On the logical foundations of moral agency

Emiliano Lorini

On the logical foundations of moral agency*

Emiliano Lorini
Université de Toulouse, IRIT-CNRS, France

Abstract

The aim of this work is to provide a logical analysis of moral agency. Although this concept has been extensively studied in social philosophy and in social sciences, it has been far less studied in the field of deontic logic and multi-agent systems (MASs). We discuss different aspects of moral agency such as the distinction between desires and moral values and the concept of moral agent. All these concepts are formalized in a variant of STIT logic with explicit actions.

1 Introduction

Although the concepts of morality and moral agency have been extensively studied in social philosophy and in social sciences, they have been far less studied in the areas of multi-agent systems (MASs) and normative multi-agent systems (NorMAS). Some works have been done on the extension of the BDI (Belief, Desire, Intention) model with normative concepts such as obligation [6, 10], but none of them have really focused on the integration of moral aspects into the architecture of a cognitive agent. Developing formal models of cognitive agents integrating a moral dimension is a promising research avenue for these two areas. Indeed, as shown by social scientists [9, 8], decisions of human agents are often affected by moral sentiments and moral concerns (*e.g.*, concerns for fairness or equity). Therefore, to take the presence of moral attitudes into account becomes extremely important when developing formal and computational models of social procedures to be applied to human societies, and when developing logical models of artificial agents which are expected to interact with human agents (*e.g.*, trading agents, recommender systems, tutoring agents, etc.). The aim of this paper is to propose a logical framework in which different aspects of moral agency can be formalized such as the distinction between desires and moral attitudes and the concept of moral agent.

The rest of the paper is organized as follows. Section 2 establishes the conceptual basis of the logical analysis of moral agency developed in the second part of the paper, whereas in Section 3 the logic DL-MA (*Dynamic Logic of Mental attitudes and joint Actions*) is presented. As we will show, DL-MA can be seen as a variant of STIT logic — the logic of "Seeing To It That" [4, 14] — with explicit actions. In the second part

*The proceedings version of the paper published by Springer-Verlage (LNCS series) contains a mistake in Definition 2. The mistake has been amended in this version of the paper.

of the paper (Section 4), the logic DL-MA is used to develop a logical analysis of the different aspects of moral agency discussed in Section 2.

2 Moral agency: conceptual basis

Some background and clarifications of the notion of moral agency are needed in order to ground the logical analysis presented in Section 4 on a solid conceptual basis.

Desires vs. moral values A model of moral agency should be able to explain the two different origins of an agent's motivations. Some motivations originate from the agents's desires. A desire can be conceived as an agent's attitude consisting in an anticipatory mental representation of a pleasant state of affairs (representational dimension of desires) that motivates the agent to achieve it (motivational dimension of desires). In this perspective, the motivational dimension of an agent's desire is realized through its representational dimension. For example when an agent desires to be at the Japanese restaurant eating sushi, he imagines himself eating sushi at the Japanese restaurant and this representation gives him plesaure. This pleasant representation motivates him to go to the Japanese restaurant in order to eat sushi. Agents are motivated not only by their desires but also by their moral values. Moral values, and more generally moral attitudes (ideals, standards, etc.), originate from an agent's capability of discerning what from his point of view is (morally) *good* from what is (morally) *bad*. If an agent has a certain ideal φ, then he thinks that the realization of the state of affairs φ ought to be promoted because φ is *good* in itself. A similar distinction has also been made by philosophers and by social scientists. For instance, Searle [19] has recently proposed a theory of how an agent may want something without desiring it and on the problem of reasons for acting based on moral values and independent from desires. In his theory of morality [12], Harsanyi distinguishes a person's *ethical preferences* from her *subjective preferences* and argues that a moral choice is a choice that is based on ethical preferences. [1]

The distinction between desires and moral values allows to identify two different kinds of moral dilemmas. The first kind of moral dilemma is the one which originates from the logical conflict between two moral values. The paradigmatic example is the situation of a soldier during a war. As a member of the army, the soldier feels obliged to kills his enemies, if this is the only way to defend his country. But, as a catholic, he thinks that human life should be respected. Therefore, he feels morally obliged not to kill other people. The other kind of moral dilemma is the one which originates from the logical conflict between desires and moral values. The paradigmatic example is that of Adam and Eve in the garden of Eden. They are tempted by the desire to eat the forbidden fruit and, at the same time, they have a moral obligation not to do it.

[1] An alternative to Harsanyi's dual theory of morality is Sen's theory [20]. According to Sen, moral attitudes are kind of meta-level attitudes. In particular, moral judgments are *rankings of preference rankings*. In this sense, differently from Harsanyi's theory, in Sen's theory subjective preferences and ethical preferences are not at the same level. A comparison between these two theories of morality goes beyond the objectives of the present work.

Moral agents An important distinction for a logical theory of morality is *self-regarding* agent versus *moral* agent. A purely *self-regarding* agent is an agent who acts in order to maximize the satisfaction of his own desires, while a purely *moral agent* is an agent who acts in order to maximize the fulfilment of his own moral values. In other words, if an agent is purely self-regarding, the utility of this act for him coincides with the personal good the agent will obtain by performing this action — where the agent's personal good coincides with the satisfaction of the agent's own desires —. If an agent is purely moral, the utility of this act for him coincides with the moral good the agent will promote by performing this action — where the agent's promotion of the moral good coincides with the accomplishment of his own moral values —. The notion of self-regarding agent should not be confused with the *rationality assumption* of classical decision and game theory: according to classical decision and game theory, individuals are rational in the sense that they maximize their utility. The notions of self-regarding agent and moral agent are not in contradiction with this assumption. We can safely say that a given agent acts to maximize his utility even even though he does not act to maximize the satisfaction of his own desires (as he also cares about the fulfilment of his own moral values). Therefore, the person's act is moral even though the person is still acting to maximize her utility. Of course, purely self-regarding agents and purely moral agents are just extremes cases. An agent is more or less moral depending on whether the utility of a given option for him is more or less affected by his moral values. More precisely, the higher is the influence of the agent's moral values in evaluating the utility of a given decision option, more moral the agent is. The extent to which an agent's utility is affected by his moral values can be called *degree of moral sensitivity*.[2]

3 Logical framework

In the following sections the logic DL-MA (*Dynamic Logic of Mental attitudes and joint Actions*) is presented. DL-MA is a modal logic which supports reasoning about actions and capabilities of agents and of coalitions of agents. DL-MA also allows to describe epistemic states of agents as well as their desires and moral values. We first present the syntax and the semantics of DL-MA (Sections 3.1, 3.2 and 3.3). A complete axiomatization of the logic is given in Section 3.4. As we will show in Section 3.5, DL-MA is tightly related to STIT logic, as it can be seen as a variant of STIT with explicit actions.

3.1 Syntax

Assume a countable set of atomic propositions $Atm = \{p, q, \ldots\}$, a finite set of agents $Agt = \{i_1, \ldots, i_{|Agt|}\}$, a finite set of atomic action types $Act = \{a, b, \ldots\}$, and a finite set of natural numbers $Num = \{x \in \mathbb{N} : 0 \leq x \leq \mathsf{max}\}$, with $\mathsf{max} \in \mathbb{N} \setminus \{0\}$. We assume that the set Act includes the (in)action skip, *i.e.*, the action of doing nothing.

We define $Prop$ to be the set of propositional formulas, that is, the set of all Boolean combinations of atomic propositions.

[2]For a similar idea in current economic models of morality see, *e.g.*, [3, 1].

For each agent $i \in Agt$, the set $Act_i = \{a_i : a \in Act\}$ identifies the set of agent i's actions. $2^{Agt*} = 2^{Agt} \setminus \{\emptyset\}$ is the set of non-empty sets of agents, also called *coalitions*. Elements of 2^{Agt*} are denoted by symbols $H, J, \ldots$ Let

$$JAct = \{\{a_{i_1}, \ldots, a_{i_{|Agt|}}\} : a_{i_k} \in Act_{i_k} \text{ for all } i_k \in Agt\}$$

be the set of all possible *joint actions* of all agents. One might think of $JAct$ as the set of all possible *strategy profiles* in the game theoretic sense. Just as in game theory we suppose that at a given time point every agent performs exactly one action, and that all actions of different agents occur in parallel. Elements of $JAct$ are denoted by symbols $\delta, \delta', \delta'', \ldots$ We let δ_i denote the element in δ corresponding to agent i. Given a coalition H and a joint action δ of all agents, $\delta_H = \{a_i : i \in H \text{ and } a_i \in \delta\}$ is coalition H's part in the joint action δ. Let $JAct_H = \{\delta_H : \delta \in JAct\}$ be the set of all possible joint actions of coalition H. Finally, let $JAct^*$ be the set of all (possibly infinite) sequences of joint actions. Elements of $JAct^*$ are denoted by symbols ϵ, ϵ', $\epsilon'', \ldots$ For every $\epsilon_1, \epsilon_2 \in JAct^*$, we write $\epsilon_1 \sqsubseteq \epsilon_2$ to mean that either $\epsilon_1 = \epsilon_2$ or ϵ_1 is an initial subsequence of ϵ_2, *i.e.*, there is $\epsilon_3 \in JAct^*$ such that $\epsilon_2 = \epsilon_1;\epsilon_3$. The empty sequence of joint actions is denoted by nil. Infinite sequences of joint actions are called *histories*. Let $Hist$ be the set of all histories. Elements of $Hist$ are denoted by symbols $h, h', \ldots$

The language of DL-MA is defined by the following grammar in Backus-Naur Form:

$$\begin{array}{lcl} \epsilon & ::= & \delta \mid \epsilon;\epsilon \\ \varphi & ::= & p \mid \mathsf{occ}(\epsilon) \mid \mathsf{des}_{i,\mathsf{h}} \mid \mathsf{idl}_{i,\mathsf{h}} \mid \neg\varphi \mid \varphi_1 \wedge \varphi_2 \mid [\![\delta]\!]\varphi \mid \Box\varphi \mid \mathsf{K}_i\varphi \end{array}$$

where p ranges over Atm, i ranges over Agt, δ ranges over $JAct$, ϵ ranges over $JAct^*$ and h ranges over Num. The other Boolean constructions $\top$, $\bot$, $\vee$, $\rightarrow$ and $\leftrightarrow$ are defined from p, $\neg$ and $\wedge$ in the standard way.

DL-MA has special atomic formulas of three different kinds. The atomic formulas $\mathsf{occ}(\epsilon)$ represent information about occurrences of joint action sequences. The formula $\mathsf{occ}(\epsilon)$ has to be read "the joint action sequence ϵ is going to occur".

The other atomic formulas $\mathsf{des}_{i,\mathsf{h}}$ and $\mathsf{idl}_{i,\mathsf{h}}$ are used to rank the histories that an agent can imagine at a given world according to their *desirability* degree and to their *ideality* degree for the agent. *Desirability* captures the quantitative dimension of desires (*i.e.*, how much a given history promotes the satisfaction of the agent's desires), whereas *ideality* captures the quantitative dimension of moral values (*i.e.*, how much a given history promotes the fulfilment of the agent's moral values). Formula $\mathsf{des}_{i,\mathsf{h}}$ has to be read "the current history has for agent i a degree of desirability equal to h" while formula $\mathsf{idl}_{i,\mathsf{h}}$ has to be read "the current history has for agent i a degree of ideality equal to h". The following two abbreviations express respectively that "the current world has for agent i a degree of desirability equal to or higher than h" and "the current world has for agent i a degree of ideality equal to or higher than h":

$$\mathsf{des}_{i,\geq\mathsf{h}} \stackrel{\mathtt{def}}{=} \bigvee_{\mathsf{h}\leq\mathsf{k}\leq\mathsf{max}} \mathsf{des}_{i,\mathsf{k}} \qquad\qquad \mathsf{idl}_{i,\geq\mathsf{h}} \stackrel{\mathtt{def}}{=} \bigvee_{\mathsf{h}\leq\mathsf{k}\leq\mathsf{max}} \mathsf{idl}_{i,\mathsf{k}}$$

The logic DL-MA has three kinds of modal operators: $[\![\delta]\!]$, $\Box$ and K_i. $\Box$ is the historic necessity operator of STIT logic. The formula $\Box\varphi$ has to be read "φ is true

in all histories passing through the current moment". We define $\Diamond$ to be the dual of $\Box$, *i.e.*, $\Diamond\varphi \stackrel{\texttt{def}}{=} \neg\Box\neg\varphi$ where $\Diamond\varphi$ has to be read "φ is true in at least one history passing through the current moment".

$[\![\delta]\!]$ is a dynamic operator describing the fact that if the joint action δ is performed then it will lead to a state in which a given state of affairs holds. In particular, $[\![\delta]\!]\varphi$ has to be read "if the joint δ is performed, then φ will be true after its execution". We define $\langle\!\langle\delta\rangle\!\rangle$ to be the dual of $[\![\delta]\!]$, *i.e.*, $\langle\!\langle\delta\rangle\!\rangle\varphi \stackrel{\texttt{def}}{=} \neg[\![\delta]\!]\neg\varphi$, where $\langle\!\langle\delta\rangle\!\rangle\varphi$ has to be read "the joint δ is performed and φ will be true after its execution".

Finally, K_i is a modal operator characterizing the concept of *ex ante* (or *choice-independent*) knowledge in the sense of Aumann & Dreze [2] (see also [18]). The formula $\mathsf{K}_i\varphi$ has to be read "agent i knows that φ is true independently from his current choice" or "agent i thinks that φ is true for any choice he could have made". The dual of the operator K_i is denoted by $\widehat{\mathsf{K}}_i$, *i.e.*, $\widehat{\mathsf{K}}_i\varphi \stackrel{\texttt{def}}{=} \neg\mathsf{K}_i\neg\varphi$. Aumann & Dreze distinguish *ex ante* knowledge from *interim* knowledge. *Ex ante* knowledge characterizes an agent's knowledge assuming that no decision has yet been made by him, whereas *interim* knowledge characterizes an agent's knowledge assuming that the agent has made his decision about which action to take, but might still be uncertain about the decisions of others. The concept of *interim* knowledge is expressed in DL-MA by the following operator K_i^*:

$$\mathsf{K}_i^*\varphi \stackrel{\texttt{def}}{=} \bigwedge_{a_i \in Act_i} (\mathsf{choose}(a_i) \rightarrow \mathsf{K}_i(\mathsf{choose}(a_i) \rightarrow \varphi))$$

where $\mathsf{K}_i^*\varphi$ has to be read "the agent i knows that φ is true given his current choice".

The following abbreviations will be useful to axiomatize DL-MA. For any coalition H, joint action sequence ϵ and joint action δ_H we define:

$$\mathsf{choose}(\epsilon,\delta_H) \stackrel{\texttt{def}}{=} \bigvee_{\delta \in JAct:\delta_H \subseteq \delta} \mathsf{occ}(\epsilon;\delta) \qquad \mathsf{able}(\epsilon,\delta_H) \stackrel{\texttt{def}}{=} \Diamond\mathsf{choose}(\epsilon,\delta_H)$$

$\mathsf{choose}(\epsilon,\delta_H)$ has to be read "at the end of the joint action sequence ϵ, the agents in H will choose the joint action δ_H", whereas $\mathsf{able}(\epsilon,\delta_H)$ has to be read "the agents in H will be able to perform the joint action δ_H, at the end of the joint action sequence ϵ". For notational convenience, when $\epsilon = nil$, we write $\mathsf{choose}(\delta_H)$ instead of $\mathsf{choose}(nil,\delta_H)$ and $\mathsf{able}(\delta_H)$ instead of $\mathsf{able}(nil,\delta_H)$. Besides, we simply write $\mathsf{choose}(\epsilon,a_i)$ instead of $\mathsf{choose}(\epsilon,\{a_i\})$, $\mathsf{able}(\epsilon,a_i)$ instead of $\mathsf{able}(\epsilon,\{a_i\})$, $\mathsf{choose}(a_i)$ instead of $\mathsf{choose}(\{a_i\})$ and $\mathsf{able}(a_i)$ instead of $\mathsf{able}(\{a_i\})$.

3.2 Action description

Similarly to Situation Calculus [17], in DL-MA actions are described in terms of their positive and negative effect preconditions. In particular, we introduce two functions:

$$\gamma^+ : Agt \times Act \times Atm \longrightarrow Prop$$
$$\gamma^- : Agt \times Act \times Atm \longrightarrow Prop$$

mapping agents, actions and atomic propositions to propositional formulas. The formula $\gamma^+(i,a,p)$ describes the *positive effect preconditions* of action a performed by

agent i with respect to p, whereas $\gamma^-(i,a,p)$ describes the *negative effect preconditions* of action a performed by agent i with respect to p. Formula $\gamma^+(i,a,p)$ represents the conditions under which agent i will make p *true* by performing action a, if no other agent interferes with i's action; while $\gamma^-(i,a,p)$ represents the conditions under which agent i will make p *false* by performing action a, if no other agent interferes with i's action. We assume that "making p true" means changing the truth value of p from false to true, whereas "making p false" means changing the truth value of p from true to false. We make the following *coherence assumption*:

(COH_γ) for every $i \in Agt$, $a \in Act$ and $p \in Atm$, $\gamma^+(i,a,p)$ and $\gamma^-(i,a,p)$ must be logically inconsistent.

COH_γ ensures that actions do not have contradictory effects. As to the (in)action skip of doing nothing, we assume that for every agent $i \in Agt$, $\gamma^+(i,\mathsf{skip},p) = \gamma^-(i,\mathsf{skip},p) = \bot$, *i.e.*, an agent cannot change the truth value of p by doing nothing.

3.3 Semantics

The semantics of DL-MA is a possible world semantics with accessibility relations associated with each modal operators, with functions for *desirability* and *ideality*, and with a function designating the history starting in a given world.

Definition 1 *A DL-MA model is a tuple* $M = \langle W, \mathcal{H}, \equiv, \{\mathcal{E}_i : i \in Agt\}, \kappa_{\mathsf{des}}, \kappa_{\mathsf{idl}}, \mathcal{V}\rangle$ *where:*

- W *is a set of states (or worlds),*
- $\mathcal{H}$ *is a total function* $\mathcal{H} : W \longrightarrow Hist$*,*
- $\equiv$ *and every* $\mathcal{E}_i$ *are equivalence relations between states in* W*,*
- $\kappa_{\mathsf{des}} : W \times Agt \longrightarrow Num$ *and* $\kappa_{\mathsf{idl}} : W \times Agt \longrightarrow Num$ *are total functions mapping worlds and agents to natural numbers in* Num*,*
- $\mathcal{V} : W \longrightarrow 2^{Atm}$ *is a valuation function.*

As usual $p \in \mathcal{V}(w)$ means that proposition p is true at world w.

For every world $w \in W$, $\mathcal{H}(w)$ identifies the history starting in w. For notational convenience, for all $\epsilon \in JAct^*$, $i \in Agt$ and $a_i \in Act_i$ we write $\epsilon;a_i \sqsubseteq \mathcal{H}(w)$ if there is $\delta \in JAct$ such that $a_i \in \delta$ and $\epsilon;\delta \sqsubseteq \mathcal{H}(w)$.

$\equiv$-equivalence classes are called *moments*. If w and v belong to the same moment (*i.e.*, $w \equiv v$), then the history starting in w (*i.e.*, $\mathcal{H}(w)$) and the history starting in v (*i.e.*, $\mathcal{H}(v)$) are said to be alternative histories (*viz.*, histories starting at the same moment). The concept of moment is the one used in STIT logic [4, 14] and, more generally, in the Ockhamist theory of time [21, 23].

The equivalence relations $\mathcal{E}_i$ are used to interpret the *ex ante* epistemic operators K_i. They can be viewed as functions from W to 2^W. Therefore, we can write $\mathcal{E}_i(w) = \{v \in W : w\mathcal{E}_i v\}$. The set $\mathcal{E}_i(w)$ is the agent i's *information state* at world w: the

set of worlds that at w agent i considers epistemically possible independently from his current choice or, more shortly, agent i's set of epistemic alternatives at w. As $\mathcal{E}_i$ is an equivalence relation, if $w\mathcal{E}_i v$ then agent i has the same information state at w and v.

The functions κ_{des} and κ_{idl} represent respectively desirability gradings and ideality gradings of the possible worlds and are used to interpret the atomic formulas $\mathsf{des}_{i,\mathsf{h}}$ and $\mathsf{idl}_{i,\mathsf{h}}$. $\kappa_{\mathsf{des}}(w,i) = \mathsf{h}$ means that according to the agent i the world w has a degree of desirability h, whereas $\kappa_{\mathsf{idl}}(w,i) = \mathsf{h}$ means that according to the agent i the world w has a degree of ideality h.

We impose the following constraints on DL-MA models. For all $w \in W$, $\delta \in JAct$, $\epsilon \in JAct^*$, $i \in Agt$ and $a_i \in Act_i$ we have:

(**C1**) if for all $i \in Agt$ there is u_i such that $w \equiv u_i$ and $\epsilon;\delta_i \sqsubseteq \mathcal{H}(u_i)$, then there is u such that $w \equiv u$ and $\epsilon;\delta \sqsubseteq \mathcal{H}(u)$;

(**C2**) if there is v such that $w \equiv v$ and $\epsilon;a_i \sqsubseteq \mathcal{H}(v)$ then, for all u such that $w\mathcal{E}_i u$, there is z such that $u \equiv z$ and $\epsilon;a_i \sqsubseteq \mathcal{H}(z)$;

(**C3**) if there is v such that $w \equiv v$ and $\epsilon;a_i \sqsubseteq \mathcal{H}(v)$, then there is u such that $w\mathcal{E}_i u$ and $\epsilon;a_i \sqsubseteq \mathcal{H}(u)$.

According to the Constraint **C1**, if every individual action in a joint action δ can be chosen at the end of the joint action sequence ϵ, then the individual actions in δ can be chosen simultaneously at the end of ϵ. The Constraint **C1** is a variant of the assumption of *independence of agents* of STIT logic. More intuitively, this means that agents can never be deprived of choices due to the choices made by other agents. The Constraint **C2** is a basic assumption about agents' knowledge over their abilities: if an agent i can perform the action a at the end of the joint action sequence ϵ, then he knows this. In other words, an agent has perfect knowledge about his ability to perform a given action. The Constraint **C3** characterizes the basic property of *ex ante* knowledge: if an agent i can perform the action a at the end of the joint action sequence ϵ, then there is a history that the agent considers possible independently from his current choice in which he performs the action a at the end of the joint action sequence ϵ. In other words, for every action that an agent is able to perform, there is a history that the agent considers possible in which he performs this action. The rules defining truth conditions of DL-MA formulas are the standard ones for Boolean constructions p, $\neg$ and $\wedge$ *plus*:

- $M, w \models \mathsf{occ}(\epsilon)$ iff $\epsilon \sqsubseteq \mathcal{H}(w)$
- $M, w \models \mathsf{des}_{i,\mathsf{h}}$ iff $\kappa_{\mathsf{des}}(w,i) = \mathsf{h}$
- $M, w \models \mathsf{idl}_{i,\mathsf{h}}$ iff $\kappa_{\mathsf{idl}}(w,i) = \mathsf{h}$
- $M, w \models [\![\delta]\!]\varphi$ iff if $M, w \models \mathsf{occ}(\delta)$ then $M^\delta, w \models \varphi$
- $M, w \models \Box\varphi$ iff $M, v \models \varphi$ for all v such that $w \equiv v$
- $M, w \models \mathsf{K}_i\varphi$ iff $M, v \models \varphi$ for all v such that $w\mathcal{E}_i v$

where model M^δ is defined according to Definition 2 below.

Definition 2 (Update via joint action) *Given a DL-MA model $M = \langle W, \mathcal{H}, \equiv, \{\mathcal{E}_i : i \in Agt\}, \kappa_{\mathsf{des}}, \kappa_{\mathsf{idl}}, \mathcal{V}\rangle$, the update of M by δ is defined to be $M^\delta = \langle W^\delta, \mathcal{H}^\delta, \equiv^\delta, \{\mathcal{E}_i^\delta : i \in Agt\}, \kappa_{\mathsf{des}}^\delta, \kappa_{\mathsf{idl}}^\delta, \mathcal{V}^\delta\rangle$ where for all $i \in Agt$ and $h \in Hist$:*

$$
\begin{array}{lcl}
W^\delta & = & \{w \in W : M, w \models \mathsf{occ}(\delta)\} \\
\mathcal{H}^\delta(w) & = & h \textit{ if } \mathcal{H}(w) = \delta'; h \textit{ for some } \delta' \in JAct \\
\equiv^\delta & = & \equiv \cap (W^\delta \times W^\delta) \\
\mathcal{E}_i^\delta & = & \mathcal{E}_i \cap (W^\delta \times W^\delta) \\
\kappa_{\mathsf{des}}^\delta(w, i) & = & \kappa_{\mathsf{des}}(w, i) \\
\kappa_{\mathsf{idl}}^\delta(w, i) & = & \kappa_{\mathsf{idl}}(w, i) \\
\mathcal{V}^\delta(w) & = & (\mathcal{V}(w) \setminus \{p : \exists a_i \in \delta \textit{ such that } M, w \models \gamma^-(i, a, p) \textit{ and} \\
 & & \nexists b_j \in \delta \textit{ such that } M, w \models \gamma^+(j, b, p)\}) \cup \\
 & & \{p : \exists a_i \in \delta \textit{ such that } M, w \models \gamma^+(i, a, p) \textit{ and} \\
 & & \nexists b_j \in \delta \textit{ such that } M, w \models \gamma^-(j, b, p)\}
\end{array}
$$

The performance of a joint action δ modifies the physical facts via the positive effect preconditions and the negative effect preconditions, defined in Section 3.2 (see the definition of $\mathcal{V}^\delta$). In particular, if there is an action in the joint action δ whose positive effect preconditions with respect to p hold and there is no other action in the joint action δ whose negative effect preconditions with respect to p hold, then p will be true after the occurrence of δ; if there is an action in the joint action δ whose negative effect preconditions with respect to p hold and there is no other action in the joint action δ whose positive effect preconditions with respect to p hold, then p will be false after the occurrence of δ. Besides, the occurrence of the joint action δ makes the current history advance one step forward (see the definition of $\mathcal{H}^\delta$). As to the equivalence relations $\equiv$ and $\mathcal{E}_i$ for historic necessity and *ex ante* knowledge, they are restricted to the set of worlds in which the joint action δ occurs (see the definitions of $\equiv^\delta$ and $\mathcal{E}_i^\delta$). Finally, the joint action δ does not modify the agents' desirability and ideality gradings over the histories (see the definitions of $\kappa_{\mathsf{des}}^\delta$, $\kappa_{\mathsf{idl}}^\delta$ and W^δ). As stated by the following proposition, the update via a joint action preserves the constraints on DL-MA-models.

Proposition 1 *If M is a DL-MA-model then M^δ is a DL-MA-model too.*

PROOF. As an example, we prove that the model M^δ satisfies the Constraint **C1** after assuming that M is a DL-MA model. Let us suppose that for all $i \in Agt$ there is u_i such that $w \equiv^\delta u_i$ and $\epsilon;\delta_i' \sqsubseteq \mathcal{H}^\delta(u_i)$. The latter implies that for all $i \in Agt$ there is u_i such that $w \equiv u_i$ and $\delta;\epsilon;\delta_i' \sqsubseteq \mathcal{H}(u_i)$. As M is a DL-MA model which satisfies the Constraint **C1**, it follows that there is u such that $w \equiv u$ and $\delta;\epsilon;\delta' \sqsubseteq \mathcal{H}(u)$. The latter implies that there is u such that $w \equiv^\delta u$ and $\epsilon;\delta' \sqsubseteq \mathcal{H}^\delta(u)$.

3.4 Axiomatization

Our logic DL-MA has so-called reduction axioms. These axioms allow to eliminate all the dynamic operators $[\![\delta]\!]$ from formulas. That elimination provides an axiomatics.

Proposition 2 *The following formulas are DL-MA valid for every $i \in Agt$, $a_i \in Act$, $\delta, \delta' \in JAct$ and $\epsilon, \epsilon' \in JAct^*$:*

$$\mathsf{occ}(\epsilon) \rightarrow \bigvee_{\delta \in JAct} \mathsf{occ}(\epsilon;\delta) \qquad \textbf{(OneJAct)}$$

$$\mathsf{occ}(\epsilon;\delta) \rightarrow \neg\mathsf{occ}(\epsilon;\delta') \textit{ if } \delta \neq \delta' \qquad \textbf{(UniqueJAct)}$$

$$\mathsf{occ}(\epsilon) \rightarrow \mathsf{occ}(\epsilon') \textit{ if } \epsilon' \sqsubseteq \epsilon \qquad \textbf{(SubSeqJAct)}$$

$$\bigvee_{\mathsf{h} \in Num} \mathsf{des}_{i,\mathsf{h}} \qquad \textbf{(ComplDes)}$$

$$\mathsf{des}_{i,\mathsf{h}} \rightarrow \neg\mathsf{des}_{i,\mathsf{k}} \textit{ if } \mathsf{h} \neq \mathsf{k} \qquad \textbf{(UniqueDes)}$$

$$\bigvee_{\mathsf{h} \in Num} \mathsf{idl}_{i,\mathsf{h}} \qquad \textbf{(ComplIdl)}$$

$$\mathsf{idl}_{i,\mathsf{h}} \rightarrow \neg\mathsf{idl}_{i,\mathsf{k}} \textit{ if } \mathsf{h} \neq \mathsf{k} \qquad \textbf{(UniqueIdl)}$$

$$(\bigwedge_{i \in Agt} \mathsf{choose}(\epsilon,\delta_i)) \rightarrow \mathsf{occ}(\epsilon;\delta) \qquad \textbf{(IndepAgt)}$$

$$\mathsf{able}(\epsilon,a_i) \rightarrow \mathsf{K}_i\mathsf{able}(\epsilon,a_i) \qquad \textbf{(KnowCan)}$$

$$\mathsf{able}(\epsilon,a_i) \rightarrow \widehat{\mathsf{K}}_i\mathsf{choose}(\epsilon,a_i) \qquad \textbf{(ExAnteKnow)}$$

Proposition 3 *The following equivalences are DL-MA valid for all $p \in Atm$, $i \in Agt$, $\delta \in JAct$, $\epsilon \in JAct^*$ and $\mathsf{h} \in Num$:*

$$\begin{array}{lcl}
[\![\delta]\!]\neg\varphi & \leftrightarrow & (\mathsf{occ}(\delta) \rightarrow \neg[\![\delta]\!]\varphi) \\
[\![\delta]\!](\varphi \wedge \psi) & \leftrightarrow & ([\![\delta]\!]\varphi \wedge [\![\delta]\!]\psi) \\
[\![\delta]\!]p & \leftrightarrow & (\mathsf{occ}(\delta) \rightarrow ((\bigvee_{a_i \in \delta} \gamma^+(i,a,p) \wedge \bigwedge_{b_j \in \delta} \neg\gamma^-(j,b,p)) \vee \\
 & & (p \wedge \bigwedge_{a_i \in \delta} \neg\gamma^-(i,a,p)) \vee (p \wedge \bigvee_{a_i \in \delta} \gamma^+(i,a,p)))) \\
[\![\delta]\!]\mathsf{occ}(\epsilon) & \leftrightarrow & (\mathsf{occ}(\delta) \rightarrow \mathsf{occ}(\delta;\epsilon)) \\
[\![\delta]\!]\mathsf{des}_{i,\mathsf{h}} & \leftrightarrow & (\mathsf{occ}(\delta) \rightarrow \mathsf{des}_{i,\mathsf{h}}) \\
[\![\delta]\!]\mathsf{idl}_{i,\mathsf{h}} & \leftrightarrow & (\mathsf{occ}(\delta) \rightarrow \mathsf{idl}_{i,\mathsf{h}}) \\
[\![\delta]\!]\Box\varphi & \leftrightarrow & (\mathsf{occ}(\delta) \rightarrow \Box(\mathsf{occ}(\delta) \rightarrow [\![\delta]\!]\varphi)) \\
[\![\delta]\!]\mathsf{K}_i\varphi & \leftrightarrow & (\mathsf{occ}(\delta) \rightarrow \mathsf{K}_i(\mathsf{occ}(\delta) \rightarrow [\![\delta]\!]\varphi))
\end{array}$$

Let L-MA be the fragment of the logic DL-MA without dynamic operators $[\![\delta]\!]$. that is, let the language of L-MA be the set of formulas defined by the following grammar:

$$\varphi \quad ::= \quad p \mid \mathsf{occ}(\epsilon) \mid \mathsf{des}_{i,\mathsf{h}} \mid \mathsf{idl}_{i,\mathsf{h}} \mid \neg\varphi \mid \varphi_1 \wedge \varphi_2 \mid \Box\varphi \mid \mathsf{K}_i\varphi$$

where p ranges over Atm, i ranges over Agt, ϵ ranges over $JAct^*$ and h ranges over Num. As the rule of replacement of equivalents preserves validity, the equivalences of Proposition 3 together with the rule of replacement of equivalents allow to reduce every DL-MA formula to an equivalent L-MA formula. Call red the mapping which iteratively applies the above equivalences from the left to the right, starting from one of the innermost modal operators. red pushes the dynamic operators inside the formula, and finally eliminates them when facing an atomic formula.

Proposition 4 *Let φ be a formula in the language of DL-MA. Then*

1. $red(\varphi)$ *has no dynamic operators* $[\![\delta]\!]$
2. $red(\varphi) \leftrightarrow \varphi$ *is* DL-MA *valid*
3. $red(\varphi)$ *is* DL-MA *valid iff* $red(\varphi)$ *is* L-MA *valid.*

PROOF.[Sketch] The first item is clear. The second item is proved using Proposition 3 (and the rule of replacement of equivalents). The last item follows from the second item and the fact that DL-MA is a conservative extension of L-MA.

Theorem 1 *The validities of* DL-MA *are completely axiomatized by*

- *all principles of classical propositional logic*
- *axiomatization of the normal modal logic S5 for the historic necessity operator* $\Box$
- *axiomatization of the normal modal logic S5 for each epistemic operator* K_i
- *the schemas of Proposition 2*
- *the reduction axioms of Proposition 3*
- *the rule of replacement of equivalents*

$$\textit{from} \quad \psi_1 \leftrightarrow \psi_2 \quad \textit{infer} \quad \varphi \leftrightarrow \varphi[\psi_1/\psi_2]$$

PROOF.[Sketch] To prove soundness is just a routine exercise. The completeness proof proceeds as follows. By standard canonical model argument, it is routine to show that the axioms and rules of inference of the normal modal logic S5 for the operator $\Box$ and for every epistemic operator K_i together with the principles in Proposition 2 and all principles of classical propositional logic provide a complete axiomatization for L-MA. Now, suppose φ is DL-MA valid. Then $red(\varphi)$ is valid in L-MA due to Proposition 4. By the completeness of L-MA, $red(\varphi)$ is also provable there. DL-MA being a conservative extension of L-MA, $red(\varphi)$ is provable in DL-MA, too. As the reduction axioms and the rule of replacement of equivalents are part of our axiomatics, the formula φ must also be provable in DL-MA.

We write $\vdash \varphi$ if φ is a DL-MA-theorem.

3.5 Relationships between DL-MA and STIT

As pointed out in the introduction, DL-MA can be seen as a variant of STIT with explicit actions: while in STIT an action is identified with the result brought about by a coalition (*i.e.*, in STIT one can only express that a given coalition H sees to it that φ), in DL-MA an action is identified both with the result brought about by the coalition and with the means used by the coalition to bring about the result (*i.e.*, in DL-MA one can also express that a given coalition H sees to it that φ by choosing the joint action

δ_H).[3] In DL-MA we can express different concepts of agency which have been studied in the context of STIT theory [4, 14].

For instance, the so-called 'Chellas' operator CStit_H of STIT can be defined in DL-MA as follows:

$$\mathsf{CStit}_H\varphi \stackrel{\mathtt{def}}{=} \bigvee_{\delta_H \in JAct_H} (\mathsf{choose}(\delta_H) \wedge \bigwedge_{\delta'_{Agt\setminus H} \in JAct_{Agt\setminus H}} \Box(\mathsf{occ}(\delta_H \cup \delta'_{Agt\setminus H}) \rightarrow \varphi))$$

This means that the coalition H sees to it that φ if and only if, the agents in H choose some joint action δ_H such that, no matter what the agents outside H choose, if the agents in H choose δ_H then φ will be true. The so-called 'deliberative' operator DStit_H can be defined in DL-MA as follows:

$$\mathsf{DStit}_H\varphi \stackrel{\mathtt{def}}{=} \mathsf{CStit}_H\varphi \wedge \neg\Box\varphi$$

DL-MA integrates a temporal dimension which allows to express that a given coalition H sees to it that φ is true in the next state of the system. In particular, in DL-MA one can simulate the so-called XSTIT operator XStit_H proposed by Broersen [5]:

$$\mathsf{XStit}_H\varphi \stackrel{\mathtt{def}}{=} \mathsf{CStit}_H\mathsf{X}\varphi$$

where X is the operator 'next' of linear temporal logic (LTL) which is defined as:

$$\mathsf{X}\varphi \stackrel{\mathtt{def}}{=} \bigvee_{\delta \in JAct} \langle\!\langle \delta \rangle\!\rangle \varphi$$

4 Moral agency: a logical formalization

In what follows the logic DL-MA is applied to the formalization of the different aspects of moral agency discussed in Section 2. Section 4.1 provides a logical formalization of desires and moral values. In Section 4.2 we define the concept of moral sensitivity as well as a concept of preference based on desires and moral values.

4.1 Desires and moral values

For every agent $i \in Agt$, we define two types of dyadic operators, one for desirability and the other for ideality:

$$\psi \leq_i^{\mathsf{des}} \varphi \stackrel{\mathtt{def}}{=} \bigwedge_{\mathsf{h} \in Num} (\widehat{\mathsf{K}}_i(\mathsf{des}_{i,\geq \mathsf{h}} \wedge \psi) \rightarrow \widehat{\mathsf{K}}_i(\mathsf{des}_{i,\geq \mathsf{h}} \wedge \varphi))$$

$$\psi \leq_i^{\mathsf{idl}} \varphi \stackrel{\mathtt{def}}{=} \bigwedge_{\mathsf{h} \in Num} (\widehat{\mathsf{K}}_i(\mathsf{idl}_{i,\geq \mathsf{h}} \wedge \psi) \rightarrow \widehat{\mathsf{K}}_i(\mathsf{idl}_{i,\geq \mathsf{h}} \wedge \varphi))$$

$\psi \leq_i^{\mathsf{des}} \varphi$ has to be read "φ is for agent i at least as desirable as ψ" and $\psi \leq_i^{\mathsf{idl}} \varphi$ has to be read "φ is for agent i at least as ideal as ψ". Corresponding strict orderings are

[3] See also [13] for a variant of STIT with explicit actions.

defined in the expected way as follows: $\psi <_i^{\mathsf{des}} \varphi \stackrel{\mathtt{def}}{=} (\psi \leq_i^{\mathsf{des}} \varphi) \wedge \neg(\varphi \leq_i^{\mathsf{des}} \psi)$ and $\psi <_i^{\mathsf{idl}} \varphi \stackrel{\mathtt{def}}{=} (\psi \leq_i^{\mathsf{idl}} \varphi) \wedge \neg(\varphi \leq_i^{\mathsf{idl}} \psi)$.

As the following proposition highlights the comparative statements $\psi \leq_i^{\mathsf{des}} \varphi$ and $\psi \leq_i^{\mathsf{idl}} \varphi$ might also be read as "for every epistemically possible ψ-state there is an epistemically possible φ-state which is at least as desirable" and "for every epistemically possible ψ-state there is an epistemically possible φ-state which is at least as ideal".

Proposition 5 *For every $i \in Agt$ we have:*

- $M, w \models \psi \leq_i^{\mathsf{des}} \varphi$ *if and only if for all* $v \in \mathcal{E}_i(w)$*, if* $M, v \models \psi$ *then there is* $u \in \mathcal{E}_i(w)$ *such that* $\kappa_{\mathsf{des}}(v, i) \leq \kappa_{\mathsf{des}}(u, i)$ *and* $M, u \models \varphi$*,*
- $M, w \models \psi \leq_i^{\mathsf{idl}} \varphi$ *if and only if for all* $v \in \mathcal{E}_i(w)$*, if* $M, v \models \psi$ *then there is* $u \in \mathcal{E}_i(w)$ *such that* $\kappa_{\mathsf{idl}}(v, i) \leq \kappa_{\mathsf{idl}}(u, i)$ *and* $M, u \models \varphi$*.*

As pointed out by [16], there is no consensus in the literature on how preferential statements between formulas should be defined. In [22] other kinds of preference comparisons between formulas (*i.e.*, between sets of states) are defined. For instance, the previous $\forall\exists$-reading of preference statements should be distinguished from a $\forall\forall$-reading ("for every ψ-state and for every φ-state the φ-state is at least as desirable/ideal as the ψ-state") and a $\exists\exists$-reading ("there are a ψ-state and a φ-state such that the φ-state is at least as desirable/ideal as the ψ-state"). A logical analysis of such alternative readings of preference statements is postponed to future work.

Proposition 6 captures some central properties of the operators $\leq_i^{\mathsf{des}}$ and $\leq_i^{\mathsf{idl}}$.

Proposition 6 *Let $x \in \{\mathsf{des}, \mathsf{idl}\}$. Then:*

$$\vdash \psi \leq_i^x \psi \tag{6a}$$

$$\vdash ((\varphi_1 \leq_i^x \varphi_2) \wedge (\varphi_2 \leq_i^x \varphi_3)) \rightarrow (\varphi_1 \leq_i^x \varphi_3) \tag{6b}$$

$$\vdash (\varphi_1 \leq_i^x \varphi_2) \vee (\varphi_2 \leq_i^x \varphi_1) \tag{6c}$$

$$\vdash \bot \leq_i^x \top \tag{6d}$$

$$\textit{if } \vdash \varphi \rightarrow (\psi_1 \vee ... \vee \psi_s) \textit{ then } \vdash (\varphi \leq_i^x \psi_1) \vee ... \vee (\varphi \leq_i^x \psi_s) \tag{6e}$$

PROOF. We prove theorem (6c) with $x = \mathsf{des}$ as an example. $\neg(\varphi_1 \leq_i^x \varphi_2)$ is equivalent to $\bigvee_{\mathsf{h} \in Num}(\widehat{\mathsf{K}}_i(\mathsf{des}_{i,\geq \mathsf{h}} \wedge \varphi_1) \wedge \mathsf{K}_i(\mathsf{des}_{i,\geq \mathsf{h}} \rightarrow \neg\varphi_2))$ which in turn implies $\bigwedge_{\mathsf{k} \in Num}(\widehat{\mathsf{K}}_i(\mathsf{des}_{i,\geq \mathsf{k}} \wedge \varphi_2) \rightarrow \widehat{\mathsf{K}}_i(\mathsf{des}_{i,\geq \mathsf{k}} \wedge \varphi_1))$. The latter is equivalent to $(\varphi_2 \leq_i^x \varphi_1)$. We have proved that $\neg(\varphi_1 \leq_i^x \varphi_2) \rightarrow (\varphi_2 \leq_i^x \varphi_1)$ which is equivalent to $(\varphi_1 \leq_i^x \varphi_2) \vee (\varphi_2 \leq_i^x \varphi_1)$.

Theorems (6a)-(6c) highlight that $\leq_i^{\mathsf{des}}$ and $\leq_i^{\mathsf{idl}}$ are total preorders. Theorems (6b)-(6d) are the three fundamental principles of Lewis's conditional logic [15].

4.2 Moral sensitivity and preference based on desires and moral values

We extend the logic DL-MA with special constructions of the form $\mathsf{moralSensit}(i,\mathsf{m})$ which has to be read "agent i's degree of moral sensitivity is equal to m" with $\mathsf{m} \in$

Num. We call DL-MA$^+$ the resulting logic. A DL-MA$^+$ model is a tuple $\langle M, \mathcal{S}\rangle$ where M is a DL-MA model and $\mathcal{S}$ is a total function:

$$\mathcal{S} : W \times Agt \longrightarrow Num$$

capturing the moral sensitivity of an agent at a given state. We assume that an agent is aware of his current degree of moral sensitivity, that is, for every $i \in Agt$ and $w \in W$ we suppose that:

(C4) if $\mathcal{S}(w, i) = \mathsf{m}$ then, for all v such that $w\mathcal{E}_i v$, $\mathcal{S}(v, i) = \mathsf{m}$.

Constructions $\mathsf{moralSensit}(i,\mathsf{m})$ are interpreted by means of the function $\mathcal{S}$ as follows:

$$M, w \models \mathsf{moralSensit}(i,\mathsf{m}) \text{ if and only if } \mathcal{S}(w, i) = \mathsf{m}$$

It is straightforward to adapt the proof of Theorem 1 in order to prove that the logic DL-MA$^+$ is completely axiomatized by the axioms and rules of inference of the logic DL-MA plus the following axiom schemas:

$$\bigvee_{\mathsf{m}\in Num} \mathsf{moralSensit}(i,\mathsf{m}) \qquad \textbf{(ComplMoral)}$$

$$\mathsf{moralSensit}(i,\mathsf{m}) \rightarrow \neg\mathsf{moralSensit}(i,\mathsf{l}) \text{ if } \mathsf{m} \neq \mathsf{l} \qquad \textbf{(UniqueMoral)}$$

$$\mathsf{moralSensit}(i,\mathsf{m}) \rightarrow \mathsf{K}_i\mathsf{moralSensit}(i,\mathsf{m}) \qquad \textbf{(KnowMoral)}$$

We use the degree of moral sensitivity as a parameter for calculating the utility of a given history for an agent.

Definition 3 (Utility) *Given a DL-MA$^+$ model $M = \langle W, \mathcal{H}, \equiv, \{\mathcal{E}_i : i \in Agt\}, \kappa_{\mathsf{des}}, \kappa_{\mathsf{idl}}, \mathcal{V}, \mathcal{S}\rangle$, the utility for agent i of the history starting in the world w, denoted by $\kappa_{\mathsf{util}}(w, i)$, is defined as follows:*

$$\kappa_{\mathsf{util}}(w, i) = \mathcal{S}(w, i) \times \kappa_{\mathsf{idl}}(w, i) + (\mathsf{max} - \mathcal{S}(w, i)) \times \kappa_{\mathsf{des}}(w, i)$$

Moreover, we define

$$UScale = \{\mathsf{x} : \exists \mathsf{h}_1, \mathsf{h}_2, \mathsf{h}_3 \in Num \textit{ such that } \mathsf{x} = \mathsf{h}_1 \times \mathsf{h}_2 + (\mathsf{max} - \mathsf{h}_1) \times \mathsf{h}_3\}$$

to be the agents' utility scale.[4]

According to Definition 3, the utility of a history is a function of both the degree of desirability and the degree of the ideality of the history. Degree of moral sensitivity captures the extent to which the utility of a given history is affected by moral values: the higher is the agent's moral sensitivity and the higher is the influence of the degree of ideality in determining the utility of the history; the lower is the agent's moral sensitivity and the higher is the influence of the degree of desirability in determining the utility of the history.

The next step in the analysis is to define the concept of preferred history, as a history that an agent envisages and that has a maximal degree of utility for him. More formally:

Definition 4 (Preferred histories) *Given a DL-MA$^+$ model $M = \langle W, \mathcal{H}, \equiv, \{\mathcal{E}_i : i \in Agt\}, \kappa_{\mathsf{des}}, \kappa_{\mathsf{idl}}, \mathcal{V}, \mathcal{S}\rangle$, agent i's set of preferred histories at w, denoted by $\mathcal{P}_i(w)$, is defined as follows:*

$$\mathcal{P}_i(w) = \operatorname*{argmax}_{v \in W : w\mathcal{E}_i v} \kappa_{\mathsf{util}}(v, i)$$

[4]Note that $UScale$ is finite because Num is finite.

We use the notion of preferred history to define the concept of preference. We say that an agent prefers φ to be true if and only if, φ is true in all histories that he prefers. More generally, we assume that each agent envisages a given set of possible histories and he evaluates which are the best ones among them. This leads to a *realistic* notion of preference in the sense that an agent's set of preferred histories is a subset of the set of histories that the agent envisages.

Definition 5 (Preference, Pref_i) *At world w agent i prefers φ to be true,* i.e.*, $M, w \models \mathsf{Pref}_i\varphi$, if and only if $M, v \models \varphi$ for all $v \in \mathcal{P}_i(w)$.*

As the following proposition highlights, the concept of preference semantically defined in Definition 5 is syntactically expressible in the logic $\mathsf{DL\text{-}MA}^+$.

Proposition 7 *For all $i \in Agt$ we have $M, w \models \mathsf{Pref}_i\varphi$ if and only if*

$$M, w \models \bigvee_{\mathsf{x} \in UScale}(\widehat{\mathsf{K}}_i\mathsf{util}_{i,\mathsf{x}} \wedge \bigwedge_{\mathsf{y} \in UScale:\mathsf{y}>\mathsf{x}} \mathsf{K}_i\neg\mathsf{util}_{i,\mathsf{y}} \wedge \mathsf{K}_i(\mathsf{util}_{i,\mathsf{x}} \rightarrow \varphi))$$

where

$$\mathsf{util}_{i,\mathsf{x}} \stackrel{\mathtt{def}}{=} \bigvee_{\mathsf{k},\mathsf{l},\mathsf{m} \in Num:\mathsf{x}=\mathsf{m}\times\mathsf{l}+(\mathsf{max}-\mathsf{m})\times\mathsf{k}}(\mathsf{moralSensit}(i,\mathsf{m}) \wedge \mathsf{des}_{i,\mathsf{k}} \wedge \mathsf{idl}_{i,\mathsf{l}})$$

Our operator of preference corresponds to Cohen & Levesque's goal operator [7]. However, differently from $\mathsf{DL\text{-}MA}$, their logic does not explain the relationships between goals and desires, and between goals and moral values.

Proposition 8 captures some basic properties of our preference operator.

Proposition 8 *For every $i \in Agt$ we have:*

$$\vdash (\mathsf{Pref}_i\varphi \wedge \mathsf{Pref}_i\psi) \rightarrow \mathsf{Pref}_i(\varphi \wedge \psi) \tag{8a}$$

$$\vdash \neg(\mathsf{Pref}_i\varphi \wedge \mathsf{Pref}_i\neg\varphi) \tag{8b}$$

$$\text{if } \vdash \varphi \text{ then } \vdash \mathsf{Pref}_i\varphi \tag{8c}$$

$$\vdash \mathsf{Pref}_i\varphi \rightarrow \mathsf{K}_i\mathsf{Pref}_i\varphi \tag{8d}$$

$$\vdash \neg\mathsf{Pref}_i\varphi \rightarrow \mathsf{K}_i\neg\mathsf{Pref}_i\varphi \tag{8e}$$

$$\vdash \mathsf{K}_i\varphi \rightarrow \mathsf{Pref}_i\varphi \tag{8f}$$

$$\vdash ((\neg\varphi <_i^{\mathsf{des}} \varphi) \wedge \mathsf{moralSensit}(i,0)) \rightarrow \mathsf{Pref}_i\varphi \tag{8g}$$

$$\vdash ((\neg\varphi <_i^{\mathsf{idl}} \varphi) \wedge \mathsf{moralSensit}(i,\mathsf{max})) \rightarrow \mathsf{Pref}_i\varphi \tag{8h}$$

Theorems (8a) and (8b) together with the rule of inference (8c) highlights that Pref_i is a KD normal modal operator. Theorems (8d) and (8e) are properties of positive and negative introspection for preferences. Theorem (8f) is the syntactic counterpart of the *realism* principle for preferences discussed above. A similar property is satisfied by Cohen & Levesque's goal operator. Finally, theorems (8g) and (8h) highlight the logical relationship between desires and preferences, and moral values and preferences. According to (8g), if an agent i has a minimal degree of moral sensitivity and considers φ strictly more desirable than $\neg\varphi$, then i prefers φ to be true. According to (8h), if an agent i has a maximal degree of moral sensitivity and considers φ strictly more ideal than $\neg\varphi$, then i prefers φ to be true.

The last aspect of moral agency we consider is the relationship between preferences and choices. To this aim we need to introduce a notion of preference-based rationality.

We say that a given agent i is rational, denoted by Rat_i, if and only if for every action a, if i prefers to do a and knows that is able to do it, then he decides to do a; and if he prefers not to do a, then he does not decide to do a. In other words, an agent is rational if his decision to perform a given action is completely determined by his own preferences:

$$\mathsf{Rat}_i \stackrel{\mathtt{def}}{=} \bigwedge_{a \in Act} (((\mathsf{Pref}_i \mathsf{choose}(a_i) \wedge \mathsf{K}_i \mathsf{able}(a_i)) \rightarrow \mathsf{choose}(a_i)) \wedge (\mathsf{Pref}_i \neg \mathsf{choose}(a_i) \rightarrow \neg \mathsf{choose}(a_i)))$$

The following two theorems, which follow from theorems (8g) and (8h) in Proposition 8, explain how desires and moral values motivate an agent to perform a given action:

$\vdash ((\neg \mathsf{choose}(a_i) <_i^{\mathsf{des}} \mathsf{choose}(a_i)) \wedge \mathsf{moralSensit}(i,0) \wedge \mathsf{Rat}_i \wedge \mathsf{K}_i \mathsf{able}(a_i)) \rightarrow \mathsf{choose}(a_i)$

$\vdash ((\neg \mathsf{choose}(a_i) <_i^{\mathsf{idl}} \mathsf{choose}(a_i)) \wedge \mathsf{moralSensit}(i,\mathsf{max}) \wedge \mathsf{Rat}_i \wedge \mathsf{K}_i \mathsf{able}(a_i)) \rightarrow \mathsf{choose}(a_i)$

According to the preceding two theorems, if agent i is rational, has a minimal/maximal degree of moral sensitivity, considers the situation in which he performs action a strictly more desirable/ideal than the situation in which he does not perform a and knows that he is able to do a, then i decides to do a.

5 Conclusion

We have devised a logic which supports reasoning about actions and capabilities of agents and coalitions, epistemic states of agents as well as their desires and moral values. We have used it to provide a formal analysis of the concept of moral agency.

Directions of future work are manifold. For instance, there are important aspects of moral agency that have not been addressed in this work and that we intend to study in the future. One of them is the concept of moral emotion [11]. Moral emotions such as guilt, moral pride and reproach are emotions which are based either on the fulfillment or on the violation of an agent's moral values by the agent himself or by another agent. Another issue we plan to investigate, and which has been briefly mentioned in Section 2, is the relationships between an agent's moral values and external norms (*e.g.*, obligations, prohibitions, etc.). As to the logical part, we have provided a complete axiomatization of the logic DL-MA. Future work will be devoted to prove its decidability.

6 Acknowledgements

This research has been supported by the French ANR project EmoTES "Emotions in strategic interaction: theory, experiments, logical and computational studies", contract No. 11-EMCO-004-01.

References

[1] I. Alger and J. W. Weibull. Homo moralis: preference evolution under incomplete information and assortative matching. Technical report, Toulouse School of Economics (TSE), 2012.

[2] R.J. Aumann and J.H. Dreze. Rational expectations in games. *American Economic Review*, 98(1):72–86, 2008.

[3] P. Battigalli and M. Dufwenberg. Guilt in games. *The American Economic Review*, 97(2):170–176, 2007.

[4] N. Belnap, M. Perloff, and M. Xu. *Facing the future: agents and choices in our indeterminist world*. Oxford University Press, New York, 2001.

[5] J. Broersen. A logical analysis of the interaction between 'obligation-to-do' and 'knowingly doing'. In *Proc. of the Ninth International Conference on Deontic Logic in Computer Science (DEON'08)*, volume 5076 of *LNCS*, pages 140–154. Springer-Verlag, 2008.

[6] J. Broersen, M. Dastani, J. Hulstijn, and L. van der Torre. Goal generation in the BOID architecture. *Cognitive Science Quarterly*, 2(3-4):428–447, 2002.

[7] P. R. Cohen and H. J. Levesque. Intention is choice with commitment. *Artificial Intelligence*, 42:213–261, 1990.

[8] E. Fehr and K. M. Schmidt. Theories of fairness and reciprocity: Evidence and economic applications. In *Advances in Economics and Econometrics*. Cambridge University Press, 2003.

[9] H. Gintis, S. Bowles, R. Boyd, and E. Fehr, editors. *Moral sentiments and material interests*. MIT Press, Cambridge, 2005.

[10] G. Governatori and A. Rotolo. BIO logical agents: Norms, beliefs, intentions in defeasible logic. *Journal of Autonomous Agents and Multi Agent Systems*, 17(1):36–69, 2008.

[11] J. Haidt. The moral emotions. In R. J. Davidson, K. R. Scherer, and H. H. Goldsmith, editors, *Handbook of Affective Sciences*, pages 852—870. Oxford University Press, 2003.

[12] J. Harsanyi. Morality and the theory of rational behaviour. In A. K. Sen and B. Williams, editors, *Utilitarianism and beyond*. Cambridge University Press, Cambridge, 1982.

[13] A. Herzig and E. Lorini. A dynamic logic of agency I: STIT, abilities and powers. *Journal of Logic, Language and Information*, 19(1):89–121, 2010.

[14] J. F. Horty. *Agency and Deontic Logic*. Oxford University Press, Oxford, 2001.

[15] D. Lewis. *Counterfactuals*. Harvard University Press, 1973.

[16] F. Liu. *Changing for the better: Preference dynamics and agent diversity*. PhD thesis, University of Amsterdam, The Netherlands, 2008.

[17] R. Reiter. *Knowledge in action: logical foundations for specifying and implementing dynamical systems*. MIT Press, Cambridge, 2001.

[18] O. Roy. Epistemic logic and the foundations of decision and game theory. *Journal of the Indian Council of Philosophical Research*, 27(2):283–314, 2010.

[19] J. Searle. *Rationality in Action*. MIT Press, Cambridge, 2001.

[20] A. K. Sen. Rational fools: a critique of the behavioral foundations of economic theory. *Philosophy and public affairs*, 6:317–344, 1977.

[21] R. Thomason. Combinations of tense and modality. In D. Gabbay and F. Guenthner, editors, *Handbook of Philosophical Logic*. Reidel, Dordrecht, 1984.

[22] J. van Benthem, P. Girard, and O. Roy. Everything else being equal: a modal logic for *ceteris paribus* preferences. *Journal of Philosophical Logic*, 38(1):83–125, 2009.

[23] A. Zanardo. Branching-time logic with quantification over branches: The point of view of modal logic. *Journal of Symbolic Logic*, 61(1):143–166, 1996.