Logique & Analyse **230** (2015), 177-218

A LOGIC FOR REASONING ABOUT MORAL AGENTS

Emiliano Lorini

Abstract

The aim of this work is to provide a logical analysis of moral agency. Although this concept has been extensively studied in moral philosophy and in economics, it has been far less studied in the areas of logics of agents and multi-agent systems. We discuss different aspects of moral agency such as the distinction between desires and moral values and the concept of moral choice. All these concepts are formalized in a logic of actions and agents' mental attitudes including knowledge, desires, moral values and preferences.

1. Introduction

Since the seminal work of Cohen & Levesque [11] aimed at implementing Bratman's philosophical theory of intention [7], many formal logics for reasoning about mental attitudes of agents such as beliefs, desires, goals and intentions have been developed. Among them we should mention the logics developed by [23, 30, 32, 38, 41, 40, 25]. These logics, commonly referred to as BDI (Belief, Desire, Intention) logics, provide the formal specification for the design and implementation of artificial cognitive agents that are capable of forming beliefs through sense perception and reasoning, making decisions on the basis of their beliefs and goals, and performing those actions that they have chosen through deliberation. Some work has been done on the extension of BDI logics with normative concepts such as obligation [8, 17], but none of it has really focused on the integration of moral aspects into the architecture of a cognitive agent. The aim of this paper is to provide such a kind of integration by building a connection between the contemporary debate on morality in philosophy and in economics, and the area of logical modeling of agents and multi-agent systems (MASs).

We propose a logical framework in which different aspects of morality can be formalized such as the distinction between desires and moral attitudes and the concept of moral choice. Specifically, our logical account of moral agency clarifies the role of morality both at the level of preference formation and at the level of choice, that is, it explains:

doi: 10.2143/LEA.230.0.3141808.

- how an agent's preference over alternative courses of action is affected both by the agent's desires and by motivations based on ethical and moral issues and,
- how a (rational) agent's choice over alternative courses of action is determined by his preferences.

We believe that integrating a moral dimension into logical models of cognitive agents is a promising research avenue. Indeed, as shown by social scientists [16, 15], 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 interesting when designing artificial agents which are expected to understand and to simulate the role of morality in human reasoning and human decision-making.

The rest of the paper is organized as follows. Section 2 establishes the conceptual basis of the logical analysis of morality developed in the second part of the paper. We address the distinction between desires and moral attitudes as well as the concept of moral choice, as it has been discussed in the contemporary debate in moral philosophy and in economics. In Section 3 the syntax and the semantics of the logic LAMA (*Logic of Actions and Mental Attitudes*) is presented. A complete axiomatization for this logic is given. In the second part of the paper (Section 4), the logic LAMA is used to develop a logical analysis of the different aspects of moral agency discussed in Section 2. Proofs are given in a technical annex at the end of the paper.

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' desires. Following the Humean conception, a desire can be viewed 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 pleasure. 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.[1] A similar distinction has also been made by philosophers and by social scientists. For instance, Searle [35] 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 [22, 21], Harsanyi distinguishes a person's *ethical preferences* from his *subjective preferences* and argues that a moral choice is a choice that is based on ethical preferences.

The distinction between desires and moral values allows us to identify two different kinds of moral dilemmas. The first kind of moral dilemma is the one which originates from a 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, thereby feeling morally obliged not to kill other people. The other kind of moral dilemma is the one which originates from a 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.[2]

Dual view of moral choice. Existing theories of moral choice have concentrated on the way decisions of rational agents are influenced by their moral attitudes and motivations. One of the most prominent approaches to moral choice is the so-called dual view proposed among the others by the economist John Harsanyi [21] and by the sociologist Howard Margolis [28], according to which desires and moral attitudes of an agent are two different

[1] There are different ways to explain the origin of moral values. Social scientists (*e.g.*, [6]) have defended the idea that there exist innate moral principles in humans such as fairness which are the product of biological evolution. Other moral values are the product of the internalization of some external norm. A possible explanation is based on the hypothesis that moral judgments are true or false only in relation to and with reference to one or another agreement between people forming a group or a community. More precisely, an agent's ideals are simply norms of the group or community to which the agent belongs that have been internalized by the agent. This is the essence of the philosophical doctrine of moral relativism (see, *e.g.*, [20]).

[2] Note a third option can envisaged, namely two of an agent's desires conflicting with each other. However, this is not properly a moral dilemma as it does not involve a moral value of the agent.

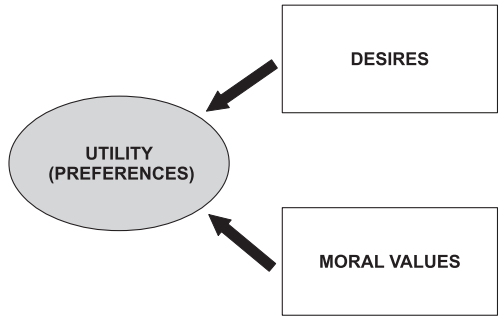

Figure 1: Dual view of moral choice: desires and moral values of an agent are two different parameters affecting the agent's preference (or utility) over alternatives.

parameters affecting the agent's preference (or utility) over alternatives (see Figure 1).[3] An alternative to Harsanyi's dual view of moral choice is the meta-cognitive view defended among the others by Sen [37, 36]. According to the meta-cognitive view, moral judgments are *rankings of preference rankings*. In particular, given a set of possible outcomes X and the set O of all possible orderings of the elements of X, a moral judgments can be defined as a reflexive and transitive relation (*i.e.*, a quasi-ordering) over the elements of O. As emphasized by Sen, one might interpret this relation as a moral value to have one preference pattern over outcomes rather than another. We here concentrate on the dual view of moral choice, leaving for future work a comparison with the meta-cognitive view.

The dual view of moral choice allows us to distinguish between *self-regarding* agents and *moral* agents. 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 fulfillment of his own moral values. In other words, if an agent is purely self-regarding, the utility of an action 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 an action 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. Of course, purely self-regarding agents and purely moral agents are just extremes cases. An agent is more or less moral depending on whether

[3] In his theory of morality, Harsanyi provides support for an utilitarian interpretation of moral motivation. Specifically, Harsanyi argues that an agent's moral motivation coincides with the goal of maximizing the collective utility represented by the weighted sum of the individuals' utilities.

the utility of a given option for him is more or less affected by his moral values. More precisely, the higher the influence of moral values on the utility of a given decision option, the more moral the agent. The extent to which an agent's utility is affected by his moral attitudes can be called *degree of moral sensitivity*. More generally, this is a numerical value that describes how much an agent's decisions are influenced by the agent's moral considerations.[4] It will be a fundamental building block of the logical model of morality developed in Section 4. Note that an agent cannot be purely moral and purely self-regarding at the same time, because being purely moral coincides with having a *maximal* degree of moral sensitivity, being purely self-regarding coincides with having a *minimal* degree of moral sensitivity, and an agent cannot have a *maximal* degree of moral sensitivity and a *minimal* degree of moral sensitivity at the same time.

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. A self-regarding/moral agent generates his utilities in a certain way. Therefore, classical decision theory and ethical considerations can be perfectly consistent with one another. This view is supported among the others by the philosopher John Broome [9], according to whom utility is conceived as that which represents an agent's preferences and rationality is defined as acting according to the agent's preferences. This suggests that preferences may also incorporate ethical issues. However, the fact that classical decision theory and ethical considerations are consistent does not exclude situations in which the two make contradicting suggestions. This would mean that an agent's utilities are mainly influenced by his desires (*i.e.*, the agent has a low degree of moral sensitivity) and the utilities thus generated contradict what his morality would demand. For example, taking up the paradigmatic example of Adam and Eve again, Eve decides to take the forbidden fruit because her utilities are mainly influenced by the desire to eat the forbidden fruit. Eve's utilities thus generated contradict with her moral obligation.

3. Logical framework

In the following sections the logic LAMA (*Logic of Actions and Mental Attitudes*) is presented. LAMA is a modal logic which supports reasoning about actions of agents and of coalitions of agents. LAMA also allows us to describe epistemic states of agents as well as their desires and moral values.

[4] For a similar idea in current economic models of moral choice and moral emotions see, *e.g.*, [5, 1].

We first present the syntax and the semantics of LAMA (Sections 3.1, 3.2 and 3.3). A complete axiomatization of the logic is given in Section 3.4. In Section 4 the logic LAMA will be used to provide a logical analysis of the different aspects of morality discussed in Section 2.

3.1. *Syntax*

Assume a countable set of atomic propositions $Atm = \{p, q, \ldots\}$, a countable set of atomic action types $Act = \{a, b, \ldots\}$, a finite set of agents $Agt = \{i, j, \ldots\}$ and a finite set of natural numbers $Num = \{x \in \mathbb{N} : 0 \leq x \leq \mathsf{maxVal}\}$, with $\mathsf{maxVal} \in \mathbb{N}^+ = \mathbb{N} \setminus \{0\}$.

The function $Rep : Agt \rightarrow 2^{Act}$ associates to every agent i a *finite* repertoire of actions $Rep(i) \subseteq Act$. The property that $Rep(i)$ is finite is justified by the reasonable assumption that the action repertoire of either a human agent or an artificial agent (*e.g.*, a robot, a virtual agent) includes finitely many action units and basic behaviors.

$2^{Agt^*} = 2^{Agt} \setminus \{\emptyset\}$ is the set of non-empty sets of agents, also called *coalitions*. Elements of 2^{Agt^*} are denoted by symbols H, J,… For every $H \in 2^{Agt^*}$, let

$$JAct_H = \prod_{i \in H} Rep(i)$$

be the set of all possible *joint actions* of coalition H. Elements of $JAct_H$ are denoted by symbols δ_H, $\delta_{H'}$, $\delta_{H''}$,… Every δ_H in $JAct_H$ should be understood a total function $\delta_H : H \rightarrow \bigcup_{i \in H} Rep(i)$ such that $\delta_H(i) \in Rep(i)$ for all $i \in H$. For notational convenience we write $JAct$ instead of $JAct_{Agt}$. Elements of $JAct$ are denoted by symbols δ, δ', δ'',… Being δ a function it can be represented as a set of mappings $\{i \mapsto \delta(i) : i \in Agt\}$. 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.

Finally, let $JAct^*$ be the set of all (possibly infinite) sequences of joint actions in $JAct$. Elements of $JAct^*$ are denoted by symbols ε, ε', ε'',… Note that $JAct^*$ also contains the empty sequence. For every $\varepsilon_1, \varepsilon_2 \in JAct^*$, we write $\varepsilon_1 \sqsubseteq \varepsilon_2$ to mean that ε_1 is an initial subsequence (a prefix) of ε_2, *i.e.*, there is $\varepsilon_3 \in JAct^*$ such that $\varepsilon_2 = \varepsilon_1; \varepsilon_3$, where the symbol ";" denotes the operation of composition (or concatenation) of sequences.[5] For notational convenience, in what follows we write δ instead of (δ) to denote the length-one sequence.

[5] Given two sequences .. and $\varepsilon_3 = (\delta'_1, \ldots, \delta'_m)$ in $JAct^*$, $\varepsilon_1; \varepsilon_3$ is defined to be the sequence $(\delta_1, \ldots, \delta_n, \delta'_1, \ldots, \delta'_m)$.

Let $Hist \subseteq JAct^*$ be the set of all infinite sequences of joint actions. Elements of *Hist* are called *histories* are denoted by symbols $h, h', \ldots$. Note that a history h can also be seen as a total function $h : \mathbb{N}^+ \rightarrow JAct$. Elements of *Hist* are denoted by symbols $h, h', \ldots$

The language $\mathcal{L}_{\mathsf{LAMA}}$ of LAMA is defined by the following grammar in Backus-Naur Form:

$$\varphi ::= p \mid \mathsf{occ}_{\varepsilon} \mid \mathsf{pls}_{i,\mathsf{k}} \mid \mathsf{idl}_{i,\mathsf{k}} \mid \neg\varphi \mid \varphi_1 \wedge \varphi_2 \mid [\![\delta]\!]\varphi \mid \mathsf{K}_i\varphi$$

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

We define LAMA^- to be the fragment of LAMA without dynamic operators $[\![\delta]\!]$. Its language $\mathcal{L}_{\mathsf{LAMA}^-}$ is defined by the following grammar:

$$\varphi ::= p \mid \mathsf{occ}_{\varepsilon} \mid \mathsf{pls}_{i,\mathsf{k}} \mid \mathsf{idl}_{i,\mathsf{k}} \mid \neg\varphi \mid \varphi_1 \wedge \varphi_2 \mid \mathrm{K}_i\varphi$$

where p ranges over *Atm*, i ranges over *Agt*, ε ranges over $JAct^*$ and k ranges over *Num*.

The logic LAMA has special atomic formulas of three different kinds. The atomic formulas $\mathsf{occ}_{\varepsilon}$ represent information about occurrences of joint action sequences. The formula $\mathsf{occ}_{\varepsilon}$ has to be read "the joint action sequence ε is going to occur". For notational convenience we provide the following abbreviation:

$$\mathsf{choose}_{i,a} \stackrel{\mathsf{def}}{=} \bigvee_{\delta \in JAct:\, \delta(i)=a} \mathsf{occ}_{\delta}$$

$\mathsf{choose}_{i,a}$ has to be read "agent i decides to perform the action a" or, more simply, "agent i chooses action a".

The atomic formulas $\mathsf{pls}_{i,\mathsf{k}}$ and $\mathsf{idl}_{i,\mathsf{k}}$ are used to rank the histories that an agent can imagine at a given world according to their *pleasantness* degree and to their *ideality* degree for the agent. *Pleasantness* 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 fulfillment of the agent's moral values). Formula $\mathsf{pls}_{i,\mathsf{k}}$ has to be read "the current history has for agent i a degree of pleasantness equal to k" while formula $\mathsf{idl}_{i,\mathsf{k}}$ has to be read "the current history has for agent i a degree of ideality equal to k". The restriction that the set *Num* should be bounded by some integer maxVal is necessary in order to be able to define in LAMA many concepts such as desires and moral values (see Section 4.1) and preferences over actions induced by desires and moral values (see Section 4.2)

without using an infinitary language. However, this restriction of the model is also justified by the fact that the idea that an individual may have an *unbounded* value scale poses a conceptual problem and is not even compatible with classical formulations of decision theory (see [3] for more details).[6] However, to make our approach more general, it would be interesting to allow different agents to have scales of degrees of pleasantness and ideality bounded by different integers. In order to do this, we shall associate to every agent i a finite set of integers Num_i. Such a generalization of the logic LAMA is subject of future work.

The logic LAMA has two kinds of modal operators: $[\![\delta]\!]$ and K_i. $[\![\delta]\!]$ 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 proposition holds. In particular, $[\![\delta]\!]\varphi$ has to be read "if the joint action 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.*, $\langle\!\langle\delta\rangle\!\rangle\varphi \stackrel{\text{def}}{=} \neg[\![\delta]\!]\neg\varphi$, where $\langle\!\langle\delta\rangle\!\rangle\varphi$ has to be read "the joint action δ is performed and φ will be true after its execution". It is worth noting that the reading of the operator $[\![\delta]\!]$ is a bit unusual. In fact, in propositional dynamic logic (PDL) [19] $[\alpha]\varphi$ means that "every execution of α leads to a state in which φ is true". However, the reading of the operator $[\![\delta]\!]$ makes perfect sense, as in LAMA it is assumed that time evolves linearly without branching. Specifically, it is assumed that every state has a unique *actual* result state that is reached by the execution of an *actual* (not merely *potential*) joint action δ. Due to the underlying assumption about linearity of time and the fact that the set of joint actions *JAct* is finite, LAMA allows us to the define the operator *next* of linear temporal logic as follows:

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

[6] A classical argument against the idea that people may have unbounded utilities is the famous St. Petersburg paradox. Consider the following lottery: a coin is tossed repeatedly until tails comes up. If this happens in the kth toss, the outcome out_k is obtained whose utility is 2^{k-1}. This implies that the utility function u is unbounded (*i.e.*, there is no $h\in\mathbb{N}$ such that $u(out_k)\leq h$ for all outcomes out_k). Since the probability of outcome out_k is $\frac{1}{2^k}$, the expected utility of this lottery is $\frac{1}{2}\cdot 1+\frac{1}{4}\cdot 2+\frac{1}{8}\cdot 4+\ldots=\frac{1}{2}+\frac{1}{2}+\frac{1}{2}+\ldots=\infty$. Therefore, according to expected utility theory (EUT), an individual should be willing to give up everything for the opportunity to play this lottery, that seems a "...patently absurd conclusion..." [29, pag. 185]. There are other incompatibilities between the idea of unbounded utility and classical decision theory. For instance, the monotonicity axiom in Luce & Raiffa's axiomatization of expected utility [27] implies that if (1) one prefers outcome out_1 to outcome out_2 (*i.e.*, $u(out_1) > u(out_2)$), (2) lottery 1 consists in a probability $\mathbf{p}$ of obtaining outcome out_1 and a probability $1-\mathbf{p}$ of obtaining outcome out_2, (3) lottery 2 consists in a probability $\mathbf{q}$ of obtaining outcome out_1 and a probability $1-\mathbf{q}$ of obtaining outcome out_2, and (4) $\mathbf{p} > \mathbf{q}$, then one must prefer lottery 1 over lottery 2. But if the utility of outcome out_1 is infinite then the expected utility of lottery 1 is equal to the expected utility of lottery 2 which is in contradiction with the monotonicity axiom.

$\mathsf{X}\varphi$ has to read "φ will be true in the next state".

Finally, K_i is the S5 epistemic modal operator which is commonly used in computer science [14] and in game theory [4] to model the notion of knowledge. We assume that K_i characterizes the concept of *ex ante* knowledge in the sense of Aumann & Dreze [2] (see also [34]). The formula $\mathsf{K}_i\varphi$ has to be read "agent i knows that φ" or "φ is true in all worlds that agent i envisages". The dual of the operator K_i is denoted by $\hat{\mathsf{K}}_i$, *i.e.*, $\hat{\mathsf{K}}_i\varphi \stackrel{\text{def}}{=} \neg\mathsf{K}_i\neg\varphi$. 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 LAMA by the following operator K_i^*:

$$\mathsf{K}_i^*\varphi \stackrel{\text{def}}{=} \bigwedge_{a\in Rep(i)} (\mathsf{choose}_{i,a} \rightarrow \mathsf{K}_i(\mathsf{choose}_{i,a} \rightarrow \varphi))$$

where $\mathsf{K}_i^*\varphi$ has to be read "φ is true in all worlds that agent i envisages and that are compatible with his current choice" or, more shortly, "the agent i knows that φ is true, given his current choice". The logical relationship between *ex ante* knowledge and *interim* knowledge will be discussed in more detail at the end of Section 3.3.

Before concluding this section, a remark is in order. One might object that the readings of the LAMA formulas $\mathsf{choose}_{i,a}$ and $\langle\langle\delta\rangle\rangle\varphi$ are problematic, as symbols a and δ denote respectively action types and joint action types. Since action types cannot be performed, an agent cannot decide to perform an action type. Our reply to this criticism is that formulas $\mathsf{choose}_{i,a}$ and $\langle\langle\delta\rangle\rangle\varphi$ can be interpreted, without loss of generality, as describing the fact that "an action *token* of the action *type* a is performed by agent i" and the fact that "a joint action *token* of the joint action *type* δ is performed and φ will be true after its occurrence". Indeed, although the logic LAMA does not represent action tokens as first-class objects and cannot represent them in the object language (*i.e.*, LAMA cannot represent the fact that a specific action *token* t of the action type a occurs), it does not prevent from interpreting $\langle\langle\delta\rangle\rangle\varphi$ as a formula describing the consequence φ of an action token of the joint action type δ and $\mathsf{choose}_{i,a}$ as a formula describing the fact that agent i decides to perform an action token of the action type a. Note that a similar comment applies to propositional dynamic logic PDL [19] in which state transitions in a model are labeled with symbols denoting action types (PDL models are nothing but labeled graphs where labels are action types). Suppose that in a certain PDL model there is a transition labeled with an

action type α from a state w to a state v. It is not necessary to interpret this as meaning that "the state v is a consequence of the action type α executed in the state w". The PDL semantics allows us to interpret this as "the state v is a consequence of an action token of the action type α executed in the state w". The only limitation of PDL is that it does not allow us to name the action token which is responsible for the transition from state w to state v. Consequently, one does not need to assume that states in a PDL model are state types.

3.2. *Action description*

Similarly to Situation Calculus [33], in LAMA actions are described in terms of their positive and negative effect preconditions. In particular, we introduce an action description (γ^+, γ^-) as a specific parameter of the logic LAMA, where γ^+ and γ^- are two functions:[7]

$$\gamma^+ \colon Agt \times Act \times Atm \rightarrow \mathcal{L}_{\mathsf{LAMA}^-}$$

$$\gamma^- \colon Agt \times Act \times Atm \rightarrow \mathcal{L}_{\mathsf{LAMA}^-}$$

mapping agents, actions and atomic propositions to formulas in the fragment of LAMA without dynamic operators $[\![\delta]\!]$ defined above.

The formula $\gamma^+(i,a,p)$ describes the *positive effect precondition* of action a performed by agent i with respect to p, whereas $\gamma^-(i,a,p)$ describes the *negative effect precondition* of action a performed by agent i with respect to p. In particular, 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 formula $\gamma^-(i,a,p)$ 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. For example, the positive effect precondition for Mary to hit the target by shooting is that Mary's arm is free and Mary's gun is loaded, *i.e.*, γ^+ (*Mary*, *shoot*, *hitTarget*) = *MaryArmfree* $\wedge$ *MaryGunLoaded*; the negative effect precondition for Mary to make the door closed by opening it is that Mary's arm is free and the door is not locked, *i.e.*, γ^- (*Mary, open, doorClosed*) = *MaryArmfree* $\wedge$ $\neg$*doorLocked.*

[7] The action description (γ^+, γ^-) is a metalogical entity that is not part of the object language or the LAMA semantics. Thus, LAMA could also be conceived as a "family" of logics, each of which is parameterized by a certain action description (γ^+, γ^-).

In order to avoid misunderstandings, let us emphasize that $\gamma^+(i,a,p)$ and $\gamma^+(i,a,p)$ *are not* primitive formulas with a special semantics. The symbol $\gamma^+(i,a,p)$ simply denotes the formula φ of the language $\mathcal{L}_{\mathsf{LAMA}^-}$ such that $\gamma^+(i,a,p)=\varphi$, while the symbol $\gamma^-(i,a,p)$ simply denotes the formula φ of the language $\mathcal{L}_{\mathsf{LAMA}^-}$ such that $\gamma^-(i,a,p)=\varphi$.

3.3. *Semantics*

The semantics of LAMA is a possible world semantics with accessibility relations associated with each modal operator, with functions for *pleasantness* and *ideality*, and with a function designating the history starting in a given world.

Definition 1. *A LAMA model is a tuple* $M=\langle W,\mathcal{H},\{\mathcal{E}_i\}_{i\in Agt},\mathcal{P},\mathcal{I},\mathcal{V}\rangle$ *where:*

- W *is a non-empty set of states (or worlds),*
- $\mathcal{H}$ *is a total function* $\mathcal{H}:W\rightarrow Hist$,
- *every* $\mathcal{E}_i$ *is an equivalence relation between states in W,*
- $\mathcal{P}:W\times Agt\rightarrow Num$ *and* $\mathcal{I}:W\times Agt\rightarrow Num$ *are total functions mapping worlds and agents to natural numbers in Num,*
- $\mathcal{V}:W\rightarrow 2^{Atm}$ *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, we introduce the function $\mathcal{C}:W\times Agt\rightarrow Act$ such that, for all $w\in W$, for all $i\in Agt$ and for all $a\in Act$: $\mathcal{C}(w,i)=a$ if and only if there is $\delta\in JAct$ such that $\delta(i)=a$ and $\delta\sqsubseteq\mathcal{H}(w)$. $\mathcal{C}(w,i)$ is the action chosen by agent i at world w.

The equivalence relations $\mathcal{E}_i$ are used to interpret the 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 set* 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 set at w and v. For every agent $i\in Agt$ and for every world $w\in W$ let

$$\|\varphi\|_{w,i}=\{v\in W: M, v\vDash\varphi \text{ and } w\mathcal{E}_i v\}$$

be the subset of agent i's epistemic alternatives at w in which φ is true.

The functions $\mathcal{P}$ and $\mathcal{I}$ represent respectively pleasantness grading and ideality grading of the possible worlds and are used to interpret the atomic

formulas $\mathsf{pls}_{i,\mathsf{k}}$ and $\mathsf{idl}_{i,\mathsf{k}}$. $\mathcal{P}(w,i)=\mathsf{k}$ means that according to the agent i the history starting in the world w has a degree of pleasantness k, whereas $\mathcal{I}(w,i)=\mathsf{k}$ means that according to the agent i the history starting in the world w has a degree of ideality k. (Remember that every world w in a LAMA model is identified with the history $\mathcal{H}(w)$ starting in w). Since the functions $\mathcal{P}$ and $\mathcal{I}$ are total, every agent can compare every two states with respect to their pleasantness (and ideality) value. This assumption comes with the other assumption underlying the LAMA semantics, namely that in every state an agent decides to perform an action, *i.e.*, for all $w\in W$ and $i\in Agt$ there is $a\in Act$ such that $\mathcal{C}(w,i)=a$. This presupposes that agents are always capable of comparing the utility of different decision options and of choosing the best one among them, or one of the best ones if there are more than one (see Section 4.2 for more details about the notion of utility). In future developments of the logical theory, we plan to relax these assumptions in order to have a more realistic account of moral choice in which an agent may be ignorant about the pleasantness or ideality of a state and may be incapable of comparing the utility of two different decision options.

The rules defining truth conditions of LAMA formulas are the standard ones for atomic propositions and for the Boolean operations *plus* the following ones:

$$
\begin{array}{lll}
M,w\vDash \mathsf{occ}_{\varepsilon} & \text{iff} & \varepsilon\sqsubseteq\mathcal{H}(w)\\
M,w\vDash \mathsf{pls}_{i,\mathsf{k}} & \text{iff} & \mathcal{P}(w,i)=\mathsf{k}\\
M,w\vDash \mathsf{idl}_{i,\mathsf{k}} & \text{iff} & \mathcal{I}(w,i)=\mathsf{k}\\
M,w\vDash [\![\delta]\!]\varphi & \text{iff} & \text{if } M,w\vDash \mathsf{occ}_{\delta} \text{ then } M^{\delta},w\vDash\varphi\\
M,w\vDash \mathsf{K}_i\varphi & \text{iff} & \forall v\in\mathcal{E}_i(w): M,v\vDash\varphi
\end{array}
$$

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

Definition 2 (Update via joint action)**.** *Given a LAMA model* $M=\langle W,\mathcal{H},\{\mathcal{E}_i\}_{i\in Agt},\mathcal{P},\mathcal{I},\mathcal{V}\rangle$*, the update of* M *by* δ *is defined to be* $M^{\delta}=\langle W^{\delta},\mathcal{H}^{\delta},\{\mathcal{E}_i^{\delta}\}_{i\in Agt},\mathcal{P}^{\delta},\mathcal{I}^{\delta},\mathcal{V}^{\delta}\rangle$ *where:*

$$W^{\delta}=\{w\in W: M,w\vDash \mathsf{occ}_{\delta}\}$$

and for all $i\in Agt$, $w\in W^{\delta}$, $h\in Hist$*:*

$$
\begin{array}{lll}
\mathcal{E}_i^{\delta} & = & \mathcal{E}_i\cap(W^{\delta}\times W^{\delta})\\
\mathcal{H}^{\delta}(w) & = & h \textit{ iff } \mathcal{H}(w)=\delta;h\\
\mathcal{P}^{\delta}(w,i) & = & \mathcal{P}(w,i)
\end{array}
$$

$$\begin{aligned}
\mathcal{I}^{\delta}(w,i) &= \mathcal{I}(w,i)\\
\mathcal{V}^{\delta}(w) &= (\mathcal{V}(w)\setminus\{p:(\exists i\in Agt\ \exists a\in Act:\delta(i)=a \text{ and } M,w\models\gamma^{-}(i,a,p))\\
&\quad \text{and } (\forall j\in Agt\ \forall b\in Act, \text{if } \delta(j)=b \text{ then } M,w\models\neg\gamma^{+}(j,b,p))\})\cup\\
&\quad \{p:(\exists i\in Agt\ \exists a\in Act:\delta(i)=a \text{ and } M,w\models\gamma^{+}(i,a,p))\\
&\quad \text{and } (\forall j\in Agt\ \forall b\in Act, \text{if } \delta(j)=b \text{ then } M,w\models\neg\gamma^{-}(j,b,p))\}
\end{aligned}$$

Definition 2 guarantees that the performance of a joint action δ restricts the model M to the worlds in which the joint action occurs (see the definition of W^{δ}). Moreover, it modifies the atomic propositions via the positive effect preconditions and the negative effect preconditions, as 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 precondition with respect to p holds and there is no other action in the joint action δ whose negative effect precondition with respect to p holds, then p will be true after the occurrence of δ; if there is an action in the joint action δ whose negative effect precondition with respect to p holds and there is no other action in the joint action δ whose positive effect precondition with respect to p holds, 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 $\mathcal{H}^{\delta}$ where $\delta;h$ is the composition of the length-one sequence δ and the infinite sequence/history h). As to the epistemic accessibility relations $\mathcal{E}_i$, they are restricted to the new set of worlds W^{δ} (see the definition of $\mathcal{E}_i^{\delta}$). Finally, the joint action δ does not modify the agents' pleasantness and ideality grading over the histories (see the definitions of $\mathcal{P}^{\delta}$ and $\mathcal{I}^{\delta}$).

As stated by the following proposition, the update via a joint action preserves the constraints on LAMA models. Indeed, it is trivial to show that $\mathcal{H}^{\delta}$, $\mathcal{P}^{\delta}$ and $\mathcal{I}^{\delta}$ are total functions and that every $\mathcal{E}_i^{\delta}$ is an equivalence relation.

Proposition 1. *If M is a LAMA model then M^{δ} is a LAMA model too.*

Figure 2 provides an illustration of the overall LAMA semantics presented above. The initial model M has four worlds w_1, w_2, w_3 and w_4 and four corresponding histories (*i.e.*, the history starting at w_1, the history starting at w_2, the history starting at w_3 and the history starting at w_4). Moreover, it has two information sets for agent i, *i.e.*, $\{w_1,w_2,w_3\}$ and $\{w_4\}$. The four histories are ranked by agent i according to their pleasantness degrees and to their ideality degrees. For example, the history starting at w_1 has for agent i a degree of pleasantness equal to 0 and a degree of ideality equal to 2. After the occurrence of the joint action δ at world w_1 in model M, the set of possible worlds is restricted to the worlds in in which δ occurs. The resulting model is M^{δ}. After the occurrence of the joint action δ' at world w_1 in model M^{δ}, the set of possible worlds and agent i's information set are

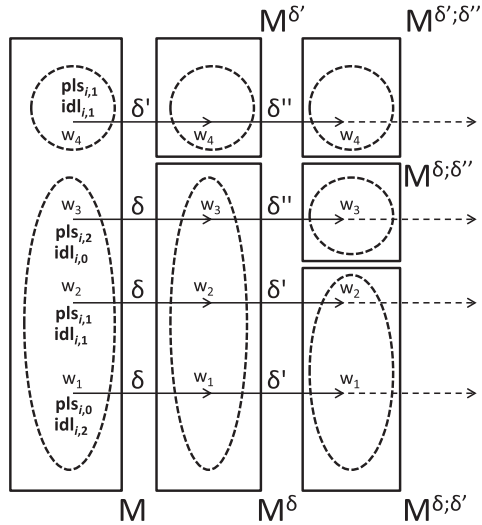

Figure 2: Example of LAMA semantics:
dotted ellipses represent agent i's information sets.

restricted to the worlds in which δ' occurs. The resulting model is $(M^{\delta})^{\delta'}$ and the resulting information set of agent i is $\{w_1, w_2\}$.

Note that the LAMA semantics relies on a linear (non-branching) conception of history where every pointed model (M, w) occurs (at most) once in a history and where every transition from a pointed model (M, w) to a pointed model (M^{δ}, w) is labeled with a joint action type δ. As emphasized at the end of Section 3.1, one can interpret this as "the pointed model (M, w) is a consequence of the occurrence of a token of the joint action type δ" and does not need to interpret it as "the pointed model (M, w) is a consequence of the occurrence of the joint action type δ".

In what follows, we say that a LAMA formula φ is true in a LAMA model M if φ is true in all worlds of the model M. Moreover, we write $\vDash \varphi$ to mean that formula φ is LAMA valid, *i.e.*, φ is true in every LAMA model.

The following are examples of LAMA validities which highlight some properties of the *interim* knowledge operator and some of its relationships with the *ex ante* knowledge operator.

Proposition 2. *For every $i \in Agt$ we have:*

$$\vDash (\mathsf{K}_i^* \varphi \wedge \mathsf{K}_i^* \psi) \leftrightarrow \mathsf{K}_i^* (\varphi \wedge \psi) \tag{2}$$

$$\text{if } \vDash \varphi \text{ then } \vDash \mathsf{K}_i^* \varphi \tag{3}$$

$$\vDash \mathsf{K}_i^* \varphi \rightarrow \varphi \tag{4}$$

$$\models \mathsf{K}_i^* \varphi \rightarrow \mathsf{K}_i^* \mathsf{K}_i^* \varphi \tag{5}$$

$$\models \neg \mathsf{K}_i^* \varphi \rightarrow \mathsf{K}_i^* \neg \mathsf{K}_i^* \varphi \tag{6}$$

$$\models \mathsf{choose}_{i,a} \rightarrow \mathsf{K}_i^* \, \mathsf{choose}_{i,a} \tag{7}$$

$$\models \mathrm{K}_i \varphi \rightarrow \mathsf{K}_i^* \varphi \tag{8}$$

Like *ex ante* knowledge operators, *interim* knowledge operators are S5 normal modal operators (validities (2)-(6)). Moreover, an agent has always *interim* knowledge of the action he chooses (validity (7)). Finally, *ex ante* knowledge is stronger than *interim* knowledge, *i.e.*, knowing that φ in an *ex ante* sense implies knowing that φ in an *interim* sense (validity (8)).

The following validity captures the basic property of the temporal operator *next* defined in Section 3.1:

$$\models \mathsf{X}\varphi \leftrightarrow \neg \mathsf{X} \neg \varphi$$

This validity follows from two more basic validities of the LAMA: namely $\models \langle\!\langle \delta \rangle\!\rangle \varphi \rightarrow [\![\delta']\!] \varphi$ for all $\delta, \delta' \in \mathit{JAct}$, and $\models \bigvee_{\delta \in \mathit{JAct}} \langle\!\langle \delta \rangle\!\rangle \top$.

3.4. *Axiomatization*

Our logic LAMA has so-called reduction axioms. These axioms allow us to eliminate all the dynamic operators $[\![\delta]\!]$ from formulas. Moreover, together with a theory for the atomic formulas and an S5 axiomatics for the epistemic operators K_i, they provide an axiomatization for the logic LAMA.

Proposition 3. *The following formulas are LAMA valid for all* $\delta, \delta' \in \mathit{JAct}$ *such that* $\delta \neq \delta'$*, for all* $\varepsilon, \varepsilon' \in \mathit{JAct}^*$ *such that* $\varepsilon' \sqsubseteq \varepsilon$ *and for all* $\mathsf{k}, \mathsf{g} \in \mathit{Num}$ *such that* $\mathsf{k} \neq \mathsf{g}$*:*

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

$$\mathsf{occ}_{\varepsilon;\delta} \rightarrow \neg \mathsf{occ}_{\varepsilon;\delta'} \qquad \textbf{(UniqueJAct)}$$

$$\mathsf{occ}_{\varepsilon} \rightarrow \mathsf{occ}_{\varepsilon'} \qquad \textbf{(SubSeqJAct)}$$

$$\bigvee_{\mathsf{k} \in \mathit{Num}} \mathsf{pls}_{i,\mathsf{k}} \qquad \textbf{(ComplDes)}$$

$$\mathsf{pls}_{i,\mathsf{k}} \rightarrow \neg \mathsf{pls}_{i,\mathsf{g}} \qquad \textbf{(UniqueDes)}$$

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

$$\mathsf{idl}_{i,\mathsf{k}} \rightarrow \neg \mathsf{idl}_{i,\mathsf{g}} \qquad \textbf{(UniqueIdl)}$$

Note that there is an infinite number of axioms represented by the axiom schemata (**OneJAct**), (**UniqueJAct**) and (**SubSeqJAct**), as the set of joint action sequences $JAct^*$ is infinite.

Proposition 4. *The following equivalences are* LAMA *valid for all* $p \in Atm$, $i \in Agt$, $\delta \in JAct$, $\varepsilon \in JAct^*$ *and* $\mathsf{k} \in Num$:

$$
\begin{array}{lcl}
[\![\delta]\!]p & \leftrightarrow & (\mathsf{occ}_\delta \rightarrow ((\bigvee_{i\in Agt}\gamma^+(i,\delta(i),p) \wedge \bigwedge_{j\in Agt}\neg\gamma^-(j,\delta(j),p)) \vee \\
 & & (p \wedge \bigwedge_{i\in Agt}\neg\gamma^-(i,\delta(i),p)) \vee (p \wedge \bigvee_{i\in Agt}\gamma^+(i,\delta(i),p))))\\
[\![\delta]\!]\neg\varphi & \leftrightarrow & (\mathsf{occ}_\delta \rightarrow \neg[\![\delta]\!]\varphi)\\
[\![\delta]\!](\varphi \wedge \psi) & \leftrightarrow & ([\![\delta]\!]\varphi \wedge [\![\delta]\!]\psi)\\
[\![\delta]\!]\,\mathsf{occ}_\varepsilon & \leftrightarrow & (\mathsf{occ}_\delta \rightarrow \mathsf{occ}_{\delta;\varepsilon})\\
[\![\delta]\!]\,\mathsf{pls}_{i,\mathsf{k}} & \leftrightarrow & (\mathsf{occ}_\delta \rightarrow \mathsf{pls}_{i,\mathsf{k}})\\
[\![\delta]\!]\,\mathsf{idl}_{i,\mathsf{k}} & \leftrightarrow & (\mathsf{occ}_\delta \rightarrow \mathsf{idl}_{i,\mathsf{k}})\\
[\![\delta]\!]\,\mathsf{K}_i\varphi & \leftrightarrow & (\mathsf{occ}_\delta \rightarrow \mathsf{K}_i[\![\delta]\!]\varphi)
\end{array}
$$

As the rule of replacement of equivalents preserves validity, the equivalences of Proposition 4 together with the rule of replacement of equivalents allow us to reduce every LAMA formula to an equivalent formula in LAMA$^-$, *i.e.*, the fragment of LAMA without dynamic operators $[\![\delta]\!]$ defined in Section 3.1.

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 5. *Let* φ *be a formula in the language of* LAMA. *Then*

1. $red(\varphi)$ *has no dynamic operators* $[\![\delta]\!]$
2. $red(\varphi) \leftrightarrow \varphi$ *is* LAMA *valid*

Note that the second item is proved using Proposition 4 and the rule of replacement of equivalents.

Theorem 1. *The validities of* LAMA *are completely axiomatized by*

- *all principles of classical propositional logic*
- *axioms and rules of inference of the normal modal logic S5 for each epistemic operator* K_i
- *the schemas of Proposition 3*
- *the reduction axioms of Proposition 4*
- *the rule of replacement of equivalents*

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

4. Moral agency: a logical formalization

In what follows the logic LAMA 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*

In order to define the concepts of desire and moral value, we extend the pleasantness and the ideality degrees of a possible world to the pleasantness and the ideality degrees of a formula viewed as a set of worlds. We provide two different definitions of desire and moral value depending on whether the assessment of pleasantness or ideality of a given state of affairs is optimistic or pessimistic. When assessing the pleasantness/ideality of a formula φ in an optimistic way, an agent focuses on his epistemic alternatives with maximal degree of pleasantness/ideality in which φ is true. On the contrary, when assessing the pleasantness/ideality of a formula φ in a pessimistic way, an agent focuses on his epistemic alternatives with minimal degree of pleasantness/ideality in which φ is true. So, while an optimistic evaluation consists in a kind of best case analysis of the epistemic alternatives in which φ is true, a pessimistic evaluation consists in a kind of worst case analysis.

Definition 3 (Optimistic and pessimistic desires and moral values)**.** *Given a LAMA model* $M = \langle W, \mathcal{H}, \{\mathcal{E}_i\}_{i \in Agt}, \mathcal{P}, \mathcal{I}, \mathcal{V} \rangle$ *we say that at world w in M:*

- *agent i has an optimistic desire that* φ *with strength* k *(or agent i has a desire that* φ *with strength* k *based on an optimistic evaluation), i.e.,* $M, w \vDash \mathsf{ODes}_i^{\mathsf{k}}\varphi$*, if and only if* $\mathcal{P}_{w,i}^{\mathsf{opt}}(\varphi) = \mathsf{k}$*,*
- *agent i has an optimistic moral value that* φ *with strength* k *(or agent i has a moral value that* φ *with strength* k *based on an optimistic evaluation), i.e.,* $M, w \vDash \mathsf{OVal}_i^{\mathsf{k}}\varphi$*, if and only if* $\mathcal{I}_{w,i}^{\mathsf{opt}}(\varphi) = \mathsf{k}$*,*
- *agent i has a pessimistic desire that* φ *with strength* k *(or agent i has a desire that* φ *with strength* k *based on a pessimistic evaluation), i.e.,* $M, w \vDash \mathsf{PDes}_i^{\mathsf{k}}\varphi$*, if and only if* $\mathcal{P}_{w,i}^{\mathsf{pess}}(\varphi) = \mathsf{k}$*,*
- *agent i has a pessimistic moral value that* φ *with strength* k *(or agent i has a moral value that* φ *with strength* k *based on a pessimistic evaluation), i.e.,* $M, w \vDash \mathsf{PVal}_i^{\mathsf{k}}\varphi$*, if and only if* $\mathcal{I}_{w,i}^{\mathsf{pess}}(\varphi) = \mathsf{k}$*,*

with:

$$\mathcal{P}_{w,i}^{\mathsf{opt}}(\varphi) \quad = \quad \max_{v \in \|\varphi\|_{w,i}} \mathcal{P}(v,i)$$

$$\mathcal{I}^{\mathsf{opt}}_{w,i}(\varphi) = \max_{v \in \|\varphi\|_{w,i}} \mathcal{I}(v,i)$$

$$\mathcal{P}^{\mathsf{pess}}_{w,i}(\varphi) = \min_{v \in \|\varphi\|_{w,i}} \mathcal{P}(v,i)$$

$$\mathcal{I}^{\mathsf{pess}}_{w,i}(\varphi) = \min_{v \in \|\varphi\|_{w,i}} \mathcal{I}(v,i)$$

and with the convention that $\max_{v \in \varnothing} \mathcal{P}(v,i) = \max_{v \in \varnothing} \mathcal{I}(v,i) = 0$ *and* $\min_{v \in \varnothing} \mathcal{P}(v,i) = \min_{v \in \varnothing} \mathcal{I}(v,i) = \mathsf{maxVal}$.

As the following proposition highlights, the four concepts of optimistic desire, optimistic ideal, pessimistic desire and pessimistic ideal semantically defined in Definition 3 are all syntactically expressible in the logic LAMA.

Proposition 6. *For all* $i \in Agt$ *and for all* $\mathsf{k} \in Num$ *we have:*

- $M, w \models \mathsf{ODes}^{\mathsf{k}}_i \varphi$ *if and only if:*
  - *if* $\mathsf{k} > 0$ *then* $M, w \models \hat{\mathsf{K}}_i(\mathsf{pls}_{i,\mathsf{k}} \wedge \varphi) \wedge \mathsf{K}_i((\bigvee_{\mathsf{g} \in Num : \mathsf{k} < \mathsf{g}} \mathsf{pls}_{i,\mathsf{g}}) \rightarrow \neg\varphi)$
  - *if* $\mathsf{k} = 0$ *then* $M, w \models \mathsf{K}_i((\bigvee_{\mathsf{g} \in Num : 0 < \mathsf{g}} \mathsf{pls}_{i,\mathsf{g}}) \rightarrow \neg\varphi)$
- $M, w \models \mathsf{OVal}^{\mathsf{k}}_i \varphi$ *if and only if:*
  - *if* $\mathsf{k} > 0$ *then* $M, w \models \hat{\mathsf{K}}_i(\mathsf{idl}_{i,\mathsf{k}} \wedge \varphi) \wedge \mathsf{K}_i((\bigvee_{\mathsf{g} \in Num : \mathsf{k} < \mathsf{g}} \mathsf{idl}_{i,\mathsf{g}}) \rightarrow \neg\varphi)$
  - *if* $\mathsf{k} = 0$ *then* $M, w \models \mathsf{K}_i((\bigvee_{\mathsf{g} \in Num : 0 < \mathsf{g}} \mathsf{idl}_{i,\mathsf{g}}) \rightarrow \neg\varphi)$
- $M, w \models \mathsf{PDes}^{\mathsf{k}}_i \varphi$ *if and only if:*
  - *if* $\mathsf{k} < \mathsf{maxVal}$ *then* $M, w \models \hat{\mathsf{K}}_i(\mathsf{pls}_{i,\mathsf{k}} \wedge \varphi) \wedge \mathsf{K}_i((\bigvee_{\mathsf{g} \in Num : \mathsf{g} < \mathsf{k}} \mathsf{pls}_{i,\mathsf{g}}) \rightarrow \neg\varphi)$
  - *if* $\mathsf{k} = \mathsf{maxVal}$ *then* $M, w \models \mathsf{K}_i((\bigvee_{\mathsf{g} \in Num : \mathsf{g} < \mathsf{maxVal}} \mathsf{pls}_{i,\mathsf{g}}) \rightarrow \neg\varphi)$
- $M, w \models \mathsf{PVal}^{\mathsf{k}}_i \varphi$ *if and only if:*
  - *if* $\mathsf{k} < \mathsf{maxVal}$ *then* $M, w \models \hat{\mathsf{K}}_i(\mathsf{idl}_{i,\mathsf{k}} \wedge \varphi) \wedge \mathsf{K}_i((\bigvee_{\mathsf{g} \in Num : \mathsf{g} < \mathsf{k}} \mathsf{idl}_{i,\mathsf{g}}) \rightarrow \neg\varphi)$
  - *if* $\mathsf{k} = \mathsf{maxVal}$ *then* $M, w \models \mathsf{K}_i((\bigvee_{\mathsf{g} \in Num : \mathsf{g} < \mathsf{maxVal}} \mathsf{idl}_{i,\mathsf{g}}) \rightarrow \neg\varphi)$

with the convention that a disjunction over an empty set is false (e.g., $\bigvee_{\mathsf{g} \in Num : \mathsf{maxVal} < \mathsf{g}} \mathsf{pls}_{i,\mathsf{g}} = \bot$*).*

It is worth noting that the operators of optimistic desire and optimistic ideal are operators of weak possibility (or potential possibility) in the sense of possibility theory, while the operators of pessimistic desire and pessimistic ideal are operators of strong possibility (or actual possibility) [12, 13]. In

the context of possibility theory they are also called operator Π and operator Δ.[8] These operators are characterized by the following decomposability properties.

Proposition 7. *For every $i \in Agt$ we have:*

$$\models (\mathsf{ODes}_i^{\mathsf{k}}\varphi \wedge \mathsf{ODes}_i^{\mathsf{g}}\psi) \rightarrow \mathsf{ODes}_i^{\max\{\mathsf{k},\mathsf{g}\}}(\varphi \vee \psi) \tag{7}$$

$$\models (\mathsf{OVal}_i^{\mathsf{k}}\varphi \wedge \mathsf{OVal}_i^{\mathsf{g}}\psi) \rightarrow \mathsf{OVal}_i^{\max\{\mathsf{k},\mathsf{g}\}}(\varphi \vee \psi) \tag{8}$$

$$\models (\mathsf{PDes}_i^{\mathsf{k}}\varphi \wedge \mathsf{PDes}_i^{\mathsf{g}}\psi) \rightarrow \mathsf{PDes}_i^{\min\{\mathsf{k},\mathsf{g}\}}(\varphi \vee \psi) \tag{9}$$

$$\models (\mathsf{PVal}_i^{\mathsf{k}}\varphi \wedge \mathsf{PVal}_i^{\mathsf{g}}\psi) \rightarrow \mathsf{PVal}_i^{\min\{\mathsf{k},\mathsf{g}\}}(\varphi \vee \psi) \tag{10}$$

$$\models (\mathsf{ODes}_i^{\mathsf{k}}\varphi \wedge \mathsf{ODes}_i^{\mathsf{g}}\psi) \rightarrow \mathsf{ODes}_i^{\leq\min\{\mathsf{k},\mathsf{g}\}}(\varphi \wedge \psi) \tag{11}$$

$$\models (\mathsf{OVal}_i^{\mathsf{k}}\varphi \wedge \mathsf{OVal}_i^{\mathsf{g}}\psi) \rightarrow \mathsf{OVal}_i^{\leq\min\{\mathsf{k},\mathsf{g}\}}(\varphi \wedge \psi) \tag{12}$$

$$\models (\mathsf{PDes}_i^{\mathsf{k}}\varphi \wedge \mathsf{PDes}_i^{\mathsf{g}}\psi) \rightarrow \mathsf{PDes}_i^{\geq\max\{\mathsf{k},\mathsf{g}\}}(\varphi \wedge \psi) \tag{13}$$

$$\models (\mathsf{PVal}_i^{\mathsf{k}}\varphi \wedge \mathsf{PVal}_i^{\mathsf{g}}\psi) \rightarrow \mathsf{PVal}_i^{\geq\max\{\mathsf{k},\mathsf{g}\}}(\varphi \wedge \psi) \tag{14}$$

where for $X \in \{\mathsf{ODes}, \mathsf{OVal}, \mathsf{PDes}, \mathsf{PVal}\}$:

$$X_i^{\geq\mathsf{k}}\varphi \stackrel{\text{def}}{=} \bigvee_{\mathsf{g}\in Num:\mathsf{g}\geq\mathsf{k}} X_i^{\mathsf{g}}\varphi$$

and

$$X_i^{\leq\mathsf{k}}\varphi \stackrel{\text{def}}{=} \bigvee_{\mathsf{g}\in Num:\mathsf{g}\leq\mathsf{k}} X_i^{\mathsf{g}}\varphi$$

Note that the first four validities use the "definite" values $\max\{\mathsf{k},\mathsf{g}\}$ / $\min\{\mathsf{k},\mathsf{g}\}$ while the last four validities use the "at least/at most" constructions $\leq\min\{\mathsf{k},\mathsf{g}\}$ / $\geq\max\{\mathsf{k},\mathsf{g}\}$ because: (1a) the maximum of the union of two sets is equal to the maximum of the maxima of the two sets and (1b) the minimun of the union of two sets is equal to the minimum of the minima of the two sets, while (2a) the maximum of the intersection of two sets is at most equal to the minimum of the maxima of the two sets but not necessarily equal and (2b) the minimun of the intersection of two sets is at least equal to the maximum of the minima of the two sets but not necessarily equal.

[8] See [10, 26] for a recent application of the operator Δ to the logical modeling of desires.

For every agent $i \in Agt$, we define four types of dyadic operators, for pleasantness and for ideality, both for the case of optimistic evaluation and for the case of pessimistic evaluation.

$$\psi \leq_i^{\mathsf{OPls}} \varphi \stackrel{\text{def}}{=} \bigvee_{\mathsf{k} \in Num} (\mathsf{ODes}_i^{\mathsf{k}} \varphi \wedge \bigwedge_{\mathsf{g} \in Num : \mathsf{g} > \mathsf{k}} \neg \mathsf{ODes}_i^{\mathsf{g}} \psi)$$

$$\psi \leq_i^{\mathsf{OIdl}} \varphi \stackrel{\text{def}}{=} \bigvee_{\mathsf{k} \in Num} (\mathsf{OVal}_i^{\mathsf{k}} \varphi \wedge \bigwedge_{\mathsf{g} \in Num : \mathsf{g} > \mathsf{k}} \neg \mathsf{OVal}_i^{\mathsf{g}} \psi)$$

$$\psi \leq_i^{\mathsf{PPls}} \varphi \stackrel{\text{def}}{=} \bigvee_{\mathsf{k} \in Num} (\mathsf{PDes}_i^{\mathsf{k}} \varphi \wedge \bigwedge_{\mathsf{g} \in Num : \mathsf{g} > \mathsf{k}} \neg \mathsf{PDes}_i^{\mathsf{g}} \psi)$$

$$\psi \leq_i^{\mathsf{PIdl}} \varphi \stackrel{\text{def}}{=} \bigvee_{\mathsf{k} \in Num} (\mathsf{PVal}_i^{\mathsf{k}} \varphi \wedge \bigwedge_{\mathsf{g} \in Num : \mathsf{g} > \mathsf{k}} \neg \mathsf{PVal}_i^{\mathsf{g}} \psi)$$

$\psi \leq_i^{\mathsf{OPls}} \varphi$ has to be read "according to agent i's optimistic evaluation, φ is at least as pleasant as ψ", $\psi \leq_i^{\mathsf{OIdl}} \varphi$ has to be read "according to agent i's optimistic evaluation, φ is at least as ideal as ψ", $\psi \leq_i^{\mathsf{PPls}} \varphi$ has to be read "according to agent i's pessimistic evaluation, φ is at least as pleasant as ψ" and $\psi \leq_i^{\mathsf{PIdl}} \varphi$ has to be read "according to agent i's pessimistic evaluation, φ is at least as ideal as ψ", Corresponding strict orderings are defined in the expected way as follows:

$$\psi <_i^{\mathsf{OPls}} \varphi \stackrel{\text{def}}{=} (\psi \leq_i^{\mathsf{OPls}} \varphi) \wedge \neg(\varphi \leq_i^{\mathsf{OPls}} \psi),$$

$$\psi <_i^{\mathsf{OIdl}} \varphi \stackrel{\text{def}}{=} (\psi \leq_i^{\mathsf{OIdl}} \varphi) \wedge \neg(\varphi \leq_i^{\mathsf{OIdl}} \psi),$$

$$\psi <_i^{\mathsf{PPls}} \varphi \stackrel{\text{def}}{=} (\psi \leq_i^{\mathsf{PPls}} \varphi) \wedge \neg(\varphi \leq_i^{\mathsf{PPls}} \psi) \text{ and}$$

$$\psi <_i^{\mathsf{PIdl}} \varphi \stackrel{\text{def}}{=} (\psi \leq_i^{\mathsf{PIdl}} \varphi) \wedge \neg(\varphi \leq_i^{\mathsf{PIdl}} \psi).$$

As the following Proposition 8 highlights, the comparative statements $\psi \leq_i^{\mathsf{OPls}} \varphi$ and $\psi \leq_i^{\mathsf{OIdl}} \varphi$ might also be read "for every epistemically possible ψ-state there is an epistemically possible φ-state which is at least as pleasant" and "for every epistemically possible ψ-state there is an epistemically possible φ-state which is at least as ideal", whereas the comparative statements $\psi \leq_i^{\mathsf{PPls}} \varphi$ and $\psi \leq_i^{\mathsf{PIdl}} \varphi$ might also be read "for every epistemically possible φ-state there is an epistemically possible ψ-state which is equally or less pleasant" and "for every epistemically possible φ-state there is an epistemically possible ψ-state which is equally or less ideal".[9]

[9] Other kinds of preference comparisons between formulas (*i.e.*, between sets of states) could be defined. For instance, following [24, 39], a $\forall\exists$-reading of preference statements

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

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

4.2. Moral sensitivity and preference based on desires and moral values

We extend the logic LAMA with special constructions of the form $\mathsf{moral}_{i,\mathsf{m}}$ which has to be read "agent i's degree of moral sensitivity is equal to m" with $\mathsf{m} \in Num$. We call LAMA^+ the resulting logic. A LAMA^+ model is a tuple $\langle M, \mathcal{S} \rangle$ where M is a LAMA model and $\mathcal{S}$ is a total function:

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

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:

(**Constr**) 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 $\mathsf{moral}_{i,\mathsf{m}}$ are interpreted by means of the function $\mathcal{S}$ as follows:

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

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

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

can 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 comparative statements between formulas is postponed to future work.

$$\mathsf{moral}_{i,\mathsf{m}} \rightarrow \neg \mathsf{moral}_{i,\mathsf{j}} \text{ if } \mathsf{m} \neq \mathsf{j} \qquad \textbf{(UniqueMoral)}$$

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

We use the degree of moral sensitivity as a parameter for calculating the utility of a given history for a certain agent. In particular, following the dual view of moral choice discussed in Section 2, we assume that pleasantness (as a measure of desire) and ideality (as a measure of moral value) are two different parameters determining the utility of a given history for a certain agent.

Definition 4 (Utility)**.** *Given a* LAMA$^+$ *model* $M = \langle W, \mathcal{H}, \equiv, \{\mathcal{E}_i\}_{i \in Agt}, \mathcal{P}, \mathcal{I}, \mathcal{V}, \mathcal{S} \rangle$*, the utility for agent i of the history starting in the world w, denoted by* $\mathcal{U}(w,i)$*, is defined as follows:*

$$\mathcal{U}(w,i) = \mathcal{S}(w,i) \times \mathcal{I}(w,i) + (\mathsf{maxVal} - \mathcal{S}(w,i)) \times \mathcal{P}(w,i)$$

Moreover, we define

$$\mathit{UScale} = \{\mathsf{y} : \exists \mathsf{k}_1, \mathsf{k}_2, \mathsf{k}_3 \in \mathit{Num} \text{ such that } \mathsf{y} = \mathsf{k}_1 \times \mathsf{k}_2 + (\mathsf{maxVal} - \mathsf{k}_1) \times \mathsf{k}_3\}$$

to be the agents' utility scale.[10]

According to Definition 4, the utility of a history is a function of both the degree of pleasantness and the degree of 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 the agent's moral sensitivity, the higher the influence of the degree of ideality on the utility of the history; the lower the agent's moral sensitivity, the higher the influence of the degree of pleasantness on the utility of the history. Note that the minimal value in the utility scale *UScale* is 0. The maximal value in *UScale* is denoted by max *UScale*.

The next step in the analysis is to define a concept of preference based on the preceding notion of utility. Following Broome [9], utility is here conceived as that which represents an agent's preference over a set of alternatives.

Definition 5 (Optimistic and pessimistic preferences)**.** *Given a* LAMA$^+$ *model* $M = \langle W, \mathcal{H}, \equiv, \{\mathcal{E}_i\}_{i \in Agt}, \mathcal{P}, \mathcal{I}, \mathcal{V}, \mathcal{S} \rangle$*, we say that at world w in M:*

- *agent i has an optimistic preference about* φ *with strength* y *(or agent i has a preference about* φ *with strength* y *based on an optimistic evaluation), i.e.,* $M, w \models \mathsf{OPref}_i^{\mathsf{y}} \varphi$*, if and only if* $\mathcal{U}^{\mathsf{opt}}_{w,i}(\varphi) = \mathsf{y}$*,*

[10] Note that *UScale* is finite because *Num* is finite.

- *agent i has a pessimistic preference about φ with strength* y *(or agent i has a preference about φ with strength* y *based on a pessimistic evaluation), i.e., $M,w \models \mathsf{PPref}_i^{\mathsf{y}}\varphi$, if and only if $\mathcal{U}_{w,i}^{\mathsf{pess}}(\varphi) = \mathsf{y}$,*

with:

$$\mathcal{U}_{w,i}^{\mathsf{opt}}(\varphi) = \max_{v \in \|\varphi\|_{w,i}} \mathcal{U}(v,i)$$

$$\mathcal{U}_{w,i}^{\mathsf{pess}}(\varphi) = \min_{v \in \|\varphi\|_{w,i}} \mathcal{U}(v,i)$$

and with the convention that $\max_{v \in \emptyset} \mathcal{U}(v,i) = 0$ *and* $\min_{v \in \emptyset} \mathcal{U}(v,i) = \max UScale$.

The concepts of optimistic preference and pessimistic preference semantically defined in Definition 5 are both syntactically expressible in the logic LAMA^{+}.

Proposition 9. *For all $i \in Agt$ and for all $\mathsf{y} \in UScale$ we have:*

- $M,w \models \mathsf{OPref}_i^{\mathsf{y}}\varphi$ *if and only if:*
  - *if* $\mathsf{y} > 0$ *then* $M,w \models \hat{\mathsf{K}}_i(\mathsf{util}_{i,\mathsf{y}} \wedge \varphi) \wedge \mathsf{K}_i((\bigvee_{\mathsf{z} \in UScale:\, \mathsf{y} < \mathsf{z}} \mathsf{util}_{i,\mathsf{z}}) \rightarrow \neg\varphi)$
  - *if* $\mathsf{y} = 0$ *then* $M,w \models \mathsf{K}_i((\bigvee_{\mathsf{z} \in UScale:\, 0 < \mathsf{z}} \mathsf{util}_{i,\mathsf{z}}) \rightarrow \neg\varphi)$
- $M,w \models \mathsf{PPref}_i^{\mathsf{y}}\varphi$ *if and only if:*
  - if $\mathsf{y} < \max$ *UScale then* $M,w \models \hat{\mathsf{K}}_i(\mathsf{util}_{i,\mathsf{y}} \wedge \varphi) \wedge \mathsf{K}_i((\bigvee_{\mathsf{z} \in UScale:\, \mathsf{z} < \mathsf{y}} \mathsf{util}_{i,\mathsf{z}}) \rightarrow \neg\varphi)$
  - if $\mathsf{y} = \max$ *UScale then* $M,w \models \mathsf{K}_i((\bigvee_{\mathsf{z} \in UScale:\, \mathsf{z} < \max UScale} \mathsf{util}_{i,\mathsf{z}}) \rightarrow \neg\varphi)$

where for all $\mathsf{y} \in UScale$:

$$\mathsf{util}_{i,\mathsf{y}} \stackrel{\mathsf{def}}{=} \bigvee_{\mathsf{k},\mathsf{g},\mathsf{m} \in Num:\, \mathsf{y} = \mathsf{m} \times \mathsf{g} + (\mathsf{maxVal} - \mathsf{m}) \times \mathsf{k}} (\mathsf{moral}_{i,\mathsf{m}} \wedge \mathsf{pls}_{i,\mathsf{k}} \wedge \mathsf{idl}_{i,\mathsf{g}})$$

and with the convention that a disjunction over an empty set is false (e.g., $\bigvee_{\mathsf{z} \in UScale:\, \mathsf{z} < 0} \mathsf{util}_{i,\mathsf{z}} = \bot$).

Like the optimistic desire operator and the optimistic ideal operator, the optimistic preference operator is an operator of weak possibility (or potential possibility) in the sense of possibility theory. On the contrary, like the pessimistic desire operator and the pessimistic ideal operator, the pessimistic preference operator is an operator of strong possibility (or actual possibility). Their main properties are listed in the following proposition.

Proposition 10. *For every $i \in Agt$ we have:*

$$\models (\mathsf{OPref}_i^{\mathsf{y}}\varphi \wedge \mathsf{OPref}_i^{\mathsf{z}}\psi) \rightarrow \mathsf{OPref}_i^{\max\{\mathsf{y},\mathsf{z}\}}(\varphi \vee \psi) \tag{10}$$

$$\models (\mathsf{PPref}_i^{\mathsf{y}}\varphi \wedge \mathsf{PPref}_i^{\mathsf{z}}\psi) \rightarrow \mathsf{PPref}_i^{\min\{\mathsf{y},\mathsf{z}\}}(\varphi \vee \psi) \tag{11}$$

$$\models (\mathsf{OPref}_i^{\mathsf{y}}\varphi \wedge \mathsf{OPref}_i^{\mathsf{z}}\psi) \rightarrow \mathsf{OPref}_i^{\leq \min\{\mathsf{y},\mathsf{z}\}}(\varphi \vee \psi) \tag{12}$$

$$\models (\mathsf{PPref}_i^{\mathsf{y}}\varphi \wedge \mathsf{PPref}_i^{\mathsf{z}}\psi) \rightarrow \mathsf{PPref}_i^{\geq \max\{\mathsf{y},\mathsf{z}\}}(\varphi \vee \psi) \tag{13}$$

where for $X \in \{\mathsf{OPref}, \mathsf{PPref}\}$*:*

$$X_i^{\geq \mathsf{y}}\varphi \stackrel{\mathsf{def}}{=} \bigvee_{\mathsf{z} \in UScale: \mathsf{z} \geq \mathsf{y}} X_i^{\mathsf{z}}\varphi$$

and

$$X_i^{\leq \mathsf{y}}\varphi \stackrel{\mathsf{def}}{=} \bigvee_{\mathsf{z} \in UScale: \mathsf{z} \leq \mathsf{y}} X_i^{\mathsf{z}}\varphi$$

As we have done for pleasantness and ideality in Section 4.1, we define two types of dyadic operators for comparison of utility, one for comparison based on an optimistic evaluation of utility and the other for comparison based on a pessimistic evaluation of utility.

$$\psi \leq_i^{\mathsf{OUtil}} \varphi \stackrel{\mathsf{def}}{=} \bigvee_{\mathsf{y} \in UScale} (\mathsf{OPref}_i^{\mathsf{y}}\varphi \wedge \bigwedge_{\mathsf{z} \in UScale: \mathsf{z} > \mathsf{y}} \neg \mathsf{OPref}_i^{\mathsf{z}}\psi)$$

$$\psi \leq_i^{\mathsf{PUtil}} \varphi \stackrel{\mathsf{def}}{=} \bigvee_{\mathsf{y} \in UScale} (\mathsf{PPref}_i^{\mathsf{y}}\varphi \wedge \bigwedge_{\mathsf{z} \in UScale: \mathsf{z} > \mathsf{y}} \neg \mathsf{PPref}_i^{\mathsf{z}}\psi)$$

$\psi \leq_i^{\mathsf{OUtil}} \varphi$ has to be read "according to agent i's optimistic evaluation, φ is at least as preferable as ψ", while $\psi \leq_i^{\mathsf{PUtil}} \varphi$ has to be read "according to agent i's pessimistic evaluation, φ is at least as preferable as ψ". Corresponding strict orderings are defined in the expected way as follows:

$$\psi <_i^{\mathsf{OUtil}} \varphi \stackrel{\mathsf{def}}{=} (\psi \leq_i^{\mathsf{OUtil}} \varphi) \wedge \neg(\varphi \leq_i^{\mathsf{OUtil}} \psi) \text{ and}$$

$$\psi <_i^{\mathsf{PUtil}} \varphi \stackrel{\mathsf{def}}{=} (\psi \leq_i^{\mathsf{PUtil}} \varphi) \wedge \neg(\varphi \leq_i^{\mathsf{PUtil}} \psi).$$

As the following Proposition 11 highlights, the comparative statements $\psi \leq_i^{\mathsf{OUtil}} \varphi$ and $\psi \leq_i^{\mathsf{PUtil}} \varphi$ might also be read "for every epistemically possible ψ-state there is an epistemically possible φ-state which is at least as useful" and "for every epistemically possible φ-state there is an epistemically possible ψ-state which is equally or less useful".

Proposition 11. *For every* $i \in Agt$ *we have:*

- $M, w \models \psi \leq_i^{\mathsf{OUtil}} \varphi$ *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* $\mathcal{U}(v,i) \leq \mathcal{U}(u,i)$ *and* $M, u \models \varphi$*,*

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

The following Proposition 12 captures some basic logical relationships among the concepts of preference, desire and moral value.

Proposition 12. *For every* $i \in Agt$ *we have:*

$$\models ((\psi <_i^{\mathsf{OPls}} \varphi) \wedge \mathsf{moral}_{i,0}) \rightarrow (\psi <_i^{\mathsf{OUtil}} \varphi) \tag{12}$$

$$\models ((\psi <_i^{\mathsf{OIdl}} \varphi) \wedge \mathsf{moral}_{i,\mathsf{maxVal}}) \rightarrow (\psi <_i^{\mathsf{OUtil}} \varphi) \tag{13}$$

$$\models ((\psi <_i^{\mathsf{PPls}} \varphi) \wedge \mathsf{moral}_{i,0}) \rightarrow (\psi <_i^{\mathsf{PUtil}} \varphi) \tag{14}$$

$$\models ((\psi <_i^{\mathsf{PIdl}} \varphi) \wedge \mathsf{moral}_{i,\mathsf{maxVal}}) \rightarrow (\psi <_i^{\mathsf{PUtil}} \varphi) \tag{15}$$

According to the preceding validities, if an agent has a minimal degree of moral sensitivity then his preferences are fully determined by his desires, whereas if an agent has a maximal degree of moral sensitivity then his preferences are fully determined by his moral values. The case in which the agent's desires/moral values are based on an optimistic evaluation is distinguished from the case in which his desires/moral values are based on a pessimistic evaluation.

The following Proposition 13 is a generalization of the preceding Proposition 12, as it considers not only the extreme cases of minimal and maximal moral sensitivity (*i.e.*, $\mathsf{moral}_{i,0}$ and $\mathsf{moral}_{i,\mathsf{maxVal}}$), but also the intermediate cases.

Proposition 13. *For every* $i \in Agt$ *we have:*

If for all $\mathsf{j} \in Num$ *we have* $\mathsf{m} \times \mathsf{k} > \mathsf{m} \times \mathsf{g} + (\mathsf{maxVal} - \mathsf{m}) \times \mathsf{j}$ *then:*

$$\models (\mathsf{OVal}_i^{\mathsf{k}} \varphi \wedge \mathsf{OVal}_i^{\mathsf{g}} \psi \wedge \mathsf{moral}_{i,\mathsf{m}}) \rightarrow (\psi <_i^{\mathsf{OUtil}} \varphi) \tag{16}$$

If for all $\mathsf{j} \in Num$ *we have* $(\mathsf{maxVal} - \mathsf{m}) \times \mathsf{k} > \mathsf{m} \times \mathsf{j} + (\mathsf{maxVal} - \mathsf{m}) \times \mathsf{g}$ *then:*

$$\models (\mathsf{ODes}_i^{\mathsf{k}} \varphi \wedge \mathsf{ODes}_i^{\mathsf{g}} \psi \wedge \mathsf{moral}_{i,\mathsf{m}}) \rightarrow (\psi <_i^{\mathsf{OUtil}} \varphi) \tag{17}$$

If for all $\mathsf{j} \in Num$ *we have* $\mathsf{m} \times \mathsf{k} > \mathsf{m} \times \mathsf{g} + (\mathsf{maxVal} - \mathsf{m}) \times \mathsf{j}$ *then:*

$$\models (\mathsf{PVal}_i^{\mathsf{k}} \varphi \wedge \mathsf{PVal}_i^{\mathsf{g}} \psi \wedge \mathsf{moral}_{i,\mathsf{m}}) \rightarrow (\psi <_i^{\mathsf{PUtil}} \varphi) \tag{18}$$

If for all $\mathsf{j} \in Num$ *we have* $(\mathsf{maxVal} - \mathsf{m}) \times \mathsf{k} > \mathsf{m} \times \mathsf{j} + (\mathsf{maxVal} - \mathsf{m}) \times \mathsf{g}$ *then:*

$$\models (\mathsf{PDes}_i^{\mathsf{k}} \varphi \wedge \mathsf{PDes}_i^{\mathsf{g}} \psi \wedge \mathsf{moral}_{i,\mathsf{m}}) \rightarrow (\psi <_i^{\mathsf{PUtil}} \varphi) \tag{19}$$

By way of example, suppose that $Num = \{0,\ldots,10\}$ and $M,w \models \mathsf{OVal}_i^8\varphi \wedge \mathsf{OVal}_i^2\psi \wedge \mathsf{moral}_{i,7}$. Then, according to validity (16), $M,w \models (\psi <_i^{\mathsf{OUtil}} \varphi)$ because $7\times 8 > 7\times 2 + (10-7)\times \mathsf{j}$ for all $\mathsf{j} \in Num$.

4.3. *Relationship between preferences and choices via the concept of rationality*

The last aspect of moral agency we consider is the relationship between preference and choice. To this aim we introduce two notions of rationality: optimistic rationality (or rationality based on an optimistic evaluation of utility) and pessimistic rationality (or rationality based on a pessimistic evaluation of utility).

We say that a given agent *i* is rational in the optimistic sense (or agent *i* is an optimistic rational agent), denoted by ORat_i, if and only if, for every action *a*, if he decides to do *a* then, according to his optimistic evaluation of utility, playing action *a* is at least as preferable as not playing action *a*:

$$\mathsf{ORat}_i \stackrel{\mathsf{def}}{=} \bigwedge_{a\in Act} (\mathsf{choose}_{i,a} \rightarrow (\neg \mathsf{choose}_{i,a} \leq_i^{\mathsf{OUtil}} \mathsf{choose}_{i,a}))$$

We say that a given agent *i* is rational in the pessimistic sense (or agent *i* is a pessimistic rational agent), denoted by PRat_i, if and only if, for every action *a*, if he decides to do *a* then, according to his pessimistic evaluation of utility, playing action *a* is at least as preferable as not playing action *a*:

$$\mathsf{PRat}_i \stackrel{\mathsf{def}}{=} \bigwedge_{a\in Act} (\mathsf{choose}_{i,a} \rightarrow (\neg \mathsf{choose}_{i,a} \leq_i^{\mathsf{PUtil}} \mathsf{choose}_{i,a}))$$

As the following Proposition 14 highlights, the preceding notions of optimistic and pessimistic rationality correspond to two well-known decision criteria in the theory of decision-making under ignorance (see, *e.g.*, [31]): the 'maximax' criterion and the 'maximin' criterion. Specifically, a given agent is rational in the optimistic sense if and only if he chooses an action whose best outcome is at least as good as the best outcome of all other courses of action; a given agent is rational in the pessimistic sense if and only if he chooses an action whose worst outcome is at least as good as the least outcome of all other courses of action.

Proposition 14. *For all $i \in Agt$ and for every LAMA$^+$ model M and world w in M, we have:*

- $M,w \models \mathsf{ORat}_i$ *if and only if there is* $a \in \operatorname{argmax}_{a\in Rep(i)} \max_{v\in \|\mathsf{choose}_{i,a}\|_{w,i}} \mathcal{U}(v,i)$ *such that* $M,w \models \mathsf{choose}_{i,a}$

- $M, w \models \mathsf{PRat}_i$ *if and only if there is* $a \in \underset{a \in Rep(i)}{\operatorname{argmax}} \min_{v \in \|\mathsf{choose}_{i,a}\|_{w,i}} \mathcal{U}(v,i)$ *such that* $M, w \models \mathsf{choose}_{i,a}$

Note that in the preceding proposition $\underset{a \in Rep(i)}{\operatorname{argmax}} \max_{v \in \|\mathsf{choose}_{i,a}\|_{w,i}} \mathcal{U}(v,i)$ denotes the set of all actions a in $Rep(i)$ such that $\max_{v \in \|\mathsf{choose}_{i,a}\|_{w,i}} \mathcal{U}(v,i) \geq \max_{v \in \|\mathsf{choose}_{i,b}\|_{w,i}} \mathcal{U}(v,i)$ for all $b \in Rep(i)$, whereas $\underset{a \in Rep(i)}{\operatorname{argmax}} \min_{v \in \|\mathsf{choose}_{i,a}\|_{w,i}} \mathcal{U}(v,i)$ denotes the set of all actions a in $Rep(i)$ such that $\min_{v \in \|\mathsf{choose}_{i,a}\|_{w,i}} \mathcal{U}(v,i) \geq \min_{v \in \|\mathsf{choose}_{i,b}\|_{w,i}} \mathcal{U}(v,i)$ for all $b \in Rep(i)$.

Moreover, note that we have $a \in \underset{a \in Rep(i)}{\operatorname{argmax}} \ldots$ instead of $\{a\} = \underset{a \in Rep(i)}{\operatorname{argmax}} \ldots$ because in a certain situation there could be *more than one* rational choice for an agent.

The following Proposition 15, which follows from Proposition 12, explains how desires and moral values motivate a rational agent to perform a given action. We consider four different cases: (1) an optimistic rational agent with a minimal degree of moral sensitivity; (2) an optimistic rational agent with a maximal degree of moral sensitivity; (3) a pessimistic rational agent with a minimal degree of moral sensitivity; (4) an pessimistic rational agent with a maximal degree of moral sensitivity.

Proposition 15. *For every* $i \in Agt$ *we have:*

$$\models ((\neg \mathsf{choose}_{i,a} <_i^{\mathsf{OPls}} \mathsf{choose}_{i,a}) \wedge \mathsf{moral}_{i,0} \wedge \mathsf{ORat}_i) \rightarrow \mathsf{choose}_{i,a} \quad (15)$$

$$\models ((\neg \mathsf{choose}_{i,a} <_i^{\mathsf{OIdl}} \mathsf{choose}_{i,a}) \wedge \mathsf{moral}_{i,\mathsf{maxVal}} \wedge \mathsf{ORat}_i) \rightarrow \mathsf{choose}_{i,a} \quad (16)$$

$$\models ((\neg \mathsf{choose}_{i,a} <_i^{\mathsf{PPls}} \mathsf{choose}_{i,a}) \wedge \mathsf{moral}_{i,0} \wedge \mathsf{PRat}_i) \rightarrow \mathsf{choose}_{i,a} \quad (17)$$

$$\models ((\neg \mathsf{choose}_{i,a} <_i^{\mathsf{PIdl}} \mathsf{choose}_{i,a}) \wedge \mathsf{moral}_{i,\mathsf{maxVal}} \wedge \mathsf{PRat}_i) \rightarrow \mathsf{choose}_{i,a} \quad (18)$$

According to the preceding validities, if a rational agent has a minimal degree of moral sensitivity then his current choice is fully determined by his desires, whereas if a rational agent has a maximal degree of moral sensitivity then his current choice is fully determined by his moral values.

The following Proposition 16, which follows from Proposition 13, provides a generalization of Proposition 15 to intermediate cases of moral sensitivity.

Proposition 16. *For every* $i \in Agt$ *we have:*

If $\mathsf{m} \times \mathsf{k} > \mathsf{m} \times \mathsf{g} + (\mathsf{maxVal} - \mathsf{m}) \times \mathsf{j}$ *for all* $\mathsf{j} \in Num$ *then:*

$$\models (\mathsf{OVal}_i^{\mathsf{k}} \mathsf{choose}_{i,a} \wedge \mathsf{OVal}_i^{\mathsf{g}} \neg \mathsf{choose}_{i,a} \wedge \mathsf{moral}_{i,\mathsf{m}} \wedge \mathsf{ORat}_i) \rightarrow \mathsf{choose}_{i,a} \quad (19)$$

If $(\mathsf{maxVal}-\mathsf{m})\times\mathsf{k}>\mathsf{m}\times\mathsf{j}+(\mathsf{maxVal}-\mathsf{m})\times\mathsf{g}$ *for all* $\mathsf{j}\in \mathit{Num}$ *then:*

$$\models (\mathsf{ODes}_i^{\mathsf{k}}\mathsf{choose}_{i,a}\wedge\mathsf{ODes}_i^{\mathsf{g}}\neg\mathsf{choose}_{i,a}\wedge\mathsf{moral}_{i,\mathsf{m}}\wedge\mathsf{ORat}_i)\rightarrow\mathsf{choose}_{i,a} \quad (20)$$

If $\mathsf{m}\times\mathsf{k}>\mathsf{m}\times\mathsf{g}+(\mathsf{maxVal}-\mathsf{m})\times\mathsf{j}$ *for all* $\mathsf{j}\in \mathit{Num}$ *then:*

$$\models (\mathsf{PVal}_i^{\mathsf{k}}\mathsf{choose}_{i,a}\wedge\mathsf{PVal}_i^{\mathsf{g}}\neg\mathsf{choose}_{i,a}\wedge\mathsf{moral}_{i,\mathsf{m}}\wedge\mathsf{PRat}_i)\rightarrow\mathsf{choose}_{i,a} \quad (21)$$

If $(\mathsf{maxVal}-\mathsf{m})\times\mathsf{k}>\mathsf{m}\times\mathsf{j}+(\mathsf{maxVal}-\mathsf{m})\times\mathsf{g}$ *for all* $\mathsf{j}\in \mathit{Num}$ *then:*

$$\models (\mathsf{PDes}_i^{\mathsf{k}}\mathsf{choose}_{i,a}\wedge\mathsf{PDes}_i^{\mathsf{g}}\neg\mathsf{choose}_{i,a}\wedge\mathsf{moral}_{i,\mathsf{m}}\wedge\mathsf{PRat}_i)\rightarrow\mathsf{choose}_{i,a} \quad (22)$$

4.4. *An example: the Prisoner's Dilemma*

Before concluding, we give an example, inspired by the famous Prisoner's Dilemma, that illustrates two interesting aspects of the logic LAMA^+, namely (1) the possibility of representing multi-agent scenarios and (2) the possibility of reasoning about the effects of the joint action of a coalition thanks to the dynamic operators $[\![\delta]\!]$ and $\langle\!\langle\delta\rangle\!\rangle$. These two features of the logic LAMA^+ turn out to be extremely useful for representing situations of strategic interaction, which are studied in game theory, and for describing how the world evolves over time.

Bonnie (B) and Clyde (C) are the members of a criminal gang who are arrested. Since the police does not have enough evidence to convict them, the prosecutor offers each prisoner the following bargain. If one of them testifies against his/her partner and the other does not, the former will go free while the latter will go to jail. If both prisoners testify against each other, both will be recommended for house arrest. Finally, if both prisoners remain silent, both will be obliged to pay a fine for firearm possession in order to be released.

In order to represent the preceding scenario in the logic LAMA^+ let us assume that:

- $Agt=\{C,B\}$,
- $Rep(C)=Rep(B)=\{silent, testify\}$,
- $Atm=\bigcup_{i\in\{C,B\}}\{free_i, fine_i, harr_i, pris_i\}$,

where *silent* is the action of remaining silent, *testify* is the action of testifying against the other prisoner, $free_i$ is a propositional atom denoting the fact that prisoner i goes free, $fine_i$ is a propositional atom denoting the fact that prisoner i is obliged to pay a fine for firearm possession in order to be released, $harr_i$ is a propositional atom denoting the fact that prisoner i is recommended for house arrest and $pris_i$ is a propositional atom denoting the fact that prisoner i is sentenced to prison.

The set of joint actions of the two prisoners is $JAct = \{\delta_1, \delta_2, \delta_3, \delta_4\}$ with

$$\delta_1 = \{B \mapsto testify, C \mapsto testify\}$$
$$\delta_2 = \{B \mapsto testify, C \mapsto silent\}$$
$$\delta_3 = \{B \mapsto silent, C \mapsto testify\}$$
$$\delta_4 = \{B \mapsto silent, C \mapsto silent\}$$

For example, δ_1 is the joint action which consists of Bonnie's individual action of testifying against Clyde and Clyde's individual action of testifying against Bonnie.

The following Table 1 collects the positive effect preconditions of the two actions *silent* and *testify* that can be performed by Bonnie and Clyde with respect to the each atomic proposition in the set *Atm*. For example, according to the function γ^+, Bonnie will get free by testifying against Clyde only if Clyde remains silent (*i.e.*, $\gamma^+(B, testify, free_B) = \mathsf{choose}_{C, silent}$) and Clyde will be sentenced to prison by remaining silent only if Bonnie testifies against him (*i.e.*, $\gamma^+(C, silent, pris_C) = \mathsf{choose}_{B, testify}$).

$\gamma^+(i, testify, free_i) = \mathsf{choose}_{j, silent}$	$\gamma^+(i, silent, free_i) = \bot$
$\gamma^+(i, testify, fine_i) = \bot$	$\gamma^+(i, silent, fine_i) = \mathsf{choose}_{j, silent}$
$\gamma^+(i, testify, harr_i) = \mathsf{choose}_{j, testify}$	$\gamma^+(i, silent, harr_i) = \bot$
$\gamma^+(i, testify, pris_i) = \bot$	$\gamma^+(i, silent, pris_i) = \mathsf{choose}_{j, testify}$
$\gamma^+(i, testify, free_j) = \bot$	$\gamma^+(i, silent, free_j) = \mathsf{choose}_{j, testify}$
$\gamma^+(i, testify, fine_j) = \bot$	$\gamma^+(i, silent, fine_j) = \mathsf{choose}_{j, silent}$
$\gamma^+(i, testify, harr_j) = \mathsf{choose}_{j, testify}$	$\gamma^+(i, silent, harr_j) = \bot$
$\gamma^+(i, testify, pris_j) = \mathsf{choose}_{j, silent}$	$\gamma^+(i, silent, pris_j) = \bot$

Table 1: Positive effect preconditions with i, j ranging over $Agt = \{B, C\}$ and $i \neq j$.

As for the negative effect preconditions, we assume that for all $i \in Agt$, $a \in Act$ and $p \in Atm$ we have:

$$\gamma^-(i, a, p) = \bot$$

In other words, Bonnie and Clyde cannot make an atom in *Atm* false by performing an action in *Act*. This is a reasonable assumption as the actions *testify* and *silent* can only change the atoms in *Atm* from false to true.

Let us abbreviate

$$HypKnow \stackrel{\text{def}}{=} \bigwedge_{i \in Agt} ((\mathsf{K}_i \bigwedge_{p \in Atm} \neg p) \wedge (\bigwedge_{\delta \in JAct} \hat{\mathsf{K}}_i \, \mathsf{occ}_\delta))$$

and

$$HypDesVal \stackrel{\text{def}}{=} \bigwedge_{i,j \in Agt : i \neq j} ((\mathsf{X}\, pris_i <_i^{\mathsf{PPls}} \mathsf{X} \neg pris_i) \wedge (\mathsf{X}(free_i \wedge \neg free_j) <_i^{\mathsf{Pldl}} \mathsf{X} \neg free_i))$$

where X is the temporal operator *next* defined in Section 3.1.

HypKnow is the hypothesis about the prisoners' epistemic states. Specifically, we assume that each prisoner knows that all atoms in *Atm* are currently false. Moreover, we assume that for every joint action δ in *JACT* both prisoners envisage a world in which δ is going to occur.

HypDesVal is the hypothesis about the prisoners' desires and moral values. Specifically, we assume that each prisoner considers the situation in which he/she will not go to prison strictly more pleasant than the situation in which he/she will go to prison, and the situation in which he/she will not get free strictly more ideal than the situation in which he/she will get free while the other will not. Indeed, we assume that each prisoner considers morally deplorable a situation in which he/she gets free to the other's detriment.

The following four validities highlight what the two prisoners are expected to do and which consequences their joint action will have, under the previous assumptions *HypKnow* and *HypDesVal* and under certain assumptions about their rationality and their moral sensitivity. For example, according to the validity (23), under the assumptions *HypKnow* and *HypDesVal*, if both Bonnie and Clyde are pessimistic rational agents and have a minimal degree of moral sensitivity then each of them will decide to testify against the other and both will be recommended for house arrest in the resulting state. According to the validity (24), under the same assumptions, if both Bonnie and Clyde are pessimistic rational agents and have a maximal degree of moral sensitivity then each of them will decide to remain silent and both will be obliged to pay a fine for firearm possession in the resulting state in order to be released.[11] The other two validities (25) and (26) capture the complementary cases.

$$\models (HypKnow \wedge HypDesVal \wedge \mathsf{moral}_{B,0} \wedge \mathsf{moral}_{C,0} \wedge \mathsf{PRat}_B \wedge \mathsf{PRat}_C) \rightarrow \langle\langle \{B \mapsto testify, C \mapsto testify\} \rangle\rangle (harr_B \wedge harr_C) \quad (23)$$

[11] This result can be interpreted as saying that in the Prisoner's Dilemma mutual defection is an equilibrium for self-interested agents while mutual cooperation is an equilibrium for moral agents.

$$\models (\mathit{HypKnow} \wedge \mathit{HypDesVal} \wedge \mathsf{moral}_{B,\mathsf{maxVal}} \wedge \mathsf{moral}_{C,\mathsf{maxVal}} \wedge \mathsf{PRat}_B \wedge \mathsf{PRat}_C) \rightarrow \langle\langle \{B \mapsto \mathit{silent}, C \mapsto \mathit{silent}\} \rangle\rangle (\mathit{fine}_B \wedge \mathit{fine}_C) \quad (24)$$

$$\models (\mathit{HypKnow} \wedge \mathit{HypDesVal} \wedge \mathsf{moral}_{B,0} \wedge \mathsf{moral}_{C,\mathsf{maxVal}} \wedge \mathsf{PRat}_B \wedge \mathsf{PRat}_C) \rightarrow \langle\langle \{B \mapsto \mathit{testify}, C \mapsto \mathit{silent}\} \rangle\rangle (\mathit{free}_B \wedge \mathit{pris}_C) \quad (25)$$

$$\models (\mathit{HypKnow} \wedge \mathit{HypDesVal} \wedge \mathsf{moral}_{B,\mathsf{maxVal}} \wedge \mathsf{moral}_{C,0} \wedge \mathsf{PRat}_B \wedge \mathsf{PRat}_C) \rightarrow \langle\langle \{B \mapsto \mathit{silent}, C \mapsto \mathit{testify}\} \rangle\rangle (\mathit{pris}_B \wedge \mathit{free}_C) \quad (26)$$

5. Conclusion

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

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 [18]. 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 for the logical part, we intend to provide a dynamic extension of the logic LAMA in which an agent's desires, moral values and degree of moral sensitivity might change. Indeed, in the current version of the logic LAMA, these three dimensions of an agent's psychological state are assumed to be static.

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.

Emiliano LORINI
Université de Toulouse
IRIT-CNRS, France

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] R.J. Aumann. St. petersburg paradox: A discussion of some recent comments. *Journal of Economic Theory*, 14:443–445, 1977.

[4] R. Aumann. Interactive epistemology I: Knowledge. *International Journal of Game Theory*, 28(3):263–300, 1999.

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

[6] K. Binmore. *Natural justice*. Oxford University Press, New York, 2005.

[7] M. Bratman. *Intentions, plans, and practical reason*. Harvard University Press, Cambridge, 1987.

[8] 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.

[9] J. Broome. Utility. *Economics and Philosophy*, 7:1–12, 1991.

[10] A. Casali, L. Godo, and C. Sierra. A graded BDI agent model to represent and reason about preferences. *Artificial Intelligence*, 175:1468–1478, 2012.

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

[12] D. Dubois and H. Prade. Possibility theory: qualitative and quantitative aspects. In D. Gabbay and P. Smets, editors, *Handbook of Defeasible Reasoning and Uncertainty Management Systems*, volume Quantified Representation of Uncertainty and Imprecision, volume 1, pages 169–226. Kluwer, 1998.

[13] D. Dubois and H. Prade. From Blanchés hexagonal organization of concepts to formal concept analysis and possibility theory. *Logica Universalis*, 6(1): 149–169, 2012.

[14] R. Fagin, J. Halpern, Y. Moses, and M. Vardi. *Reasoning about Knowledge*. MIT Press, Cambridge, 1995.

[15] 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.

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

[17] 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.

[18] 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.

[19] D. Harel, D. Kozen, and J. Tiuryn. *Dynamic Logic*. MIT Press, Cambridge, 2000.

[20] G. H. Harman. *Explaining value and other essays in moral philosophy*. Clarendon Press, Oxford, 2000.

[21] J. C. Harsanyi. Cardinal welfare, individualistic ethics, and interpersonal comparisons of utility. *Journal of Political Economy*, 63:309–321, 1955.

[22] 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.
[23] A. Herzig and D. Longin. C&L intention revisited. In *Proceeding of the 9th International Conferencene on Principles of Knowledge Representation and Reasoning (KR2004)*, pages 527–535. AAAI Press, 2004.
[24] F. Liu. *Reasoning about Preference Dynamics*, volume 354 of *Synthese Library*. Springer-Verlag, 2011.
[25] E. Lorini and A. Herzig. A logic of intention and attempt. *Synthese*, 163(1): 45–77, 2008.
[26] E. Lorini and H. Prade. Strong possibility and weak necessity as a basis for a logic of desire. In *Proceeding of the First International Workshop on Weighted Logics for AI (WL4AI)*, pages 99–103, 2012.
[27] R. D. Luce and H. Raiffa. *Games and decisions: introduction and critical survey*. Wiley, 1957.
[28] H. Margolis. *Selfishness, Altruism, and Rationality: A Theory of Social Choice*. University of Chicago Press, Chicago, 1982.
[29] M. Mas-Colell, A. Winston and J. Green. *Microeconomic Theory*. Oxford University Press, 1995.
[30] J. J. Ch. Meyer, W. van der Hoek, and B. van Linder. A logical approach to the dynamics of commitments. *Artificial Intelligence*, 113(1-2):1–40, 1999.
[31] M. Peterson. *An Introduction to Decision Theory*. Cambridge University Press, 2009.
[32] A. S. Rao and M. Georgeff. Decision procedures for BDI logics. *Journal of Logic and Computation*, 8(3):293–344, 1998.
[33] R. Reiter. *Knowledge in action: logical foundations for specifying and implementing dynamical systems*. MIT Press, Cambridge, 2001.
[34] 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.
[35] J. Searle. *Rationality in Action*. MIT Press, Cambridge, 2001.
[36] A. K. Sen. Choice, orderings and morality. In S. Korner, editor, *Practical reason*. Blackwell, Oxford, 1974.
[37] A. K. Sen. Rational fools: a critique of the behavioral foundations of economic theory. *Philosophy and public affairs*, 6:317–344, 1977.
[38] Y. Shoham. Agent-oriented programming. *Artificial Intelligence*, 60:51–92, 1993.
[39] 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.
[40] W. van der Hoek and M. Wooldridge. Towards a logic of rational agency. *Logic Journal of the IGPL*, 11:133–157, 2003.
[41] M. Wooldridge. *Reasoning about rational agents*. MIT Press, Cambridge, 2000.

APPENDIX

A. Selected proofs

A.1. *Proof of Proposition 2*

In order to prove the validities (2)-(8) it is sufficient to note that for any LAMA model M and world w in M we have $M,w \vDash \mathsf{K}_i^* \varphi$ if and only if $M,v \vDash \varphi$ for all v such that $w\mathcal{E}_i^* v$ with $\mathcal{E}_i^* = \{(w,v) \in \mathcal{E}_i : \mathcal{C}(w,i) = \mathcal{C}(v,i)\}$. The fact that K_i^* can be seen as a modal box operator for an accessibility relation $\mathcal{E}_i^*$ justifies (2) and (3). The fact that $\mathcal{E}_i^*$ is an equivalence relation justifies (4), (5) and (6). The fact that $\mathcal{E}_i^* \subseteq \mathcal{E}_i$ justifies (8).

Let us prove validity (7). Suppose $M,w \vDash \mathsf{choose}_{i,a}$. By the semantics of $\mathsf{choose}_{i,a}$, it follows that $M,w \vDash \neg \mathsf{choose}_{i,b}$ for all $b \in Rep_i$ such that $b \neq a$. Moreover, we have $M,w \vDash \mathsf{K}_i(\mathsf{choose}_{i,a} \rightarrow \mathsf{choose}_{i,a})$. Therefore, $M,w \vDash \bigwedge_{c \in Rep_i}(\neg \mathsf{choose}_{i,c} \vee \mathsf{K}_i(\mathsf{choose}_{i,c} \rightarrow \mathsf{choose}_{i,a}))$. Hence, $M, w \vDash \bigwedge_{c \in Rep_i}(\mathsf{choose}_{i,c} \rightarrow \mathsf{K}_i(\mathsf{choose}_{i,c} \rightarrow \mathsf{choose}_{i,a}))$. Hence, by definition of the *interim* knowledge operator K_i^*, $M,w \vDash \mathsf{K}_i^* \mathsf{choose}_{i,a}$.

A.2. *Proof of Proposition 4*

We prove the last validity as an example.

$M,w \vDash [\![\delta]\!]\mathsf{K}_i\varphi$ IFF,

if $M,w \vDash \mathsf{occ}_{(\delta)}$ then $M^{\delta},w \vDash \mathsf{K}_i\varphi$ IFF,

if $M,w \vDash \mathsf{occ}_{(\delta)}$ then, for all ν, if $w\mathcal{E}_i^{\delta} v$ then $M^{\delta},v \vDash \varphi$ IFF,

if $M,w \vDash \mathsf{occ}_{(\delta)}$ then, for all ν, if $w\mathcal{E}_i v$ and $M,v \vDash \mathsf{occ}_{(\delta)}$ then $M^{\delta},v \vDash \varphi$ IFF,

if $M,w \vDash \mathsf{occ}_{(\delta)}$ then, for all ν, if $w\mathcal{E}_i v$ then (if $M,v \vDash \mathsf{occ}_{(\delta)}$ then $M^{\delta},v \vDash \varphi$) IFF,

if $M,w \vDash \mathsf{occ}_{(\delta)}$ then, for all ν, if $w\mathcal{E}_i v$ then $M,v \vDash [\![\delta]\!]\varphi$ IFF,

if $M,w \vDash \mathsf{occ}_{(\delta)}$ then $M,w \vDash \mathsf{K}_i[\![\delta]\!]\varphi$ IFF,

if $M,w \vDash \mathsf{occ}_{(\delta)} \rightarrow \mathsf{K}_i[\![\delta]\!]\varphi$.

A.3. *Sketch of proof of Theorem 1*

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 every epistemic operator K_i together with the principles in Proposition 3 and all principles of classical propositional logic provide a complete axiomatization for LAMA^{-}. Now, suppose φ is LAMA valid. Then $red(\varphi)$ is valid

in LAMA⁻ due to item 2 of Proposition 5. By the completeness of LAMA⁻, *red*(φ) is also provable there. LAMA being a conservative extension of LAMA⁻, *red*(φ) is provable in LAMA, 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 LAMA.

A.4. *Proof of Proposition 6*

We only prove the first item and the third item as the other two items can be proved in a similar way. Let us remind that $\|\varphi\|_{w,i} = \{v \in W : M, v \vDash \varphi$ and $v \in \mathcal{E}_i(w)\}$.

As for the first item, we distinguish two cases: $\mathsf{k} > 0$ and $\mathsf{k} = 0$.

Let us assume that $\mathsf{k} = 0$. Then:

$M, w \vDash \mathsf{ODes}_i^0 \varphi$ IFF,

$\max\limits_{v \in \|\varphi\|_{w,i}} \mathcal{P}(v,i) = 0$ IFF,

(there is $v \in \|\varphi\|_{w,i}$ such that $\mathcal{P}(v,i) = 0$ and

for all $u \in \mathcal{E}_i(w)$, if $\mathcal{P}(u,i) > 0$ then $M, u \vDash \neg\varphi$) or ($\|\varphi\|_{w,i} = \emptyset$) IFF,

for all $u \in \mathcal{E}_i(w)$, if $\mathcal{P}(u,i) > 0$ then $M, u \vDash \neg\varphi$ IFF,

$M, w \vDash \bigwedge_{\mathsf{g} \in Num : 0 < \mathsf{g}} \mathsf{K}_i(\mathsf{pls}_{i,\mathsf{g}} \rightarrow \neg\varphi)$ IFF,

$M, w \vDash \mathsf{K}_i((\bigvee_{\mathsf{g} \in Num : 0 < \mathsf{g}} \mathsf{pls}_{i,\mathsf{g}}) \rightarrow \neg\varphi)$.

Let us assume that $\mathsf{k} > 0$. Then:

$M, w \vDash \mathsf{ODes}_i^{\mathsf{k}} \varphi$ IFF,

$\max\limits_{v \in \|\varphi\|_{w,i}} \mathcal{P}(v,i) = \mathsf{k}$ IFF,

there is $v \in \|\varphi\|_{w,i}$ such that $\mathcal{P}(v,i) = \mathsf{k}$ and

for all $u \in \mathcal{E}_i(w)$, if $\mathcal{P}(u,i) > \mathsf{k}$ then $M, u \vDash \neg\varphi$ IFF,

$M, w \vDash \hat{\mathsf{K}}_i(\mathsf{pls}_{i,\mathsf{k}} \wedge \varphi) \wedge \bigwedge_{\mathsf{g} \in Num : \mathsf{k} < \mathsf{g}} \mathsf{K}_i(\mathsf{pls}_{i,\mathsf{g}} \rightarrow \neg\varphi)$ IFF,

$M, w \vDash \hat{\mathsf{K}}_i(\mathsf{pls}_{i,\mathsf{k}} \wedge \varphi) \wedge \mathsf{K}_i((\bigvee_{\mathsf{g} \in Num : \mathsf{k} < \mathsf{g}} \mathsf{pls}_{i,\mathsf{g}}) \rightarrow \neg\varphi)$.

As for the third item, we distinguish two cases: $\mathsf{k} < \mathsf{maxVal}$ and $\mathsf{k} = \mathsf{maxVal}$. Let us assume that $\mathsf{k} = \mathsf{maxVal}$. Then:

$M, w \vDash \mathsf{PDes}_i^{\mathsf{maxVal}} \varphi$ IFF,

$\min\limits_{v \in \|\varphi\|_{w,i}} \mathcal{P}(v,i) = \mathsf{maxVal}$ IFF,

(there is $v \in \|\varphi\|_{w,i}$ such that $\mathcal{P}(v,i) = \mathsf{maxVal}$ and

for all $u \in \mathcal{E}_i(w)$, if $\mathcal{P}(u,i) < \mathsf{maxVal}$ then

$M,u \vDash \neg\varphi$) or ($\|\varphi\|_{w,i} = \emptyset$) IFF,
for all $u \in \mathcal{E}_i(w)$, if $\mathcal{P}(u,i) < \mathsf{maxVal}$ then $M,u \vDash \neg\varphi$ IFF,
$M,w \vDash \bigwedge_{\mathsf{g} \in Num:\mathsf{g}<\mathsf{maxVal}} \mathsf{K}_i(\mathsf{pls}_{i,\mathsf{g}} \rightarrow \neg\varphi)$ IFF,
$M,w \vDash \mathsf{K}_i((\bigvee_{\mathsf{g} \in Num:\mathsf{g}<\mathsf{maxVal}} \mathsf{pls}_{i,\mathsf{g}}) \rightarrow \neg\varphi)$.

Let us assume that $\mathsf{k} < \mathsf{maxVal}$. Then:

$M,w \vDash \mathsf{ODes}_i^{\mathsf{k}}\varphi$ IFF,
$\min_{v \in \|\varphi\|_{w,i}} \mathcal{P}(v,i) = \mathsf{k}$ IFF,
there is $v \in \|\varphi\|_{w,i}$ such that $\mathcal{P}(v,i) = \mathsf{k}$ and
for all $u \in \mathcal{E}_i(w)$, if $\mathcal{P}(u,i) < \mathsf{k}$ then $M,u \vDash \neg\varphi$ IFF,
$M,w \vDash \hat{\mathsf{K}}_i(\mathsf{pls}_{i,\mathsf{k}} \wedge \varphi) \wedge \bigwedge_{\mathsf{g} \in Num:\mathsf{g}<\mathsf{k}} \mathsf{K}_i(\mathsf{pls}_{i,\mathsf{g}} \rightarrow \neg\varphi)$ IFF,
$M,w \vDash \hat{\mathsf{K}}_i(\mathsf{pls}_{i,\mathsf{k}} \wedge \varphi) \wedge \mathsf{K}_i((\bigvee_{\mathsf{g} \in Num:\mathsf{g}<\mathsf{k}} \mathsf{pls}_{i,\mathsf{g}}) \rightarrow \neg\varphi)$.

A.5. *Proof of Proposition 7*

We only prove validities (7), (9), (11) and (13). Validities (8), (10), (12) and (14) can be proved in a similar way.

Proof of (7):

$M,w \vDash \mathsf{ODes}_i^{\mathsf{k}}\varphi \wedge \mathsf{ODes}_i^{\mathsf{g}}\psi$ IFF,
$\max_{v \in \|\varphi\|_{w,i}} \mathcal{P}(v,i) = \mathsf{k}$ and $\max_{v \in \|\psi\|_{w,i}} \mathcal{P}(v,i) = \mathsf{g}$ THEN,
$\max_{v \in \|\varphi\|_{w,i} \cup \|\psi\|_{w,i}} \mathcal{P}(v,i) = \max\{\mathsf{k},\mathsf{g}\}$ IFF,
$\max_{v \in \|\varphi \vee \psi\|_{w,i}} \mathcal{P}(v,i) = \max\{\mathsf{k},\mathsf{g}\}$ IFF,
$M,w \vDash \mathsf{ODes}_i^{\max\{\mathsf{k},\mathsf{g}\}}(\varphi \vee \psi)$.

Proof of (9):

$M,w \vDash \mathsf{PDes}_i^{\mathsf{k}}\varphi \wedge \mathsf{PDes}_i^{\mathsf{g}}\psi$ IFF,
$\min_{v \in \|\varphi\|_{w,i}} \mathcal{P}(v,i) = \mathsf{k}$ and $\min_{v \in \|\psi\|_{w,i}} \mathcal{P}(v,i) = \mathsf{g}$ THEN,
$\min_{v \in \|\varphi\|_{w,i} \cup \|\psi\|_{w,i}} \mathcal{P}(v,i) = \min\{\mathsf{k},\mathsf{g}\}$ IFF,
$\min_{v \in \|\varphi \vee \psi\|_{w,i}} \mathcal{P}(v,i) = \min\{\mathsf{k},\mathsf{g}\}$ IFF,
$M,w \vDash \mathsf{PDes}_i^{\min\{\mathsf{k},\mathsf{g}\}}(\varphi \vee \psi)$.

Proof of (11):

$M,w \vDash \mathsf{ODes}_i^{\mathsf{k}}\varphi \wedge \mathsf{ODes}_i^{\mathsf{g}}\psi$ IFF,

$\max\limits_{v\in\|\varphi\|_{w,i}} \mathcal{P}(v,i)=\mathsf{k}$ and $\max\limits_{v\in\|\psi\|_{w,i}} \mathcal{P}(v,i)=\mathsf{g}$ THEN,

$\max\limits_{v\in\|\varphi\|_{w,i}\cap\|\psi\|_{w,i}} \mathcal{P}(v,i)\le \min\{\mathsf{k},\mathsf{g}\}$ IFF,

$\max\limits_{v\in\|\varphi\wedge\psi\|_{w,i}} \mathcal{P}(v,i)\le \min\{\mathsf{k},\mathsf{g}\}$ IFF,

$M,w \vDash \mathsf{ODes}_i^{\le \min\{\mathsf{k},\mathsf{g}\}}(\varphi\wedge\psi)$.

Proof of (13):

$M,w \vDash \mathsf{PDes}_i^{\mathsf{k}}\varphi \wedge \mathsf{PDes}_i^{\mathsf{g}}\psi$ IFF,

$\min\limits_{v\in\|\varphi\|_{w,i}} \mathcal{P}(v,i)=\mathsf{k}$ and $\min\limits_{v\in\|\psi\|_{w,i}} \mathcal{P}(v,i)=\mathsf{g}$ THEN,

$\min\limits_{v\in\|\varphi\|_{w,i}\cap\|\psi\|_{w,i}} \mathcal{P}(v,i)\ge \max\{\mathsf{k},\mathsf{g}\}$ IFF,

$\min\limits_{v\in\|\varphi\wedge\psi\|_{w,i}} \mathcal{P}(v,i)\ge \max\{\mathsf{k},\mathsf{g}\}$ IFF,

$M,w \vDash \mathsf{PDes}_i^{\ge \max\{\mathsf{k},\mathsf{g}\}}(\varphi\wedge\psi)$.

A.6. *Proof of Proposition 8*

We only prove the first item and the third item, as the second item and the fourth item can be proved in a similar way.

$M,w \vDash \psi \le_i^{\mathsf{OPIs}} \varphi$ IFF,

$M,w \vDash \bigvee_{\mathsf{k}\in Num}(\mathsf{ODes}_i^{\mathsf{k}}\varphi \wedge \bigvee_{\mathsf{g}\in Num:\mathsf{g}>\mathsf{k}}\neg\mathsf{ODes}_i^{\mathsf{g}}\psi)$ IFF,

$\max\limits_{v\in\|\psi\|_{w,i}} \mathcal{P}(v,i)\le \max\limits_{v\in\|\varphi\|_{w,i}} \mathcal{P}(v,i)$ IFF,

for all $v\in\mathcal{E}_i(w)$, if $M,v\vDash\psi$ then

there is $u\in\mathcal{E}_i(w)$ such that $\mathcal{P}(v,i)\le\mathcal{P}(u,i)$ and $M,u\vDash\varphi$.

$M,w \vDash \psi \le_i^{\mathsf{PPIs}} \varphi$ IFF,

$M,w \vDash \bigvee_{\mathsf{k}\in Num}(\mathsf{PDes}_i^{\mathsf{k}}\varphi \wedge \bigvee_{\mathsf{g}\in Num:\mathsf{g}>\mathsf{k}}\neg\mathsf{PDes}_i^{\mathsf{g}}\psi)$ IFF,

$\min\limits_{v\in\|\psi\|_{w,i}} \mathcal{P}(v,i)\le \min\limits_{v\in\|\varphi\|_{w,i}} \mathcal{P}(v,i)$ IFF,

for all $v\in\mathcal{E}_i(w)$, if $M,v\vDash\varphi$ then

there is $u\in\mathcal{E}_i(w)$ such that $\mathcal{P}(u,i)\le\mathcal{P}(v,i)$ and $M,u\vDash\psi$.

A.7. *Proof of Proposition 9*

We only prove the first item, as the second item can proved in a way similar to the third item of Proposition 6 (see above for the proof). We distinguish two cases: $\mathsf{y} > 0$ and $\mathsf{y} = 0$. Let us remind that $\mathsf{util}_{i,\mathsf{y}} \stackrel{\text{def}}{=} \bigvee_{\mathsf{k},\mathsf{g},\mathsf{m}\in Num:\mathsf{y}=\mathsf{m}\times\mathsf{g}+(\mathsf{maxVal}-\mathsf{m})\times\mathsf{k}}(\mathsf{moral}_{i,\mathsf{m}} \wedge \mathsf{pls}_{i,\mathsf{k}} \wedge \mathsf{idl}_{i,\mathsf{g}})$.

Let us assume that $\mathsf{y} = 0$. Then:

$M,w \models \mathsf{OPref}_i^0\varphi$ IFF,

$\max_{v\in\|\varphi\|_{w,i}} \mathcal{U}(v,i) = 0$ IFF,

(there is $v\in\|\varphi\|_{w,i}$ such that $\mathcal{U}(v,i) = 0$ and

for all $u\in\mathcal{E}_i(w)$, if $\mathcal{U}(u,i) > 0$ then

$M,u \models \neg\varphi$) or ($\|\varphi\|_{w,i} = \emptyset$) IFF,

for all $u\in\mathcal{E}_i(w)$, if $\mathcal{U}(u,i) > 0$ then $M,u \models \neg\varphi$ IFF,

for all $u\in\mathcal{E}_i(w)$, if $\mathcal{S}(u,i)\times\mathcal{I}(u,i) + (\mathsf{maxVal} - \mathcal{S}(u,i))\times\mathcal{P}(u,i) > 0$

then $M,u \models \neg\varphi$ IFF,

for all $u\in\mathcal{E}_i(w)$, if there are $\mathsf{k},\mathsf{g},\mathsf{m}\in Num$ such that

$\mathsf{m}\times\mathsf{g} + (\mathsf{maxVal} - \mathsf{m})\times\mathsf{k} > 0$ and

$\mathcal{S}(u,i) = \mathsf{m}$ and $\mathcal{P}(u,i) = \mathsf{k}$ and $\mathcal{I}(u,i) = \mathsf{g}$ then $M,u \models \neg\varphi$ IFF,

$M,w \models \bigwedge_{\mathsf{z}\in UScale:0<\mathsf{z}} \mathsf{K}_i(\mathsf{util}_{i,\mathsf{z}} \rightarrow \neg\varphi)$ IFF,

$M,w \models \mathsf{K}_i((\bigvee_{\mathsf{z}\in UScale:0<\mathsf{z}} \mathsf{util}_{i,\mathsf{z}}) \rightarrow \neg\varphi)$.

Let us assume that $\mathsf{y} = 0$. Then:

$M,w \models \mathsf{OPref}_i^{\mathsf{y}}\varphi$ IFF,

$\max_{v\in\|\varphi\|_{w,i}} \mathcal{U}(v,i) = \mathsf{y}$ IFF,

there is $v\in\|\varphi\|_{w,i}$ such that $\mathcal{U}(v,i) = \mathsf{y}$ and

for all $u\in\mathcal{E}_i(w)$, if $\mathcal{U}(u,i) > \mathsf{y}$ then $M,u \models \neg\varphi$ IFF,

there is $v\in\|\varphi\|_{w,i}$ such that $\mathcal{S}(v,i)\times\mathcal{I}(v,i) + (\mathsf{maxVal} - \mathcal{S}(v,i))\times\mathcal{P}(v,i) = \mathsf{y}$

and for all $u\in\mathcal{E}_i(w)$, if $\mathcal{S}(u,i)\times\mathcal{I}(u,i) + (\mathsf{maxVal} - \mathcal{S}(u,i))\times\mathcal{P}(u,i) > \mathsf{y}$

then $M,u \models \neg\varphi$ IFF,

there are $v\in\|\varphi\|_{w,i}$ and $\mathsf{k}_1,\mathsf{g}_1,\mathsf{m}_1\in Num$ such that

$\mathsf{m}_1\times\mathsf{g}_1 + (\mathsf{maxVal} - \mathsf{m}_1)\times\mathsf{k}_1 = \mathsf{y}$ and

$\mathcal{S}(v,i) = \mathsf{m}_1$ and $\mathcal{P}(v,i) = \mathsf{k}_1$ and $\mathcal{I}(v,i) = \mathsf{g}_1$ and

for all $u \in \mathcal{E}_i(w)$, if there are $\mathsf{k}_2, \mathsf{g}_2, \mathsf{m}_2 \in Num$ such that
$\mathsf{m}_2 \times \mathsf{g}_2 + (\mathsf{maxVal} - \mathsf{m}_2) \times \mathsf{k}_2 > \mathsf{z}$ and
$\mathcal{S}(u,i) = \mathsf{m}_2$ and $\mathcal{P}(u,i) = \mathsf{k}_2$ and $\mathcal{I}(u,i) = \mathsf{g}_2$ then $M,u \models \neg\varphi$ IFF,
$M,w \models \hat{\mathsf{K}}_i(\mathsf{util}_{i,\mathsf{y}} \wedge \varphi) \wedge \bigwedge_{\mathsf{z} \in UScale:\mathsf{y}<\mathsf{z}} \mathsf{K}_i(\mathsf{util}_{i,\mathsf{z}} \rightarrow \neg\varphi)$ IFF,
$M,w \models \hat{\mathsf{K}}_i(\mathsf{util}_{i,\mathsf{y}} \wedge \varphi) \wedge \mathsf{K}_i((\bigvee_{\mathsf{z} \in UScale:\mathsf{y}<\mathsf{z}} \mathsf{util}_{i,\mathsf{z}}) \rightarrow \neg\varphi)$.

A.8. *Proof of Proposition 12*

We only prove validities (12) and (15). Validities (13) and (14) can be proved in a similar way.

Let us prove validity (12) first. $M,w \models (\psi <_i^{\mathsf{OPls}} \varphi) \wedge \mathsf{moral}_{i,0}$ implies $\max_{v \in \|\psi\|_{w,i}} \mathcal{P}(v,i) < \max_{v \in \|\varphi\|_{w,i}} \mathcal{P}(v,i)$ and $\mathcal{S}(w,i) = 0$. By item 1 of Proposition 8, we have $\max_{v \in \|\psi\|_{w,i}} \mathcal{P}(v,i) < \max_{v \in \|\varphi\|_{w,i}} \mathcal{P}(v,i)$ if and only if:

(A) for all $v \in \mathcal{E}_i(w)$, if $M,v \models \psi$ then there is $u \in \mathcal{E}_i(w)$ such that $M,u \models \varphi$ and $\mathcal{P}(v,i) < \mathcal{P}(u,i)$.

Moreover, by Constraint (**Constr**) on LAMA$^+$ models and the definition of function $\mathcal{U}$, $\mathcal{S}(w,i) = 0$ implies that:

(B) for all $v \in \mathcal{E}_i(w)$, $\mathcal{U}(v,i) = \mathsf{maxVal} \times \mathcal{P}(v,i)$.

(A) and (B) together imply that for all $v \in \mathcal{E}_i(w)$, if $M,v \models \psi$ then there is $u \in \mathcal{E}_i(w)$ such that $M,u \models \varphi$ and $\mathcal{U}(v,i) < \mathcal{U}(u,i)$. The latter, by item 1 of Proposition 11, implies that $M,w \models \psi <_i^{\mathsf{OUtil}} \varphi$.

Let us now prove validity (15). $M,w \models (\psi <_i^{\mathsf{PIdl}} \varphi) \wedge \mathsf{moral}_{i,\mathsf{maxVal}}$ implies $\min_{v \in \|\psi\|_{w,i}} \mathcal{I}(v,i) < \min_{v \in \|\varphi\|_{w,i}} \mathcal{I}(v,i)$ and $\mathcal{S}(w,i) = \mathsf{maxVal}$. By item 4 of Proposition 8, we have $\min_{v \in \|\psi\|_{w,i}} \mathcal{I}(v,i) < \min_{v \in \|\varphi\|_{w,i}} \mathcal{I}(v,i)$ if and only if:

(A) for all $v \in \mathcal{E}_i(w)$, if $M,v \models \varphi$ then there is $u \in \mathcal{E}_i(w)$ such that $M,u \models \psi$ and $\mathcal{I}(u,i) < \mathcal{I}(v,i)$.

Moreover, by Constraint (**Constr**) on LAMA$^+$ models and the definition of function $\mathcal{U}$, $\mathcal{S}(w,i) = \mathsf{maxVal}$ implies that:

(B) for all $v \in \mathcal{E}_i(w)$, $\mathcal{U}(v,i) = \mathsf{maxVal} \times \mathcal{I}(v,i)$.

(A) and (B) together imply that for all $v \in \mathcal{E}_i(w)$, if $M,v \models \varphi$ then there is $u \in \mathcal{E}_i(w)$ such that $M,u \models \psi$ and $\mathcal{U}(u,i) < \mathcal{U}(v,i)$. The latter, by item 4 of Proposition 11, implies that $M,w \models \psi <_i^{\mathsf{PUtil}} \varphi$.

A.9. *Proof of Proposition 13*

We prove (16) and (18) as an example. The proofs are by reductio ad absurdum.

Let us prove (16). Suppose that:

(A1) $\mathsf{m}\times\mathsf{k} > \mathsf{m}\times\mathsf{g} + (\mathsf{maxVal}-\mathsf{m})\times\mathsf{j}$ for all $\mathsf{j}\in \mathit{Num}$.

Moreover suppose that:

(B1) $M,w \vDash \mathsf{OVal}_i^{\mathsf{k}}\varphi \wedge \mathsf{OVal}_i^{\mathsf{g}}\psi \wedge \mathsf{moral}_{i,\mathsf{m}} \wedge \neg(\psi <_i^{\mathsf{OUtil}} \varphi)$.

(B1) means that $\max_{v\in\|\varphi\|_{w,i}} \mathcal{I}(v,i) = \mathsf{k}$ and $\max_{v\in\|\psi\|_{w,i}} \mathcal{I}(v,i) = \mathsf{g}$ and $\mathcal{S}(w,i) = \mathsf{m}$ and $\max_{v\in\|\varphi\|_{w,i}} \mathcal{U}(v,i) \leq \max_{v\in\|\psi\|_{w,i}} \mathcal{U}(v,i)$. By the Constraint (**Constr**), $\mathcal{S}(w,i) = \mathsf{m}$ implies that for all v, if $w\mathcal{E}_i v$ then $\mathcal{S}(v,i) = \mathsf{m}$. The latter together with $\max_{v\in\|\varphi\|_{w,i}} \mathcal{I}(v,i) = \mathsf{k}$ and $\max_{v\in\|\psi\|_{w,i}} \mathcal{I}(v,i) = \mathsf{g}$ implies that:

(C1) there is v such that $w\mathcal{E}_i v$ and $M,v \vDash \varphi$ and $\mathcal{U}(v,i) = \mathsf{m} \times \mathsf{k} + (\mathsf{maxVal}-\mathsf{m})\times\mathsf{j}$ for some $\mathsf{j}\in \mathit{Num}$,

(D1) for all v, if $w\mathcal{E}_i v$ and $M,v \vDash \psi$ then $\mathcal{U}(v,i) \leq \mathsf{m} \times \mathsf{g} + (\mathsf{maxVal}-\mathsf{m}) \times \mathsf{maxVal}$.

(A1), (C1) and (D1) together imply that:

(E1) there is v such that $w\mathcal{E}_i v$ and $M,v \vDash \varphi$ and, for all u, if $w\mathcal{E}_i u$ and $M,u \vDash \psi$ then $\mathcal{U}(v,i) > \mathcal{U}(u,i)$.

But (E1) is in contradiction with $\max_{v\in\|\varphi\|_{w,i}} \mathcal{U}(v,i) \leq \max_{v\in\|\psi\|_{w,i}} \mathcal{U}(v,i)$.

Let us now prove (18). Suppose that:

(A2) $\mathsf{m}\times\mathsf{k} > \mathsf{m}\times\mathsf{g} + (\mathsf{maxVal}-\mathsf{m})\times\mathsf{j}$ for all $\mathsf{j}\in \mathit{Num}$.

Moreover suppose that:

(B2) $M,w \vDash \mathsf{PVal}_i^{\mathsf{k}}\varphi \wedge \mathsf{PVal}_i^{\mathsf{g}}\psi \wedge \mathsf{moral}_{i,\mathsf{m}} \wedge \neg(\psi <_i^{\mathsf{OUtil}} \varphi)$.

(B2) means that $\min_{v\in\|\varphi\|_{w,i}} \mathcal{I}(v,i) = \mathsf{k}$ and $\min_{v\in\|\psi\|_{w,i}} \mathcal{I}(v,i) = \mathsf{g}$ and $\mathcal{S}(w,i) = \mathsf{m}$ and $\min_{v\in\|\varphi\|_{w,i}} \mathcal{U}(v,i) \leq \min_{v\in\|\psi\|_{w,i}} \mathcal{U}(v,i)$. By the Constraint (**Constr**), $\mathcal{S}(w,i) = \mathsf{m}$ implies that for all v, if $w\mathcal{E}_i v$ then $\mathcal{S}(v,i) = \mathsf{m}$. The latter together with $\min_{v\in\|\varphi\|_{w,i}} \mathcal{I}(v,i) = \mathsf{k}$ and $\min_{v\in\|\psi\|_{w,i}} \mathcal{I}(v,i) = \mathsf{g}$ implies that:

(C2) there is v such that $w\mathcal{E}_i v$ and $M,v \vDash \psi$ and $\mathcal{U}(v,i) = \mathsf{m} \times \mathsf{g} + (\mathsf{maxVal}-\mathsf{m})\times\mathsf{j}$ for some $\mathsf{j}\in \mathit{Num}$,

(D2) for all v, if $w\mathcal{E}_i v$ and $M,v \vDash \varphi$ then $\mathcal{U}(v,i) \geq \mathsf{m}\times\mathsf{k}$.

(A2), (C2) and (D2) together imply that:

(E2) there is v such that $w\mathcal{E}_i v$ and $M,v \vDash \psi$ and, for all u, if $w\mathcal{E}_i u$ and $M,u \vDash \varphi$ then $\mathcal{U}(v,i) < \mathcal{U}(u,i)$.

But (E2) is in contradiction with $\min\limits_{v\in\|\varphi\|_{w,i}} \mathcal{U}(v,i) \leq \min\limits_{v\in\|\psi\|_{w,i}} \mathcal{U}(v,i)$.

A.10. *Proof of Proposition 14*

We prove the first item as an example.

($\Rightarrow$) Suppose $M,w \vDash \mathsf{ORat}_i$ with $M = \langle W,\mathcal{H},\{\mathcal{E}_i\}_{i\in Agt},\mathcal{P},\mathcal{I},\mathcal{V},\mathcal{S}\rangle$. As $\mathcal{H}$ is a total function, we have that there is $a \in Rep_i$ such that $M,w \vDash \mathsf{choose}_{i,a}$. The latter together with $M,w \vDash \mathsf{ORat}_i$ implies that there is $a \in Rep_i$ such that $M,w \vDash \mathsf{choose}_{i,a}$ and $M,w \vDash \neg\mathsf{choose}_{i,a} \leq_i^{\mathsf{OUtil}} \mathsf{choose}_{i,a}$. But $M,w \vDash \neg\mathsf{choose}_{i,a} \leq_i^{\mathsf{OUtil}} \mathsf{choose}_{i,a}$ just means that $\max\limits_{v\in\|\neg\mathsf{choose}_{i,a}\|_{w,i}} \mathcal{U}(v,i) \leq \max\limits_{v\in\|\mathsf{choose}_{i,a}\|_{w,i}} \mathcal{U}(v,i)$. By definition of $\mathsf{choose}_{i,a}$, we have that for all $b \neq a$, $\|\mathsf{choose}_{i,b}\|_{w,i} \subseteq \|\neg\mathsf{choose}_{i,a}\|_{w,i}$. Therefore, it follows that there is $a \in Rep_i$ such that $M,w \vDash \mathsf{choose}_{i,a}$ and for all $b \neq a$, $\max\limits_{v\in\|\mathsf{choose}_{i,b}\|_{w,i}} \mathcal{U}(v,i) \leq \max\limits_{v\in\|\mathsf{choose}_{i,a}\|_{w,i}} \mathcal{U}(v,i)$. The latter implies that there is $a \in \operatorname{argmax}\limits_{a\in Rep_i} \max\limits_{v\in\|\mathsf{choose}_{i,a}\|_{w,i}} \mathcal{U}(v,i)$ such that $M,w \vDash \mathsf{choose}_{i,a}$.

($\Leftarrow$) Suppose there is $a \in \operatorname{argmax}\limits_{a\in Rep_i} \max\limits_{v\in\|\mathsf{choose}_{i,a}\|_{w,i}} \mathcal{U}(v,i)$ such that $M,w \vDash \mathsf{choose}_{i,a}$. It follows that there is $a \in Rep_i$ such that $M,w \vDash \mathsf{choose}_{i,a}$ and for all $b \neq a$, $\max\limits_{v\in\|\mathsf{choose}_{i,b}\|_{w,i}} \mathcal{U}(v,i) \leq \max\limits_{v\in\|\mathsf{choose}_{i,a}\|_{w,i}} \mathcal{U}(v,i)$. We have that $\bigcup_{b\neq a}\|\mathsf{choose}_{i,b}\|_{w,i} = \|\neg\mathsf{choose}_{i,a}\|_{w,i}$. Therefore, we can conclude that there is $a \in Rep_i$ such that $M,w \vDash \mathsf{choose}_{i,a}$ and $\max\limits_{v\in\|\neg\mathsf{choose}_{i,a}\|_{w,i}} \mathcal{U}(v,i) \leq \max\limits_{v\in\|\mathsf{choose}_{i,a}\|_{w,i}} \mathcal{U}(v,i)$. The latter implies that $M,w \vDash \mathsf{ORat}_i$ because $\max\limits_{v\in\|\neg\mathsf{choose}_{i,a}\|_{w,i}} \mathcal{U}(v,i) \leq \max\limits_{v\in\|\mathsf{choose}_{i,a}\|_{w,i}} \mathcal{U}(v,i)$ is equivalent to $M,w \vDash \neg\mathsf{choose}_{i,a} \leq_i^{\mathsf{OUtil}} \mathsf{choose}_{i,a}$.

A.11. *Proof of Proposition 15*

We only prove validity (15), as the proofs of validities (16), (17) and (18) follow the same pattern.

Suppose that $M,w \vDash (\neg\mathsf{choose}_{i,a} <_i^{\mathsf{OPls}} \mathsf{choose}_{i,a}) \wedge \mathsf{moral}_{i,0} \wedge \mathsf{ORat}_i$. By validity 12 of Proposition 12, $M,w \vDash (\neg\mathsf{choose}_{i,a} <_i^{\mathsf{OPls}} \mathsf{choose}_{i,a}) \wedge \mathsf{moral}_{i,0}$

implies $M,w \models \neg\mathsf{choose}_{i,a} <_i^{\mathsf{OUtil}} \mathsf{choose}_{i,a}$. By item 1 of Proposition 11, we have $M,w \models \neg\mathsf{choose}_{i,a} <_i^{\mathsf{OUtil}} \mathsf{choose}_{i,a}$ if and only if:

(A) for all $v \in \mathcal{E}_i(w)$, if $M,v \models \neg\mathsf{choose}_{i,a}$ then there is $u \in \mathcal{E}_i(w)$ such that $\mathcal{U}(v,i) < \mathcal{U}(u,i)$ and $M,u \models \mathsf{choose}_{i,a}$.

Moreover, we have:

(B) for all $b \in Act$ such that $b \neq a$ and for all $v \in W$, if $M,v \models \mathsf{choose}_{i,b}$ then if $M,v \models \neg\mathsf{choose}_{i,a}$.

(A) and (B) together imply that:

(C) for all $b \in Act$ such that $b \neq a$ and for all $v \in \mathcal{E}_i(w)$, if $M,v \models \mathsf{choose}_{i,b}$ then there is $u \in \mathcal{E}_i(w)$ such that $\mathcal{U}(v,i) < \mathcal{U}(u,i)$ and $M,u \models \mathsf{choose}_{i,a}$.

(C) implies that:

(D) $\underset{a \in Rep_i}{\operatorname{argmax}} \max_{v \in \|\mathsf{choose}_{i,a}\|_{w,i}} \mathcal{U}(v,i) = \{a\}$.

By item 1 of Proposition 14, (D) together with $M,w \models \mathsf{ORat}_i$ imply that $M,w \models \mathsf{choose}_{i,a}$.