Principal and Helper: Notes on Reflex Responsibility in MAS

Clara Smith[1]

Abstract. What justifies -in the head of another agent different from the one acting- the obligation to compensate is the fact that the principal agent has lengthen its own action through the implementation of a *foreign activity* for its own interests. We present two basic modal operators for representing, respectively, intentions in the interest of another agent and agency in the interest of another agent. They appear useful enough for characterizing the notion of reflex responsibility in a multi-modal multi-agent system (MAS) context.

1 MOTIVATION AND AIMS

As pointed out by Chopra and White [1], theorizing in domains such as legal and cognitive status of agents is crucial for designers of agents, especially, for the design of "on demand" MAS. Within such engineering account, a legal question arises: do the designed agents are to be autonomous enough to have rights and responsibilities? (By being autonomous we *at least* mean that agents act to achieve their own goals cf. Conte and Castelfranchi [2].)

Most works on the topic are centered on "contractual issues" (see e.g. [3,4]). Chopra and White point out four approaches as one moves up the sophistication scale of agents: three "weak" positions, based on (i) the idea of agents as mere tools of their operators, (ii) the *unilateral offer* doctrine (a contract formed by a party´s offer plus an acceptance, stipulated in the offer), and (iii) the *objective theory of contractual intention* (a contract –usually words- is an obligation which is law to the parties, who have the intention to agree), plus a fourth, radical one, which involves treating artificial agents as the legal agents of their operators. There is also a fifth position that postulates the legal systems treating agents as legal *persons*.

In this work we focus on the legal binding between a principal agent and a dependent agent. Particularly, we are interested on the dependent's performance that has its origin in *extra contractual* situations e.g. factual and/or occasional situations, trust, or courtesy. Examples of such bindings occur e.g. between the owner of a car -or any other device- and the one who drives it with the owner´s authorization (and without a proper title for using it), blog activities such as *twitting* in the name of another, or bidding in an auction in the interest of another: performed in the interest of a principal agent. In these situations is enough that the principal wills to be bind to third parties through the helper's or dependent's performance.

We therefore keep apart from our analysis situations in which performance in the interest of another has a contractual basis. This because, in contracts, function for another one and subordination may be rather straightforward to identify, mainly because there is a notion of obligation involved. If an agent gives some explicit orders or instructions to another agent which acts as his helper, or if and agent is obliged through a contract in the interest of another agent (even when agents voluntarily engage into contracts because of their own utility), or if an agent *h* forms part of agent *p*'s business organization, subordination is somehow established. We therefore exclude here cases such as mandates and any conferral of a power of representation accompanied by an obligation of representation in certain ways (e.g. a cheque is a mandate from the customer to its bank to pay the sum in question.)

We give a definition for the concept of reflex responsibility between a principal agent and a helper agent, mainly inspired on general provisions settled by Italian and Argentinean provisions for *persons*. We indeed use the terms "does", "performance", and "action" as referring to persons, although it is not entirely clear for us if it is meaningful to speak of the *actions* of devices, and artificial agents within highly automates systems; possibly a term as "executes" sounds more suitable. In what follows, "does" has the usual expected anthropomorphic meaning. The definitions we give may be useful as a step towards a specific notion of reflex responsibility of *artificial agents*.

Article 1113 of the Argentinean Civil Code states that "The obligation of the one who caused damage is extended to the damages caused by those under his dependence." In its turn, art. 1228 of the Italian Civil Code settles that "except a different will of the parties, the debtor who profits from the work of a third party for fulfilling the obligation is responsible for malicious or negligent facts carried out by that third party."

According to general doctrine and jurisprudence related to such articles, reflex responsibility has a subjective basis. This is a reason that makes reflex responsibility challenging to represent: if it had an objective basis, checking the standard legal extremes would be sufficient. Requisites for reflex responsibility are: i) the existence of a dependence relationship between principal and helper (or dependent), ii) the successful performance of an illegal action carried out by the helper, iii) that such performance was carried out while exercising a subordinate incumbency, iv) that such performance provoked a damage or injury to a third party, and that v) there must be an efficient causal relation between the helper's act and the damage caused.

Regarding the formal framework, we use as a basis a BDI multi-agent context for dealing with agents' attitudes, extended with generic obligations, as in [5]. A = {*x*, *y*, *z*...} is a finite set of

[1] Universidad Nacional de La Plata, and FACEI, Universidad Católica de La Plata, Argentina. csmith@info.unlp.edu.ar.

agents, and P = {p, q, r, ...} is a countable set of propositions. Complex expressions are formed syntactically from these, plus the following unary modalities, in the usual way: $\text{Goal}_x A$ is used to mean that "agent x has goal A", where A is a proposition. Propositions reflect particular state-of-affairs cf. B. Dunnin-Kepliçz and R. Verbrugge [6]. $\text{Int}_x A$ is meant to stand for "agent x has the intention to make A true". The doxastic (or epistemic) modality $\text{Bel}_x A$ represents that "agent x has the belief that A". The deontic operator O represents generic (legal/lawful) obligations, meaning "it is obligatory that" [7]. The operator $\text{Does}_x A$ represents successful agency in the sense given by D. Elgesem, i.e. agent x indeed brings about A [8]. For simplicity, we assume that in expressions like $\text{Does}_x A$, A denotes behavioral actions concerning only single conducts of agents such as withdrawal, inform, purchase, payment, etc. (i.e. no modalized formulas occur in the scope of a Does.) As classically established, Goal is a K_n operator, while Int and Bel are, respectively, KD_n and $KD45_n$. O is taken to be a classical KD operator. These are all normal modalities. The logic of Does, instead, is non-normal [8,9].

The rest of the paper is organized as follows. Section 2 addresses one possible characterization for a certain notion of dependence which happens to be complex enough and central to the lawful concept of reflex responsibility we deal with. We attempt four subsequent definitions, each of which improves the previous one. We go through them by using several examples. A relativized modality is introduced for dealing with oriented, coordinated intentions: an agent intends to become true a state-of-affairs A in the interest of another agent. We introduce in Section 3 another modality, a directed agency operator that binds a helper agent h to the principal agent p, and to the "oriented" action or situation A that h carries out in the interest of p. In Section 4 we formally define reflex responsibility of p regarding h with respect to an action or state-of-affairs A when: there is dependence between p and h w.r.t. A, h succeeds on carrying out A on account of p, such action constitutes an illegal act, and there is a damage a third agent t suffers which is attributable to h´s performance of A on account of p. Section 5 presents the underlying logical structure and the corresponding semantics. Conclusions end the paper.

2 DEPENDENCE

A requisite for reflex responsibility to hold is the dependence between the author of the harmful act and the agent to whom the responsibility is attributed by reflex, i.e. the principal. Such relation has two constitutive elements: 1) there is a *function* the helper carries out, on the principal´s utility; and 2) the helper is a subordinate of the principal w.r.t. the performance of such function, i.e. there is a subordinate incumbency.

Examples and non-examples of dependence relations. There is dependence between the owner of a car –or other device- and the one who drives it with the owner´s authorization (and without a proper title for using it), some blog activities such as twitting in the interest of another, or bidding in an auction in the interest of another. There is dependence between an artificial helper agent that occasionally accesses my email account profile and transfers part of its content to its (yet artificial) principal agent, which performs some data mining and later shows me tuned web ads. There is no impediment for dependence when the son works under the orders of his father, or if the daughter drives the car of her mother, who is being transported in it (there is occasional dependence). Neither parental relationships nor marriage is an impediment for the configuration of a dependence relationship.

Here then, dependence excludes delegation or mandate. There is no dependence between the car owner and the car-shop where the car is left in order to be repaired, except if the owner has authorized its use; neither between a student of a public school and the State, neither between the owner of a field and the firm in charge of its fumigation (all examples according to jurisprudence in [10].)

Definition 1. Dependence is a relation that holds according to certain internal states of agents. Let p be the principal agent and h the helper agent. Let A be a single behavioral action (e.g. pay, bid, tweet, etc.). A plausible initial characterization of dependence between p and h regarding A is: A is one of p's goals, p has the intention that h indeed carries out A, and h intends to make A true believing (knowing) that A is one of p's goals:

$$\text{Dep}^p_h A \equiv \text{Goal}_p A \wedge \text{Int}_p(\text{Does}_h A) \wedge \wedge \text{Int}_h A \wedge \text{Bel}_h(\text{Goal}_p A)\ . \qquad (1)$$

Discussion. h adopts p's goal (A) as its own intention ($\text{Int}_h A$) in the exercise of, e.g., courtesy. Based on this fact, h will carry out A. Note that the last two conjunctors in (1) are meant to capture the idea of "function for another one" (h intends to become A true because he knows it is p's goal).

Nonetheless, (1) holds when it happens to be no subordination, or is merely a coincidence, or p would like that h does A and h does it for other reasons. For instance, (1) holds in a situation where p and h are -rather than principal and helper- rivals involved in a competitive scenario, i.e. both effectively having the same goal and aiming to fulfilling it.

Example 1. **The Bach Double Concerto.** Consider the two violinists' example in [6] where two violinists intend to perform the two solo parts of the Bach Double Concerto. (The Concerto for 2 Violins, Strings and Continuo in D Minor is characterized by the subtle yet expressive relationship between the violins throughout the work.) Let us revisit the example: suppose *Peter* is a violinist who has as goal being one of the soloists. Moreover, *Peter* also has the intention that *Helen*, his past fiancée -who is also a violinist- plays as the other soloist (he would like that). But as far as *Helen* goes, she intends to become one of the chosen soloists without care of who the other soloist is (and whatsoever part she plays); nonetheless, for sure she knows that *Peter* aims to play himself as a soloist too. We get that $\text{Goal}_{peter} play \wedge \text{Int}_{peter}(\text{Does}_{helen}\ play) \wedge \text{Bel}_{helen}(\text{Goal}_{peter}\ play) \wedge \text{Int}_{helen}\ play$ holds although there is no dependence between *Peter* and *Helen* (assume that they currently have no relationship at all!): *Helen* is in competence with *Peter*.

Let us attempt an improvement for our definition.

Definition 2. There is dependence between p and h regarding A when p has A as goal, h believes on this, and such p's goal is what induces h to have the intention to carry out A:

$$\mathrm{Dep}^{p}_{h} A \equiv \mathrm{Goal}_{p} A \wedge \mathrm{Bel}_{h}(\mathrm{Goal}_{p} A) \wedge \\ \wedge (\mathrm{Goal}_{p} A \rightarrow \mathrm{Int}_{h}(\mathrm{Does}_{h} A)) \ . \tag{2}$$

Discussion. The conditional here is meant to specify that p's goal is the motive for h's intention.

Expression (2) may even hold in a rivalry scenario such as the Bach Double example. Suppose that *Helen*, knowing that *Peter* has as goal being one of the soloists, triggers her own interest in being a soloist, due to her competitive personality (and not based in any interest in *Peter*). Note also that in (2) it is sufficient that p has a goal, and that it is not necessary that the he wants h to be engaged. Then (2) also holds in a scenario where p does not want to be helped by h.

Example 2. The unwanted helper. I want my netbook to be fixed, but not by Harry who is incompetent; Harry, who does the job, satisfies my goal and qualifies as a helper.

Let us attempt a further improvement.

Definition 3. h's action will be triggered on the basis of p's intention that h does A (and h is aware of this), and not merely based on p's goal:

$$\mathrm{Dep}^{p}_{h} A \equiv \mathrm{Goal}_{p} A \wedge \mathrm{Int}_{p}(\mathrm{Does}_{h} A) \wedge \\ \wedge (\mathrm{Bel}_{h}(\mathrm{Int}_{p}\mathrm{Does}_{h} A)). \tag{3}$$

Discussion. Harry would not qualify as a helper under this definition, because I do not have the intention that he repairs my netbook (he will not carry out the task on my utility, I do not want him to).

Unfortunately, (3) still holds under rivalry between p and h w.r.t. goal A. (For a more *artificial agents*' scenario, assume any state-of-affairs in which automatic allocation of resources is in permanent dispute, and devices are not necessarily dependent one of each other.)

We next attempt a new definition that excludes rivalry situations by introducing a primitive, relativized operator, that coordinates two agents to an intention with regard to A. Binding p with h trough an "oriented" intention is what we need to exclude competitive situations.

Definition 4. Intention in the interest of another. We define a relativized operator: $\mathrm{Int}^{p}_{h} A$, meaning "h intends A to be true in the interest of p". This way, we model dependence as a *coordinated relation*, as follows. The principal indeed must have the intention that the helper performs the task, while the helper is aware. He will somehow be "activated" not only by the belief that the principal intends that s/he does the task but also with his own "oriented" intention, in the interest of p, to carry out A. Formally:

$$\mathrm{Dep}^{p}_{h} A \equiv \mathrm{Goal}_{p} A \wedge \mathrm{Int}_{p}(\mathrm{Does}_{h} A) \wedge \\ \mathrm{Bel}_{h}(\mathrm{Int}_{p}\mathrm{Does}_{h} A) \wedge \mathrm{Int}^{p}_{h}(\mathrm{Does}_{h} A) \tag{4}$$

which stands for "A is one of agent p's goals, and p intends that h performs A; h is aware of this, and intends to become A true in the interest of p".

$\mathrm{Int}^{p}_{h} A$ allows capturing custom or courtesy behavior: h may be an altruistic agent not expecting any reward, merely intending to fulfill p's expectations, even occasionally. Observe that (4) indeed reflects the power of the intention in the interest of another, as such "directed" intention defines dependence as an oriented, coordinated, non-competitive relation.

Improvements regarding the intensional basic operators have already been addressed through e.g. the concept of deadline intentions and deadline beliefs. [11,12]. For example, suppose that agent y does not believe that agent x is travelling, and says "I won't believe he is travelling until he shows the ticket to me": we write a deadline *belief* using the until operator as $\mathrm{U}(\mathrm{Does}_{x}\mathrm{ShowsTicket},\neg\mathrm{Bel}_{y}\mathrm{Travels})$ [12]. Moreover, collective intention operators for mutual and common intentions have been designed based on the basic Int operator in [6]. Relativised obligations to bearers and counterparties are defined in [13].

3 ACTION ON ACOUNT OF ANOTHER

Another requisite for the emergence of reflex responsibility is that law is violated (a legal aspect has now emerged.) The illegal act must be imputable to the helper, who is the one who materially and effectively *acts,* therefore he becomes materially responsible for the forbidden act. For the reflex responsibility to raise it is essential that the helper agent carries out the harmful activity *on account of* the principal.

We have gone through the discussion on directed intentions. It must be clear at this point that we also need an oriented/directed agency operator for coordinating h, p and the proper "oriented" action h carries out in the interest of p. Let us illustrate with an example.

Example 3. The truck driver. *d,* the occasional driver of p's truck, takes the truck off from *p's* garage on Sunday afternoon, with a view to have a ride with his friends. Due to his misguidance, his friends are injured on the occasion of this Sunday drive ($\mathrm{Does}_{d}\, \mathit{drive} \rightarrow \mathrm{injure}_{\mathrm{friends}}$).

Discussion. p has as goal that d drives his truck, and intends him to drive it, d believes in this, and d has the intention to drive the truck in the interest of p. So we get dependence between h and p regarding A (i.e. (4) holds). Now, note that provided the general obligation that states that we should not harm others ($\mathrm{O}\neg\mathrm{injure}_{t}$), *p´s* reflex responsibility is about to raise. But d drove in its own interest. What justifies in the head of another agent -different from the one acting- an obligation to compensate is that the principal agent has lengthen its own action through the implementation of a *foreign activity* for its own interests. Here is not the case, *d* drove on his own account when he provoked the accident.

We know that it is essential for reflex responsibility to hold that the performing agent carries out the task on account of another agent. We should be able, then, to distinguish those directed intentions and actions that we make in our own interest from those which we do in the interest of another one.

Definition 5. Agency in the interest of another. We introduce a relativized operator $\mathrm{Does}^{p}_{h}\, A$ to represent agency in the interest of another, meaning "*h* carries out *A* in the interest of *p*". This non-normal operator is meant to capture *performance for another one* i.e. directed material performance in the head (and/or hands, or executable code) of *d,* but on account of *p*. This way, we establish oriented agency as a basic type of event, the same way as Does is.

This relativised agency operator leads us to a more precise definition for dependence:

$$\mathrm{Dep}^{p}_{h}\, A \equiv \mathrm{Goal}_{p}\, A \wedge \mathrm{Int}_{p}(\mathrm{Does}^{p}_{h}\, A) \wedge \\ \wedge\, (\mathrm{Bel}_{h}(\mathrm{Int}_{p}(\mathrm{Does}^{p}_{h}\, A)) \wedge \mathrm{Int}^{p}_{h}\, A\ . \qquad (5)$$

Back to the truck example, we have that, that Sunday, $\mathrm{Does}^{d}_{d}\mathit{drive}$ holds and also $\neg(\mathrm{Int}_{p}(\mathrm{Does}^{p}_{d}\ \mathit{drive}))$ holds, making (5) false. (Note that, intuitively, $\mathrm{Does}^{d}_{d}\, A$ collapses to $\mathrm{Does}_{d}\, A$.)

4 REFLEX RESPONSIBILITY

We saw that another requisite for the emergence of reflex responsibility is that the helper's harmful performance provokes a damage or injury to a third party, let us say *t*, and that there must be an efficient causal relation between *h*'s performance -on account of *p*- and the damage caused to *t*: $\mathrm{Does}^{p}_{h}\, A \rightarrow \mathrm{Damage}_{t}$, with $t \neq h \neq p$.

We are now in a position to define reflex responsibility.

Definition 6. Reflex Responsibility. There is reflex responsibility of agent *p* regarding agent *h* w.r.t. the action or state-of affairs *A* when there is dependence between *p* and *h* w.r.t. *A*, *h* succeeds regarding *A* on account of *p*, such performance is an illegal act, and there is a damage *t* suffers, which is attributable to *h*'s performance:

$$\mathrm{Reflex}^{p}_{h}\, A \equiv \mathrm{Dep}^{p}_{h}\, A \wedge \mathrm{Does}^{p}_{h}\, A \wedge \\ \wedge\, \mathrm{O}\neg\, A \wedge (\mathrm{Does}^{p}_{h}\, A \rightarrow \mathrm{Damage}_{t})\ . \qquad (7)$$

Discussion. According to the analysis done in [14], reflex responsibility belongs to the category of: (i) *blameworthiness* responsibility, meaning that the principal failed to comply with the demands of the system i.e. being faulty according to the system (because $\mathrm{Does}^{p}_{h}A$ and $\mathrm{O}\neg A$); and also to the category of: (ii) *accountability* responsibility, because the principal has a particular connection to the harm (the harm can be linked to the principal) so that he has to give an explanation (an account) why the harm happened, and, of course, he may possibly be sued. According to [14], when (7) holds we can say that *p* is legally liable for the harmful event because all conditions for connecting the harm to that person are realized: note that both dependence and directed action connect *p* to the harm and thus lead to *p*'s liability.

Another relevant issue is that the responsibility of the dependent must be established before declaring the principal's responsibility by reflex. Only in a second moment the reflex can be settled. Consequently, we cannot conceive a case where the principal is responsible but the dependent is not. The exclusion of the dependent's responsibility excludes the principal's responsibility:

$$\neg(\mathrm{Does}^{p}_{h}\, A \rightarrow \mathrm{Damage}_{t}) \rightarrow \neg\mathrm{Reflex}^{p}_{h}\, A\ . \qquad (8)$$

An important ingredient for delimiting the application of reflex responsibility is the consciousness (awareness) that the injured third party has w.r.t. the fact that the helper acted beyond the subordinate incumbency. In this case, we may consider that the comitent has no responsibility even when the injuries possibly have been inflicted with devices entrusted to the helper just for being so. For example, if *d*'s friends know it is *p*'s truck (and not *d*'s truck), *p* is not to be liable. We write this limit as:

$$(\mathrm{Bel}_{t}(\neg\mathrm{Int}_{p}(\mathrm{Does}^{p}_{h}A)) \wedge (\mathrm{Does}^{p}_{h}\, A \rightarrow \mathrm{Damage}_{t})) \rightarrow \neg\mathrm{Reflex}^{p}_{h}\, A \qquad (9)$$

Also, recall that if it happens that *d* is the injured party (i.e. suppose for a moment that *d*=*t* in (7)) general provisions regarding negligence and incompetence exclude any *d*'s attempt to sue *p*.

If the harmed third party *t* is bound to the principal by means of a contract (e.g. it holds that it is obligatory for *p* in the interest of *t* that *A*: $\mathrm{O}^{t}_{p}\, A$), and the dependent's harmful performance imports the non-execution of obligations assumed by the principal w.r.t. the third party ($\mathrm{Reflex}^{p}_{h}\, A \rightarrow \neg A$), then such non-execution is imputable to the principal (contract beats reflex): here we have entered the contractual arena in which any faulty act of the subordinate is imputed to his principal, *p*. The solution is thus beyond the reflex responsibility approach. Formally we may write:

$$(\mathrm{O}^{t}_{p}\, A \wedge (\mathrm{Reflex}^{p}_{h}\, A \rightarrow \neg\, A)) \rightarrow \mathrm{O}^{t}_{p}\ \mathrm{Compensate}\ . \qquad (10)$$

Finally, if the harmed party is the principal, the dependence relationship becomes irrelevant and cannot be used as *d*'s excuse or exception: *d* is to be sued according to general rules.)

One more remark. G. Sartor et al. briefly outline in [14] the notion of *vicarious liability* in tort law. "Vicarious" refers to the idea of one person being liable for the harm caused by another. In that work, it is pointed out that Anglo/American law does not provide a general formula to deal with the requirement that the liability of the principal *p* is based on whether the servant committed the tort in the course of his duty; moreover, an "inner connection" is needed between the harmful act and the task asked by *p*.

Complex situations can be designed with the aid of a definition such as the one given here for reflex responsibility, when we use it as a building block. It may lead us to an interesting and high level of sophistication in the devise and outline of the lawful support of a system.

Example 4. Reflex responsibility and trust deception. Paul lends to me his user name and password, so as I can use the wireless connection at his university, which I am visiting. I made wrong use of some contents, a database damaged, and I –under Paul's user name- got blacklisted. Paul trusted me, now he is responsible by reflex for my misuse. His trust on me is connected to his responsibility, which for sure is now deceived with independence of the case that he manages to give an adequate explanation to whom he had to respond in order to be erased from the blacklist.

5 SEMANTICS

The semantics for this logics of reflex responsibility is based on a multi-relational frame *F*, with the following structure [5]:

$$F = <\mathrm{A}, \mathrm{W}, \{\mathrm{B_i}\}_{\mathrm{i}\in\mathrm{A}}, \{\mathrm{G_i}\}_{\mathrm{i}\in\mathrm{A}}, \{\mathrm{I_i}\}_{\mathrm{i}\in\mathrm{A}}, \{D_\mathrm{i}\}_{\mathrm{i}\in\mathrm{A}}>$$

where:

- A is the finite set of agents;
- W is a set of possible worlds;
- $\{\mathrm{B_i}\}_{\mathrm{i}\in\mathrm{A}}$ is a set of accessibility relations w.r.t. beliefs, which are transitive, euclidean and serial;
- $\{\mathrm{G_i}\}_{\mathrm{i}\in\mathrm{A}}$ is a set of accessibility relations w.r.t. goals; with standard $\mathrm{K_n}$ semantics;
- $\{\mathrm{I_i}\}_{\mathrm{i}\in\mathrm{A}}$ is a set of accessibility relations w.r.t. intentions, which are serial;
- $\{D_\mathrm{i}\}_{\mathrm{i}\in\mathrm{A}}$ is a family of sets of accessibility relations $\mathrm{D_i}$ wrt Does, which are pointwise closed under intersection, reflexive and serial [5].

Recall that we want to be able to represent directed intentions and directed actions; we should also be able to represent generic obligations. Therefore we introduce slight modifications extending *F*: the underlying structure for supporting reflex responsibility is a variant of *F*, call it *R*:

$$R = < \mathrm{A}, \mathrm{W}, \{\mathrm{B_i}\}_{\mathrm{i}\in\mathrm{A}}, \{\mathrm{G_i}\}_{\mathrm{i}\in\mathrm{A}}, \{\mathrm{I^j_i}\}_{\mathrm{i,j}\in\mathrm{A}}, \{D^\mathrm{j}_\mathrm{i}\}_{\mathrm{i,j}\in\mathrm{A}}, O >$$

where:

- $\{\mathrm{I^j_{i.}}\}_{\mathrm{i,j}\in\mathrm{A}}$ is a set of accessibility relations w.r.t. the notion of relativized intention, meaning that there is an I relation for each combination of *i*s and *j*s (which are serial); and
- $\{D^\mathrm{j}_\mathrm{i}\}_{\mathrm{i,j}\in\mathrm{A}}$ is a family of sets of accessibility relations $\mathrm{D^j_i}$ w.r.t. oriented actions, meaning that there is a set for each combination of *i*s and *j*s, which are pointwise closed under intersection, reflexive and serial; and
- *O* is the accessibility relation for the deontic modality O for obligations, which is serial (standard KD semantics).

Note that if we are to represent formulas such as (10) we also need to include modalities for relativised obligations (standard KDn semantics.)

In its turn, a multi-relational model is a structure $M = <R,\mathrm{V}>$ where *R* is a multi-relational frame as above, and V is a valuation function defined as follows:

1. standard Boolean conditions;
2. $\mathrm{V}(w, \mathrm{Bel_i}\, A) = 1$ iff $\forall v$ (if $w\ \mathrm{B_i}\ v$ then $\mathrm{V}(v, A) = 1$);
3. $\mathrm{V}(w, \mathrm{Goal_i}\, A) = 1$ iff $\forall v$ (if $w\ \mathrm{G_i}\ v$ then $\mathrm{V}(v, A) = 1$);
4. $\mathrm{V}(w, \mathrm{Int^j_i}\, A) = 1$ iff $\forall v$ (if $w\ \mathrm{I^j_i}\ v$ then $\mathrm{V}(v, A) = 1$).
5. $\mathrm{V}(w, \mathrm{Does^j_i}\, A) = 1$ iff $\exists \mathrm{D^j_i}\in D^\mathrm{j}_\mathrm{i}$ such that $\forall v$ ($w\ \mathrm{D^j_i}\ v$ iff $\mathrm{V}(v, A) = 1$);
6. $\mathrm{V}(w, \mathrm{O}\, A) = 1$ iff $\forall v$ (if $w\mathrm{O}v$ then $\mathrm{V}(v, A) = 1$);

Decidability for the logics for *R* follows directly from [15, 16]. The logics for *F* was there reorganized as a fibring in [15], this is a particular combination of logics which amounts to place one logics on top of another. In the case of *F*, the normal logic was put on top of the non-normal one. By exploiting results in regard to techniques for combining logics, it was proved in [15] that that fibred logics is complete and decidable. Therefore, we only have to extend the proofs in [15] for the new modalities in *R*.

In its turn, [16] gives a new presentation for existing theorems generalizing to neighborhood structures the well-known results regarding decidability through filtrations for Kripke structures. *F* is a special case of [16, Def. 5] because its semantics can be outlined within a neighborhood approach. Therefore it is straightforward to prove decidability for its extension *R*.

6 FINAL REMARKS

In this work we attempt to provide one step towards the issue of 'rational automatic allocation of liability' [14] within MAS. In particular, we focus on a possible logical formalization of situations and state-of-affairs where a principal agent wills to be bind to a helper for achieving his goals.

Clearly, whether one decides to include –or not- in the system the automatic detection of reflex responsibility, depends on the interest on lawfully distinguishing between principal and helpers' separate responsibilities. Such a distinction has an impact on the concept of liability underlying the system and, possibly induced by this fact, on the issue of efficient distribution of available resources among agents, due to sanctions such as obligations to repair harm. Moreover, distinguishing between helpers and principals allows to the system's users and to other agents to e.g. recognize which agent is to be sued for wrongdoing.

In the words of M. Sergot [17], it has been suggested –from, let us say the last twenty years- that interactions among multiple, independently acting artificial agents can be effectively regulated and managed by norms (or 'social laws') which, if respected, allow the agents to co-exist in a shared environment. This article attempts an answer to his question of what happens to the system behavior when 'social laws' are not respected. In our present outline, trust, altruistic, and courtesy behavior can be seen as social predispositions that may induce occasional dependence between agents, generating a bound between them, and possibly establishing a reflex responsibility. The usual expected behavior is that the entrusted agent should behave according to accepted standards, *acting good*. When this principle is broken, there is a need of lawfully repairing the wrongdoing.

From the logical viewpoint, the structure of the systems outlined in this work is a simple combination of normal and non-normal modalities. Nonetheless, the structure is suitable for representing sophisticated relationships such as occasional dependence, bridges between trust and responsibility, and bindings between agreements (such as contracts) and dependence. The logical simplicity is also a support for their usefulness and robustness, and also keeps systems manageable and suitable for further extensions.

At least two issues are left open. First, if it can be argued that artificial agents act in the same sense humans do; in particular, if they can will to be bind by other agent's performance, have directed intentions, and perform actions in the interest of another one. Second, provided that the reflex responsibility is, in this paper, allocated by the system, what are its consequences or impact on the agents' reactions. For example, what will Paul do and how will he behave from now on, now that he has been proved responsible by reflex? Will he reconsider from now on his beliefs? If so, with regard to everyone, or just with regard to me? This topic leads to the study of what G. Sartor et al. call the social consequences that are induced by allocating liabilities [14].

Finally, we are to explore more in depth the relationship between reflex responsibility and trust.

REFERENCES

[1] S. Chopra, L. White. Artificial Agents – Personhood in Law and Philosophy. ECAI, pages 635–639, 2004.

[2] Conte, R. Castelfranchi, C. Cognitive and social action. UCL Press Ltd., 1995.

[3] T. Allan, R. Widdison. Can computers make contracts? *Harvard Journal of Law and Technolgy*, 9, 25-52, 1996.

[4] I. Kerr. Ensuring the success of contract formation in agent-mediated electronic commerce. *Electronic Commerce Research*, 1 (1/2), 183-202, 2001.

[5] C. Smith, A. Rotolo. Collective trust and normative agents. Logic Journal of IGPL, 18(1), 195–213, (2010).

[6] B. Dunin-Keplicz, R. Verbrugge. Collective intentions. Fundamenta Informaticae, 271–295, (2002).

[7] A. Jones, M. Sergot. A logical framework. In Open agent societies, normative specification in multiagent systems, 2007.

[8] D. Elgesem, 'The modal logic of agency', Nordic Journal of Philosophical Logic, 2, 1–46, (1997).

[9] Guido Governatori and Antonino Rotolo, 'On the Axiomatization of Elgesem's Logic of Agency and Ability'. Journal of Philosophical Logic, 34(4), 403–431, (2005).

[10] J.J. Llambias. Civil Code with Annotations. Buenos Aires.

[11] Broersen, J., Dignum, F., Dignum, V., Meyer, J-J. Designing a Deontic Logic of Deadlines. LNCS 3065, 43-56. Springer, 2004.

[12] C. Smith, A. Rotolo, G. Sartor. Representations of time within normative MAS. Proceedings of the 2010 conference on Legal Knowledge and Information Systems: JURIX 2010: The Twenty-Third Annual Conference, 107-116. IOS Press Amsterdam, The Netherlands, 2010. ISBN: 978-1-60750-681-2.

[13] H. Herrestad and C. Krogh, Deontic Logic Relativised to Bearers and Counterparties, 453–522, J. Bing and O. Torvund, 1995.

[14] G. Sartor et al. Framework for addressing the introduction of automated technologies in socio-technical systems, in particular with regard to legal liability. E.02.13-ALIAS-D1.1. EUI, Firenze, 2011. http://dl.dropbox.com/u/10505513/Alias/E0213ALIASD11FramingtheProblemV015.pdf

[15] C. Smith, A. Ambrossio, L. Mendoza, A. Rotolo. Combinations of normal and non-normal modal logics for modeling collective trust in normative MAS, AICOL XXV IVR, Forthcoming Springer LNAI, 2012.

[16] C. Smith, L. Mendoza, A. Ambrossio. Decidability via Filtration of Neighbourhood Models for Multi-Agent Systems. Forthcoming proceedings of the SNAMAS 2012 @ AISB/IACAP World Congress, UK.

[17] M. Sergot. Norms, Action and Agency in Multi-agent Systems Deontic Logic in Computer Science Lecture Notes in Computer Science, 2010, Volume 6181/2010, 2, DOI: 10.1007/978-3-642-14183-6_2 .