Moral Systems of Agent Societies: Some Elements for their Analysis and Design

Antônio Carlos da Rocha Costa [1]

Abstract. This paper introduces elements for the foundation of a notion of *moral system of an agent society*. The paper is specially concerned with elements for the design and analysis of moral systems of agent societies that are to be embedded in social contexts involving diverse human groups.

1 Introduction

Moral systems embody *norms and values* about the *conducts* (behaviors, interactions) that are possible in a society, as well as any *knowledge* that may be available about those conducts, norms and values [14].

In this paper, we introduce the core elements of a formal foundation for *moral systems of agent societies*. In analogy to H. Kelsen's theory of legal systems [13], the formal foundation that we envisage concentrates on the principles of the *structure and operation of moral systems,* not on the *contents of their norms and values*.

We use the term "moral knowledge" to denote knowledge that an agent has about another agent's morality. The set of moral features determined by such moral knowledge constitutes the *moral model* that the former (the *moral modeler*) has about the latter (the one *morally modeled*).

A moral model specifies the moral knowledge on the basis of which an agent ag_1 analyzes both the conducts of some agent ag_2 (possibly itself) and the moral assessments that ag_2 does about the social conducts of any agent ag_3. The core of the moral model that ag_1 has of ag_2 is the set of moral norms that ag_1 believes that ag_2 has adopted.

The moral knowledge embodied by a moral model is *relativistic*, for a variety of reasons. For instance, the moral knowledge embodied in a moral model depends on which are the agents (moral modeler and morally modeled) it concerns and on the means available for the moral modeler to gather information about the agent morally modeled.

Also, moral models are *observational models*, and the moral knowledge they embody can only be acquired in a piecewise way. In consequence, at each point in time, any moral model is *tentative*, regarding the information that the moral modeler could gather, up to that time.

Thus, the moral knowledge embodied in a moral model is always *incomplete* and, so, incapable to fully morally differentiate that agent from others, morally similar agents.

[1] Programa de Pós-Graduação em Informática na Educação da UFRGS. 90.040-060 Porto Alegre, Brazil. Programa de Pós-Graduação em Computação da FURG. 96.203-900 Rio Grande, Brazil. Email: ac.rocha.costa@gmail.com .

In consequence, any moral modeling of an agent by another is, in fact, the moral modeling of a *class of agents*, always being more general than the modeling of one particular agent.

Any *moral judgment* of an individual agent is necessarily, then, a judgment based on a *moral model of a class of agents*, to which that agent is considered to belong, not about that individual agent, specifically.

So, in principle, any such moral judgment is inevitably *prejudicial*, or *stereotypical*, in the sense that it is necessarily based on a *prejudice* about the individual agent being morally judged, namely, the prejudice that the individual fully fits the *general moral features* of the class of agents to which refers the moral model used to support the moral judgment.

By the same token, the moral judgment about an agent may be *seamlessly extended*, in an even more prejudicial way, to the *totality of agents* presumed to belong to the class of agents to which that agent is itself presumed to belong (that is, the class of agents referred to by the moral model).

One sees, then, that moral models have two important effects on the conducts of agents and groups of agents. They are a necessary means for the establishment of the indispensable *minimum level of mutual moral understanding* within any group of agents that constitutes itself as a *social group*.

They are also, however, a potential source of *misconceptions* of agents and groups of agents about each other. They are also, thus, a potential source of *moral misunderstandings* (more specifically, *moral conflicts* and their consequent *moral contradictions*) among those agents and groups of agents.

1.1 The Aims and Structure of the Paper

This paper aims to introduce conceptual elements necessary for a formal account of the structure and functioning of *moral systems* in agent societies, so that methods for the *moral analysis and design* of agent societies can be soundly established.

The paper concentrates on the basic components of such moral systems, namely, *moral models*, which are the structures that embody the moral knowledge that agents and social groups may have about each other.

In Sect. 2, we review J. Halpern and Y. Moses' way of formally accounting for *knowledge* that is *about*, and *situated in*, computational systems. We specialize their conception to knowledge about, and situated in, *agent societies*, and extend it to deal with the *relativistic* nature of such knowledge.

The result is the formal concept of knowledge that we use to account for the *epistemic aspects* of the notion of *moral knowledge* that we think is appropriate to agent societies.

In Sect. 3, we formally introduce the concepts of *moral knowledge*, *moral model* and *moral judgments*, as well as the concepts of *morally assigned group identity*, *moral prejudice*, and *moral contradiction* between social groups.

Finally, in Sect. 5, the paper introduces a notion of *moral design of agent societies*, built on the conceptual framework introduced previously, and briefly relates moral design to other parts of the *organizational design* of agent societies.

For completeness, we summarize now the notion of agent society adopted here.

1.2 Agent Society, Agent Conduct

The notion of agent society that we adopt here is the one we have been using in our work (see, e.g., [7]): we take an *agent society* to be an open, organized, persistent and situated multiagent system, where:

- *openness* means that the agents of the society can freely enter and leave it;
- *organization* means that the working of the society is based on an *articulation of individual and collective conducts*[2], the collective ones performed by *groups of agents* of various kinds (institutionalized or not);
- *persistence* means that the organization persists in time, independently of the agents that enter or leave the society;
- *situatedness* means that the society exists and operates in a definite physical environment, involving physical objects that the agents and groups of agents may make use of, in the performance of their individual and collective conducts.

Formally, the *organization* of an agent society is a structure encompassing *groups of agents* (possibly singletons), together with the *conducts* that such groups of agents perform. The groups of agents constitute the *organizational units* of the society (independently of their being institutionalized or not).

2 Knowledge About an Agent Society that is Situated in that Society

We start with a *general notion of knowledge*, construed to be both *about* an agent society, and *situated* in that agent society. For that, we build on the general notion of *knowledge about a distributed computational system* that is *situated in that system*, which was introduced by Halpern and Moses [11]. We take the presentation of that notion in [10] as our basis.

Notice the crucial role that the concept of *external observer* plays in our overall conception.

2.1 General Characterization

A general characterization of *knowledge* in an agent society can be given as follows. Let:

- $G = \{ag_1, \ldots, ag_n\}$ be a finite set, composed of n agents, generically ranged over by the variables ag_i and ag_j;
- P^* be a set of *primitive propositions*, generically ranged over by variables p and p';
- $\wedge$ and $\neg$ be propositional operators that (together with the operators $\vee$ and $\Rightarrow$, defined from them) extend the set P^* to the set P of compound propositions, also generically ranged over by the variables p and p'.

We take $\mathcal{K}_{ag_1}, \ldots, \mathcal{K}_{ag_n}$ to be *epistemic operators*, such that $\mathcal{K}_{ag_i}(p)$ means that $p \in P$ *belongs to the knowledge* of the agent ag_i, that is, that agent ag_i *knows that* p.

Three additional notions of knowledge are presented in [10], besides this notion of *individual knowledge* $\mathcal{K}_{ag_i}(p)$. They refer to knowledge held by *groups of agents*:

- $\mathcal{E}_G(p)$, which means: p belongs to the *knowledge of each of* the agents of the group G;
- $\mathcal{C}_G(p)$, which means: p belongs to the recursive notion of *common knowledge* of the agents of the group G, that is: each of the agents of the group G knows that p; each of the agents of the group G knows that each of the agents of the group G knows that p; etc.;
- $\mathcal{I}_G(p)$, which means: p belongs to the *implicit knowledge* of the agents of the group G, that is, the *union of the individual knowledges* of the agents of the group G, so that an *external observer* that holds such union can deduce p if it reasons from that union, even if none of the agents can do that by reasoning from the common knowledge of G.

This paper concentrates on propositions of the form $\mathcal{K}_{ag_i}(p)$.

2.2 External Relativity

With the notions of $\mathcal{K}_{ag_i}(p)$, $\mathcal{E}_G(p)$, $\mathcal{C}_G(p)$ and $\mathcal{I}_G(p)$, Halpern and colleagues [10, 11] proceed to analyze properties of *communication* and *action coordination* protocols in distributed systems. The basis of their approach is an interpretation, in terms of the set of the *global states* of a distributed computational system, of the *semantics of possible worlds* that Hintikka introduced in [12].

We specialize their interpretation to agent societies in the following way. An agent society is characterized by a set of *objective global states*, defined as $S_O = \Gamma_O \times T$, where Γ_O is the set of all possible *configurations* of the society [3], and T is a linear structure of discrete time instants, so that each global state of the society is a pair $s = (\gamma, t) \in S_O$.

The determination of such set of global states is *objective* in the sense that it is given by an *external observer* O that has access to all the details of the society, in a way that, from O's point of view, is taken to be *complete*. However, even though *objective* (external and complete), that characterization is still *relativistic*, precisely because it depends O's *point of view*, hence the index O in Γ_O and S_O.

Regarding the individual agents, the approach assumes that - due to the *locality* of their particular points of view - each agent of the society partitions the set of global states S_O (that O is capable of fully differentiating) into *equivalence classes*. That is, each agent is lead to take as *indistinguishable* certain global states that can be *objectively distinguished* by O.

In precise terms: an agent is lead to take two objectively different global states to be indistinguishable whenever the agent's knowledge about the society is the same in the two global states. That is, whenever the two states do not allow the agent to elaborate different knowledges about the society.

[2] By a *conduct* of an agent or group of agents we understand either a *behavior* that that agent or group performs, when considered in isolation from other agents or groups, or the *part of the interaction* that an agent or group performs, when *interacting* with other agents or groups.

[3] See [8] for the notion of *configuration of agent society*.

Formally, what is defined is an epistemic structure $M_O = (S_O, P; v_O, K_{ag_1}, \ldots, K_{ag_n})$ where:

- $S_O = \Gamma_O \times T$ is the set of objective global states of the agent society, considered from the point of view of the external observer O;
- P is a set of propositions, with basic set P^*;
- $v_O : S_O \times P \to \{T, F\}$ is a *truth assignment function* that, to each global state $s \in S_o$ and each basic proposition $p \in P$, assigns a truth value $v_O(s, p) \in \{T, F\}$, according with p being objectively true or false in the state s, from the point of view of O;
- each K_{ag_i} is an *equivalence relation* on S_O, such that if $(s, s') \in K_{ag_i}$ then agent ag_i can not distinguish between the global states s and s', as O can; that is, given the knowledge that agent ag_i has about the society, the agent takes s and s' to be indistinguishable.

We denote the fact that $p \in P$ is true in the global state $s \in S_O$ by $(M_O, s) \models p$.

With those definitions, the semantics of the epistemic operators $\mathcal{K}_{ag_i}$ takes as its basis the objective truth of the primitive propositions in P, as given by the function v_O.

Formally, we have:

- For any primitive proposition $p \in P^*$:

1) $(M_O, s) \models p$ if and only if $v_O(s, p) = T$;

- For any composed proposition $p \in P$:

2) $(M_O, s) \models \neg p$ if and only if $v_O(s, p) = F$;
3) $(M_O, s) \models (p \wedge p')$ if and only if $(M_O, s) \models p$ and $s \models_{M_O} p'$;
4) $(M_O, s) \models \mathcal{K}_{ag_i}(p)$ if and only if $(M_O, s') \models p$ for each $s' \in S_O$ such that $(s, s') \in K_{ag_i}$.

That is, an agent ag_i is *objectively considered* to know that p is true, in a given global state s, if and only if p is objectively true in s and p is objectively true in every state s' that ag_i cannot distinguish from s.

Notice that the knowledge of an agent about p being true of a global state s, in which the agent finds itself, depends on p being *objectively* true in s, that is, being true from the point of view of the external observer O. That is, an agent is objectively considered to know something about its society if and only if the external observer O considers that it does.

Clearly, this *possible world semantics* makes use of an *observational* notion of knowledge of an agent, different from any *intensional* notion of knowledge, which takes as criterion the *occurrence* of p in the *knowledge base* of the agent. Accordingly, Halpern says that p is *ascribed* to the agent [10].

We call *external relativity* such condition that results from knowledge being assigned to agents on the basis of observations made by an *external observer* that also defines the set of global states that should be taken into consideration.

2.3 Internal Relativity

We introduce now a crucial modification in the formal characterization of knowledge just presented. Instead of having an *objective, external* notion of truth, given by the function $v_O : S_O \times P \to \{T, F\}$, determined by the *external* observer of the society, we introduce a *subjective, internal* notion of truth, given by a set of functions $v_{ag_i} : S_O \times P \to \{T, F\}$, one per agent (see [9]).

That is, we let each agent make use of v_{ag_i} to decide, by itself, the truth of each proposition $p \in P$, in each global state $s \in S_O$. At the same time, however, we keep the set of global states S_O determined by the external observer O, so that a minimally objective connection is preserved in the account of the different truth functions of the agents.

What we obtain can be informally summarized as follows:

- an agent society is characterized by the set S_O of its *global states*, as determined by the external observer O;
- each agent ag_i, according to the knowledge it has, establishes a *relativistic* equivalence relation $K^R_{ag_i}$ in the set of global states S_O, so that if $(s, s') \in K^R_{ag_i}$ it happens that s and s' are indistinguishable from ag_i's point of view;
- each agent ag_i, according to the knowledge it has, assigns to the primitive propositions of the set P^*, at each global state s, a truth value that is denoted by $v_{ag_i}(s, p) \in \{T, F\}$;
- the assignment of truth values to primitive propositions is extended to composed propositions in the natural way;
- the individual knowledge of each agent ag_i is characterized by the *relativistic epistemic operator* $\mathcal{K}^R_{ag_i}$;
- whenever we want to refer to the *objective knowledge* of an agent ag_i (that is, knowledge that the agent can determine, if it uses the objective truth function v_O), we make use of the *objective epistemic operator* that we have introduced above, denoted by $\mathcal{K}_{ag_i}$.

The *relativistic epistemic structure* that characterizes the knowledge of the agents of the society is, then, given by $M^R_O = (S_O, P; v_O, K_{ag_1}, \ldots, K_{ag_n}; v_{ag_1}, \ldots, v_{ag_n}, K^R_{ag_1}, \ldots, K^R_{ag_n})$.

We denote by $(M^R_O, s) \models_{ag_i} p$ the fact that the proposition p is determined to be true in the state s, by the agent ag_i, in the context of the relativistic epistemic structure M^R_O.

Under these conditions, the semantics of the relativistic epistemic operator $\mathcal{K}^R_{ag_i}$, in a society that has M^R_O as its epistemic structure, is formally given by the following rules:

- For primitive propositions $p \in P^*$:

1) $(M^R_O, s) \models_{ag_i} p$ if and only if $v_{ag_i}(s, p) = T$;

- For composed propositions $p \in P$:

2) $(M^R_O, s) \models_{ag_i} \neg p$ if and only if $v_{ag_i}(s, p) = F$;
3) $(M^R_O, s) \models_{ag_i} (p \wedge p')$ if and only if $(M^R_O, s) \models_{ag_i} p$ and $(M^R_O, s) \models_{ag_i} p'$;
4) $(M^R_O, s) \models_{ag_i} \mathcal{K}^R_{ag_i}(p)$ if and only if $(M^R_O, s') \models_{ag_i} p$ for all $(s, s') \in K^R_{ag_i}$;

This allows us to establish another crucial point in our formal model, namely, the *rule of internal relativity*, according to which an agent ag_i is allowed to assign the knowledge of p to an agent ag_j, in accordance with ag_i's own knowledge.

- *Rule of Internal Assignment*: In the global state $s \in S_O$, agent ag_i is allowed to assign the knowledge of p to an agent ag_j, denoted by $(M^R_O, s) \models_{ag_i} \mathcal{R}^K_{ag_j}(p)$, if and only if ag_i can verify that:
  1. $(M_O, s) \models K_{ag_j} p$, that is, it can be *externally* determined (i.e., from O's point of view) that agent ag_j knows p, in the global state s;
  2. $(M^R_O, s) \models_{ag_i} \mathcal{R}^K_{ag_i}(p)$, that is, ag_i *relativistically* knows that p is true, in s.

Notice that the *external* assignment of the knowledge of p to ag_j, required by the first condition, provides an *objective point of comparison* for different such assignments.

2.4 The Externalization of Internally Relativistic Knowledge, and the Rise of Objective Epistemic Contradictions Between Agents

The only way for an agent ag_i to argue that its *relativistic* (i.e., internal) truths are *objective* truths, is by the agent *externalizing* itself, that is, by ag_i considering itself to be in the role of O. In such situation, we say that ag_i has *externalized* and *objectified* its relativistic knowledge, and we denote by ag_i^O that ag_i externalized itself, and by $M^O_{ag_i}$ its "objectified" subjective and relative epistemic structure.

By intending that $M^O_{ag_i}$ holds objectively, ag_i intends that $(M^R_O, s) \models_{ag_i} \mathcal{K}^R_{ag_i}(p)$ (i.e., that ag_i relativistically knows p in s) be equated both with $(M^O_{ag_i}, s) \models p$ (i.e., that the externalized agent ag_i^O objectively knows p in s) and with $(M_O, s) \models p$ (i.e., that p is objectively true in s).

Clearly, an externalized internal observer takes itself to be a *superagent* of the society, with the power to objectively determine what is true and what is false, in that society.

But, when two agents, ag_i and ag_j, externalize themselves, at the same time, an *objective contradiction* may be established between them, concerning what is objectively true and what is objectively false in the society.

For, in such situation, for some $s \in S_{ag_i} \cap S_{ag_j}$, the agent ag_i may consider it valid to equate $(M^R_O, s) \models_{ag_i} \mathcal{K}_{ag_i}(p)$ with $(M^O_{ag_i}, s) \models p$ and $(M_O, s) \models p$ while, at the same time, the agent ag_j may consider it valid to equate $(M^R_O, s) \models_{ag_j} \mathcal{K}_{ag_j}(\neg p)$ with $(M_{ag_j^O}, s) \models_{ag_j} \mathcal{K}^R_{ag_j^O}(\neg p)$ and $(M_O, s) \models \neg p$. So that, jointly, the two agents claim both $(M_O, s) \models p$ and $(M_O, s) \models \neg p$, which characterizes (from the point of view of O) the *objective contradiction* between them.

Moreover, under $M^O_{ag_i}$ and $M^O_{ag_j}$, the agents may conclude that $M^O_{ag_i} \models \mathcal{K}_{ag_j}(\neg p)$ and $M^O_{ag_j} \models \mathcal{K}_{ag_i}(\neg p)$, each stating that the other is "objectively" wrong.

Such *objective contradiction* about a proposition p shows that (from the point of view of O) at least one of the agents involved in the contradiction is not assessing p objectively, that is, that either $(M^R_O, s) \models_{ag_i} K^R_{ag_i} p$ or $(M^R_O, s) \models_{ag_j} K^R_{ag_j} \neg p$ (or both) does not hold, so that either v_{ag_i} or v_{ag_i} (or both) is not in accordance with v_O about s.

3 Elements for Moral Systems of Agent Societies

3.1 Moral Knowledge

As indicated in the Introduction, *moral knowledge* refers both to the knowledge of *moral norms* of conducts that agents are supposed to follow and to the knowledge of *facts* involving conducts that agents have performed, are performing, or intend to perform. Moral knowledge also refers to the *moral judgments* that the agents make of their own conducts, or of the others, and to the *moral norms* with which agents perform those moral judgments.

We construe these *four types of moral knowledge* in terms of four basic types of *moral propositions* (each type admitting additional arguments and decorations):

1. *moral norms*: propositions of the forms $prohib(Ag, Cnd)$, $oblig(Ag, Cnd)$ and $permit(Ag, Cnd)$, meaning that agents of the class of agents Ag are (respectively) prohibited, obligated and permitted to perform conducts of the class of conducts Cnd;
2. *moral facts*: propositions of the form $prfrm^t(ag_i, cnd)$, meaning that, at the time t, agent ag_i performed (or is performing, or will perform) the conduct cnd;
3. *moral judgments*: propositions of the form $asgn^t(ag_i, mfct, mv)$, meaning that, at time t, agent ag_i assigns (or is assigning, or will assign) the moral value $mv \in \{prs, blm\}$ (*praise* or *blame*) to the moral fact $mfct$;
4. *moral judgment rules*: propositions of either forms:

   (a) If $cmpl(cnd, mnrm)$ and $prfrm^t(ag_j, cnd)$

   then $allowed[asgn^{t'}(ag_i, prfmd^t(ag_j, cnd), prs)]$.

   - meaning that if the conduct cnd complies[4] with the moral norm $mnrm$ and the agent ag_j performs that conduct at time t, then an agent ag_i is allowed to *morally praise*, at any time t', the agent ag_j for performing cnd at the time t;

   (b) If $\neg cmpl(cnd, mnrm)$ and $prfrm^t(ag_j, cnd)$

   then $allowed[asgn^{t'}(ag_i, pfrmd^t(ag_j, cnd), blm)]$.

   - meaning that if the conduct cnd does not comply with the moral norm $mnrm$ and the agent ag_j performs that conduct at time t then an agent ag_i is allowed to *blame*, at any time t', the agent ag_j for performing cnd at the time t.

We remark that, among the conducts that agents may perform are *moral judgments* themselves, so that agents may be morally judged for performing moral judgments.

Also, we admit extensions of those forms (moral norms, facts, judgments and judgment rules), allowing for groups of agents substituting any of the agent arguments. For instance:

- If the collective conduct $ccnd$ complies with the moral norm $mnrm$ and the group of agents Ag performs that collective conduct at time t then an agent ag' is allowed to *praise*, at any time t', the group of agents Ag for performing $ccnd$ at the time t.

3.2 Moral Model

We call *moral model* of a society any structure of the form $MMdl = (RAgs, MNrms, MJRls, MFcts, MJdgms)$ where: RAg is a set of *agents* and *groups of agents* to which the model refers; $MNrms$ is the set of *moral rules* which are valid in the model; $MJRls$ is the set of *moral judgment rules* (see Sect. **??**) that the agents and groups of agents in $RAgs$ have adopted; $MFcts$ is a set of *moral facts* involving an agent or a group of agents in $RAgs$; and $MJdgms$ is a set of *moral judgments*, each with some agent or group of agents of $RAgs$ assigning some *moral value* (praise or blame) to some moral fact. As mentioned above, we require $MJdgms \subseteq MFcts$, so that moral judgments may be applied to moral judgments.

We let each agent ag (or group of agents Ag) develop its own moral model $MMdl_{ag}$ (or $MMdl_{Ag}$), referring such model to any set $RAgs_{ag}$ (or $RAgs_{Ag}$), of its own discretion.

[4] We leave formally undefined, here, the condition of a conduct complying with a moral norm.

Of course, regarding the epistemic structure M_O^R of the society, the *knowledge embedded in a moral model* is of the *relativistic* kind, both in what concerns the existence of agents and groups of agents (in $RAgs$) and moral norms (in $MNrms$), and in what concerns the occurrence of facts (in $MFcts$) and moral judgment rules (in $MJRls$).

For instance, an agent ag may have developed a moral model $MMdl_{ag} = (RAgs_{ag}, MJRls_{ag}, MNrms_{ag}, MFcts_{ag}, MJdgms_{ag})$ embodying a relativistic moral knowledge such that, in $s \in S_O$, and from the point of view of the external observer O:

- $(M_O^R, s) \models_{ag} \mathcal{K}_{ag}^R(\{ag_1, Ag_2\} \subseteq RAgs_{ag})$
  - meaning that in the state s, from the point of view of ag, there are an agent ag_1 and a group of agents Ag_2 in the reference set $RAgs_{ag}$;
- $(M_O^R, s) \models_{ag} \mathcal{K}_{ag}^R(asgn^{t'}(ag_3, prfm^t(ag_2, cnd_1), blm) \in MAsgns_{ag})$
  - meaning that, in the state s, from the point of view of ag, it happened that, at time t', agent ag_3 blamed agent ag_2 for having realized the conduct cnd_1 at time t;
- $(M_O^R, s) \models_{ag} \mathcal{K}_{ag}^R(mrl_1 \in MRls_{ag})$
  - meaning that, in the state s, from the point of view of ag, there is a moral rule mrl_1 in the set $MRls_{ag}$ of moral rules that are applicable to the agents and groups of agents in the reference set $RAgs_{ag}$.

3.3 Moral Judgments and Moral Conflicts

We call *moral judgment* any application of a *moral judgment rule* to the realization of a conduct by an agent or group of agents, the result of the moral judgment being the assignment of a *moral value* to the realization of that conduct.

Whenever an agent ag_1 makes use of the moral judgment rule $mjrl$ to perform, at time t', a moral judgment of a conduct cnd realized by an agent ag_2 at time t, the agent ag_1 changes its current moral model $MMdl_{ag_1}$, by including:

- the agent ag_2 in the set Ags_{ag_1}, if it was not there already;
- the moral fact $prfrm^t(ag_2, cnd)$ in the set $MFcts_{ag_1}$, if it was not there already;
- the moral judgment $asgn^{t'}(ag_1, prfm^t(ag_2, cnd), mv)$ in the set $MJdgms_{ag_1}$, where $mv = blm$ if the judgment resulted in a blame, and $mv = prs$ if it resulted in a praise.

However, we require, for the agent ag_1 to be able to perform such judgment, that the moral judgment rule $mjrl$ already belonged to the set $MJRls_{ag_1}$, at the time t'.

We say that there is a *moral conflict* between two moral rules, regarding a given conduct, if the rules are *contradictory* to each other, that is, if one *permits* or *obliges* the conduct while the other *forbids* it.

3.4 Group Identity, Moral Prejudice, Moral Contradiction

As mentioned above, *moral prejudices* arise from treating individual agents on the bases of judgments founded not on moral models of the individual agents themselves, but on moral models of the groups of agents to which those individual agents appear to belong (to the eyes of the moral modeler that performs the judgment).

Such *transference of moral models of groups of agents* to *individual agents* that seem to belong to them requires that groups of agents be morally modeled in terms of *stereotypical conducts* that their members appear to be used to perform (to the eyes of the moral modeler).

The *set of stereotypical conducts* that a moral modeler assigns to a group of agents constitutes a means to characterize the group, a way for the moral modeler to distinguish that group among other groups of agents, that is, an *assigned group identity*.

Moral prejudices arise, then, when an agent judges another agent on the basis of an identity assigned to a group to which the former considers the latter to belong.

To accommodate this notion of *morally assigned group identity*, we may extend the moral models with a component $GIds$, such that for each group of agents Ag in the reference set $RAgs$, one or more tuples of the form (Ag, id_{Ag}) may be present in $GIds$, where the group identity id_{Ag} should be construed as a set of conducts considered by the moral modeler to be *typical* of the members of the group Ag.

With such addition, *moral prejudices* may be explained in terms of an operation of *substitution of conducts*, by which an individual agent is morally judged not by the particular conduct (with its precise characteristics, etc.) that it has performed, or intends to perform, but by a *stereotypical conduct* that is substituted for it, a conduct that is considered to be typical of the group of agents to which that agent is considered to belong.

On the other hand, we define a *moral contradiction* between two agents or groups of agents as a *conflict between moral judgments* made by such agents or groups of agents, on the basis of a moral conflict (objective or not) between them.

Since moral judgments are, in principle, *relativistic* judgments, moral contradictions can arise as *objective* issues, between given agents or groups of agents, only when their points of view are *externalized* and *objectified*: when they constitute their relative points of view as objective.

Only then one can characterize a moral contradiction arising from a moral contradiction as an *objective moral contradiction*.

4 The Embedding of Agent Societies in Human Social Contexts

Agent societies can operate in a *stand alone* fashion and, as any other type of *isolated* society, can develop its epistemic structure, and the moral system that it supports, in ways that are uncompromised by external conditions.

Whenever an agent society is *embedded* in a given *human social context*, however, its epistemic structure and the moral system that it supports necessarily have to take into account the points of view (both epistemic and moral) of the human agents and groups of human agents that constitute that human social context.

Moreover, when that agent society operates as an intermediary between different human groups, the agents and the groups of agents of the agent society *necessarily have to take into account* the possibility of the *externalization* of the relativistic points of view of the human agents and human groups, because those externalizations are the *objective condition* for the rise of moral contradictions among those human groups.

Figure 1 illustrates the situation of a particular agent society which embedded in a particular human social context, with interactions between humans and agents, and some accesses to the moral models that are taken to be common to all the agents of each society.

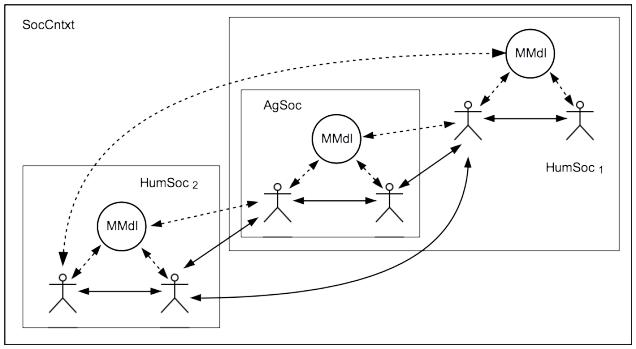

Figure 1. Agent society embeeded in a human social context.

5 The Notion of Moral Design of an Agent Society

By *moral design* of an agent society, we mean the *provision of architectural means* to support the agents and groups of agents of the agent society in their handling of moral issues (specially moral contradictions and moral contradictions).

Similarly to the *legal design* of agent societies [7], the *moral design* of agent societies belongs to the *design of the culture* of the agent society [5], and so belongs to various domains of its *architectural design* (organizational structure, symbolic environment, etc.).

In particular, it belongs to the design of the *normative system* [2] of the agent society, as the moral system is a part of the normative system of the society. Also, it belongs to the design of the *organizational intelligence* and of the *information distribution constraints* [3] of the society.

6 Conclusion

As argued in several ways by several authors (see, e.g., [1]), the *social processes of knowledge construction* are strongly conditioned by the social and historical contexts in which they occur, contexts that vary widely in time and space among different societies, and even among different social groups within a single society. So, any approach to the issue of the social construction of *moral knowledge* has to deal with the issue of *epistemic relativity*.

In this paper, we have explored in a preliminary way a formalization of the notion of *moral relativity* in agent societies, taking a particular formalization of the notion of *epistemic relativity* as its foundation.

Formal moral concepts (of *knowledge, model, judgment, prejudice, contradiction, contradiction*, *morally-based assignment of group identity*, etc.) were introduced to capture moral issues that can arise in agent societies.

Also, the paper introduced the notion of *moral design* of agent society. Moral design should be a concern specially in regard to agent societies that are embedded in human social contexts that involve a variety of *externalized* and *objectified moral models* of individuals and social groups, and that are, thus, prone to produce *objective moral contradictions* and *objective moral contradictions*.

Although we have not touched the issue in the present paper, it should be clear that the moral design of an agent society should tackle also the definition of the *content* of the moral system of the society, and should proceed hand-in-hand with the *moral design of the agents* themselves (see, e.g., [4], for the latter issue).

Finally, it should also be clear that, when considering such embedded agent societies, *moral models* (in the sense introduced here) should be articulated with *legal models* (in the sense proposed, e.g., in [6] and, more extensively, in [7]).

REFERENCES

[1] Peter L. Berger and Thomas Luckmann, *The Social Construction of Reality - A Treatise in the Sociology of Knowledge*, Anchor Books, New York, 1966.

[2] Guido Boella, Leendert van der Torre, and Harko Verhagen, 'Introduction to normative multiagent systems', *Computational and Mathematical Organization Theory*, **12**, 71–79, (2006).

[3] Karen Carley and Les Gasser, 'Computational organization theory', in *Multiagent Systems: A Modern Approach to Distributed Artificial Intelligence*, ed., Gerhard Weiss, 299–330, MIT Press, Cambridge, (1999).

[4] Helder Coelho and Antônio Carlos Rocha Costa, 'On the intelligence of moral agency', in *14th Portuguese Conference on Artificial Intelligence - EPIA'2009/Social Simulation and Modelling - SSM 2009*, pp. 439–450. University of Aveiro, (2009).

[5] Antônio Carlos Rocha Costa. The cultural level of agent societies. Invited talk at WESAAC 2011 - 5o. Workshop-School of Agent Systems, their Environments, and Applications. Curitiba, Brazil. Proceedings, 2011. (In Portuguese).

[6] Antônio Carlos Rocha Costa. On the legal aspects of agent societies. *Open publication on* www.ResearchGate.net - DOI: 10.13140/2.1.4345.7923, 2014.

[7] Antônio Carlos Rocha Costa, 'Situated legal systems and their operational semantics', *Artificial Intelligence & Law*, **43**(1), 43–102, (2015).

[8] Antônio Carlos Rocha Costa and Graçaliz Pereira Dimuro, 'A minimal dynamical organization model', in *Hanbook of Multi-Agent Systems: Semantics and Dynamics of Organizational Models*, ed., V. Dignum, 419–445, IGI Global, Hershey, (2009).

[9] Antônio Carlos da Rocha Costa, 'Relativismo epistêmico em sociedades de agentes: Uma modelagem semântica preliminar', in *Anais do Workshop-Escola de InformáticaTeórica - WEIT 2011*, pp. 122–133. UFPEL, (2012). (in Portuguese).

[10] Joseph Y. Halpern, 'Using reasoning about knowledge to analyze distributed systems', *Annual Review of Computer Science*, **2**, 37–68, (1987).

[11] Joseph Y. Halpern and Y. Moses, 'Knowledge and common knowledge in a distributed environment', in *Proc. 4th ACM Symposium on Principles of Distributed Computing*, pp. 50–61, (1984).

[12] Jaakko Hintikka, *Knowledge and Belief : An Introduction to the Logic of the Two Notions*, Cornell University Press, New York, 1962.

[13] Hans Kelsen, *Pure Theory of Law*, The Law Book Exchange, New Jersey, 2009.

[14] Émile Durkheim, 'Introduction à la morale', *Revue Philosophique*, **89**, 81–97, (1920).