2008. A Logic for Reasoning about Persuasion. *Fundamenta Informaticae* 85(1-4): 51–65.

Caridroit, T.; Lagniez, J.; Le Berre, D.; de Lima, T.; and Montmirail, V. 2017. A SAT-Based Approach for Solving the Modal Logic S5-Satisfiability Problem. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence (AAAI-17)*, 3864–3870. AAAI Press.

Cialdini, R. B. 2001. *Influence: science and practice*. Allyn & Bacon.

Cooper, M. C.; Herzig, A.; Maffre, F.; Maris, F.; and Régnier, P. 2016. A simple account of multi-agent epistemic planning. In *Proceedings of the 22nd European Conference on Artificial Intelligence (ECAI 2016)*, 193–201.

Da Costa Pereira, C.; Tettamanzi, A.; and Villata, S. 2011. Changing One's Mind: Erase or Rewind? Possibilistic Belief Revision with Fuzzy Argumentation Based on Trust. In *Proceedings of the Twenty-Second International Joint Conference on Artificial Intelligence (IJCAI 2011)*, 164–171. AAAI Press. ISBN 9781577355137.

Davidson, D. 1980. *Essays on Actions and Events*. Clarendon Press.

Fagin, R.; Halpern, J.; Moses, Y.; and Vardi, M. 1995. *Reasoning about Knowledge*. Cambridge: MIT Press.

Goldman, A. 1979. What is Justified Belief? In Pappas, G., ed., *Justification and Knowledge*, 1–25. D. Reidel.

Goldman, A. I. 2006. *Simulating Minds: The Philosophy, Psychology, and Neuroscience of Mindreading*. Oxford University Press.

Halpern, J. Y.; and Moses, Y. 1992. A guide to completeness and complexity for modal logics of knowledge and belief. *Artificial Intelligence* 54(3): 319–379.

Hunter, A. 2015. Modelling the Persuadee in Asymmetric Argumentation Dialogues for Persuasion. In *Proceedings of the 24th International Conference on Artificial Intelligence (IJCAI 2015)*, 3055–3061. AAAI Press. ISBN 9781577357384.

Kominis, F.; and Geffner, H. 2015. Beliefs in multiagent planning: from one agent to many. In Brafman, R. I.; Domshlak, C.; Haslum, P.; and Zilberstein, S., eds., *Proceedings of the 25th International Conference on Automated Planning and Scheduling (ICAPS 2015)*, 147–155. AAAI Press.

Ladner, R. E. 1977. The Computational Complexity of Provability in Systems of Modal Propositional Logic. *SIAM Journal of Computing* 6(3): 467–480.

Lê Cong, S.; Pinchinat, S.; and Schwarzentruber, F. 2018. Small Undecidable Problems in Epistemic Planning. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI 2018, July 13-19, 2018, Stockholm, Sweden*, 4780–4786. ijcai.org.

Lomuscio, A.; Qu, H.; and Raimondi, F. 2017. MCMAS: an open-source model checker for the verification of multi-agent systems. *International Journal on Software Tools for Technology Transfer* 19: 9–30.

Lorini, E. 2018. In Praise of Belief Bases: Doing Epistemic Logic Without Possible Worlds. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence (AAAI-18)*, 1915–1922. AAAI Press.

Lorini, E. 2020. Rethinking epistemic logic with belief bases. *Artificial Intelligence* 282.

Löwe, B.; Pacuit, E.; and Witzel, A. 2011. DEL planning and some tractable cases. In *Proceedings of the 3rd International International Workshop on Logic, Rationality and Interaction (LORI 2011)*, 179–192. Springer Berlin Heidelberg.

Muggleton, S.; and de Raedt, L. 1994. Inductive logic programming: theory and methods. *Journal of Logic Programming* 19-20: 629–679.

Muise, C.; Belle, V.; Felli, P.; McIlraith, S. A.; Miller, T.; Pearce, A. R.; and Sonenberg, L. 2015. Planning over multi-agent epistemic states: A classical planning approach. In *Proceedings of the 29th AAAI Conference on Artificial Intelligence (AAAI 2015)*, 3327–3334. AAAI Press.

Perloff, R. M. 2003. *The Dynamics of Persuasion: Communication and Attitudes in the 21st Century*. L. Erlbaum.

Prakken, H. 2006. Formal Systems for Persuasion Dialogue. *The Knowledge Engineering Review* 21(2): 163–188.

Proietti, C.; and Yuste-Ginel, A. 2019. Persuasive Argumentation and Epistemic Attitudes. In *Proceedings of the Second International Workshop on Dynamic Logic. New Trends and Applications (DALI 2019)*, volume 12005 of *LNCS*, 104–123. Springer-Verlag.

Rabinowitz, N. C.; Perbet, F.; Song, H. F.; Zhang, C.; Eslami, S. M. A.; and Botvinick, M. 2018. Machine Theory of Mind. In *Proceedings of the 35th International Conference on Machine Learning (ICML 2018)*, volume 80 of *Proceedings of Machine Learning Research*, 4215–4224. PMLR.

Rashotte, L. 2009. Social Influence. In Ritzer, G.; and Ryan, J. M., eds., *Concise Blackwell Encyclopedia of Sociology*. Blackwell.

Searle, J. 1969. *Speech acts: An essay in the philosophy of language*. Cambridge: Cambridge University Press.

Stalnaker, R. 2002. Common ground. *Linguistics and Philosophy* 25(5-6): 701–721.

Walton, D.; and Krabbe, E. 1995. *Commitment in Dialogue: Basic Concepts of Interpersonal Reasoning*. SUNY Series in Logic and Language. State University of New York Press.