Interval-Based Markov Decision Processes for Regulating Interactions Between Two Agents in Multi-agent Systems*

Graçaliz P. Dimuro[1] and Antônio C.R. Costa[1,2]

[1] Escola de Informática, Universidade Católica de Pelotas
Rua Felix da Cunha 412, 96010-000 Pelotas, Brazil
{liz,rocha}@ucpel.tche.br
[2] Programa de Pós-Graduação em Computação
Universidade Federal do Rio Grande do Sul
Av. Bento Gonçalves 9500, 91501-970 Porto Alegre, Brazil

Abstract. This work presents a model for Markov Decision Processes applied to the problem of keeping two agents in equilibrium with respect to the values they exchange when they interact. Interval mathematics is used to model the qualitative values involved in interactions. The optimal policy is constrained by the adopted model of social interactions. The MDP is assigned to a supervisor, that monitors the agents' actions and makes recommendations to keep them in equilibrium. The agents are autonomous and allowed to not follow the recommendations. Due to the qualitative nature of the exchange values, even when agents follow the recommendations, the decision process is non-trivial.

1 Introduction

There are many different techniques to deal with the problem of choosing optimal agent actions [11,13], some of them considering stochastic domains. The work presented in [3] deals with this problem using techniques from operations research, namely the theory of Markov Decision Processes (MDP) [5,8,12]. In this paper we introduce a qualitative version of a MDP, called *Qualitative Interval-based Markov Decision Process* (QI–MDP). The values characterizing the states and actions of the model are based on intervals and their calculation performed according to Interval Arithmetic [6]. The model is said to be qualitative in the sense that intervals are considered equivalent according to a loose equivalence relation. We apply the QI–MDP model to the analysis of the equilibrium of social exchanges between two interacting agents. The equilibrium is determined according to the balance of values the agents exchange during their interactions. The decision process pertains to a third agent, the *equilibrium supervisor*, who is in charge of giving recommendations to the agents on the best exchanges they can perform in order to keep the balance in equilibrium.

We modelled the social interactions according to Piaget's theory of exchange values [7], and derived the idea of casting the equilibrium problem in terms of a MDP from George Homans' approach to that same problem [4]. Due the lack of space, we shall

* This work was partially supported by CTINFO/CNPq and FAPERGS.

J. Dongarra, K. Madsen, and J. Waśniewski (Eds.): PARA 2004, LNCS 3732, pp. 102–111, 2006.

not consider in detail the social model based on Piaget's theory, since it was deeply explored in previous work [1,9,10]. A first application of this model in multi-agent systems was presented in [9,10]. In [1], exchange values were proposed for the modelling of collaborative on-line learning interactions.

The paper is organized as follows. The model of social interactions is presented in Sect. 2, and intervals are introduced in Sect. 3. The QI–MDP model is introduced in Sect. 4. Section 5 discusses the results. Section 6 is the Conclusion.

2 Social Reasoning About Exchange Values

The *Social Exchange Model* introduced by Piaget [7] is based on the idea that social relations can be reduced to social exchanges between individuals. Each social exchange is a service exchange between individuals and it is also concerned with an exchange of values between such individuals. The exchange values are of qualitative nature and are constrained by a *scale of values*.

A *social exchange* is assumed to be performed in two stages. Figure 1 shows a schema of the exchange stages. In the stage $\mathrm{I}_{\alpha\beta}$, the agent α realizes a service for the agent β. The values related with this exchange stage are the following: *(i)* $r_{\mathrm{I}_{\alpha\beta}}$ is the value of the *investment*[3] done by α for the realization of a service for β; *(ii)* $s_{\mathrm{I}_{\beta\alpha}}$ is the value of β's *satisfaction* due to the receiving of the service done by α; *(iii)* $t_{\mathrm{I}_{\beta\alpha}}$ is the value of β's *debt*, the debt it acquired to α for its satisfaction with the service done by α; *(iv)* $v_{\mathrm{I}_{\alpha\beta}}$ is the value of the *credit* that α acquires from β for having realized the service for β. In the stage $\mathrm{II}_{\alpha\beta}$, the agent α asks the payment for the service previously done for the agent β, and the values related with this exchange stage have similar meaning. $r_{\mathrm{I}_{\alpha\beta}}$, $s_{\mathrm{I}_{\beta\alpha}}$, $r_{\mathrm{II}_{\beta\alpha}}$ and $s_{\mathrm{II}_{\alpha\beta}}$ are called *material values*. $t_{\mathrm{I}_{\beta\alpha}}$, $v_{\mathrm{I}_{\alpha\beta}}$, $t_{\mathrm{II}_{\beta\alpha}}$ and $v_{\mathrm{II}_{\alpha\beta}}$ are the *virtual values*. The order in which the exchange stage may occur is not necessarily $\mathrm{I}_{\alpha\beta} - \mathrm{II}_{\alpha\beta}$.

Piaget's approach to social exchange was an algebraic one: what interested him was the algebraic laws that define equilibrium of social exchanges. George Homans [4] approached the subject from a different view: he was interested in explaining how and why agents strive to achieve equilibrium in such exchanges. The solution he found, based on a behavioristic explanation of the agents' decision, suggested that agents look for a maximum of profit, in terms of social values, when interacting with each other. That proposal gave the starting point for the formalization we present below, where the looking for a maximum of profit is understood as a MDP to be solved by the equilibrium supervisor.

3 Modelling Social Exchanges with Interval Values

Piaget's concept of scale of values [7] is now interpreted in terms of Interval Mathematics [6]. Consider the set $\mathbb{IR}_L = \{[a,b] \mid -L \leq a \leq b \leq L, a, b \in \mathbb{R}\}$ of real intervals bounded by $L \in \mathbb{R}$ $(L > 0)$ and let $\mathcal{IR}_L = (\mathbb{IR}_L, +, \Theta, ^{\frown})$ be a *scale of interval exchange values*, where:

[3] An investment value is always negative.

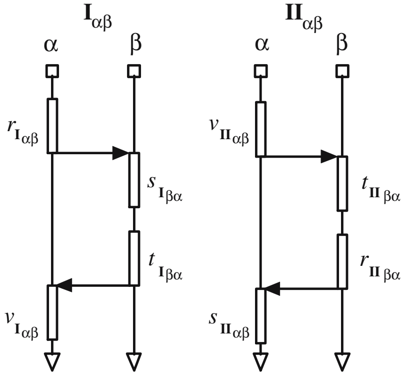
$\mathbf{I}_{\alpha\beta}$
α
β
$r_{\mathbf{I}\alpha\beta}$
$s_{\mathbf{I}\beta\alpha}$
$t_{\mathbf{I}\beta\alpha}$
$v_{\mathbf{I}\alpha\beta}$
$\mathbf{II}_{\alpha\beta}$
α
β
$v_{\mathbf{II}\alpha\beta}$
$t_{\mathbf{II}\beta\alpha}$
$r_{\mathbf{II}\beta\alpha}$
$s_{\mathbf{II}\alpha\beta}$

Fig. 1. Stages of social exchanges

(i) $+ : \mathbb{IR}_L \times \mathbb{IR}_L \rightarrow \mathbb{IR}_L$ is the *addition* operation $[a,b] + [c,d] = [\max\{-L, a + c\}, \min\{b + d, L\}]$.
(ii) A *null value* is any $[a,b] \in \mathbb{IR}_L$ such that $mid([a,b]) = 0$, where $mid([a,b]) = \frac{a+b}{2}$ is the mid point of $[a,b]$. The set of null values is denoted by Θ. $[0,0]$ is called the *absolute null value*.
(iii) A *quasi-symmetric value* for $X \in \mathbb{IR}_L$ is any $X' \in \mathbb{IR}_L$ such that $X + X' \in \Theta$. The set of quasi-symmetric values of X is denoted by $\widetilde{X}$.

$\mu\widetilde{X} \in \widetilde{X}$ is said to be the *least quasi-symmetric value* of X, if whenever there exists $S \in \widetilde{X}$ it holds that $d(\mu\widetilde{X}) \leq d(S)$, where $d([a,b]) = b - a$ is the diameter of an interval $[a,b]$. A *qualitative equivalence relation* $\approx$ is defined on $\mathbb{IR}_L$ by $X \approx Y \Leftrightarrow \exists Y' \in \widetilde{Y} : X + Y' \in \Theta$. For all $X \in \mathbb{IR}_L$, it follows that:

Proposition 1. *(i)* $\widetilde{X} = \{-[mid(X) - k, mid(X) + k] \mid k \in \mathbb{R} \wedge k \geq 0\}$*; (ii)* $\mu\widetilde{X} = -[mid(X), mid(X)]$.

Proof. $mid(X + (-[mid(X) - k, mid(X) + k])) = mid([\frac{a_1 - a_2 - 2k}{2}, \frac{a_2 - a_1 + 2k}{2}]) = 0$, for $X = [a_1, a_2]$. If $S \in \widetilde{X}$ is such that $mid(S) \neq mid(X)$, then $mid(X + S) = mid([\frac{a_1 - a_2 - 2k_2}{2}, \frac{a_2 - a_1 + 2k_1}{2}]\ ) \neq 0$, for $k_1 \neq k_2 \in \mathbb{R}$, which is a contradiction. □

For practical applications, we introduce the concept of *absolute ϵ-null value* $0_\epsilon = [-\epsilon, +\varepsilon]$, with $\epsilon \in \mathbb{R}$ ($\epsilon \geq 0$) being a given tolerance. In this case, an ϵ-null value is any $N \in \mathbb{IR}_L$ such that $mid(N) \in 0_\epsilon$. The set of ϵ-null values is denoted by Θ_ϵ. The related set of ϵ-quasi-symmetric values of $X \in \mathbb{IR}_L$ is denoted by $\widetilde{X}_\epsilon$.

Let T be a set of discrete instants of time. Let α and β be any two agents. A *qualitative interval exchange-value system* for modelling the exchanges from α to β is a structure $\mathbf{IR}_{\alpha\beta} = (\mathcal{IR}_L; r_{\mathrm{I}_{\alpha\beta}}, r_{\mathrm{II}_{\beta\alpha}}, s_{\mathrm{I}_{\beta\alpha}}, s_{\mathrm{II}_{\alpha\beta}}, t_{\mathrm{I}_{\beta\alpha}}, t_{\mathrm{II}_{\beta\alpha}}, v_{\mathrm{I}_{\alpha\beta}}, v_{\mathrm{II}_{\alpha\beta}})$ where $r_{\mathrm{I}_{\alpha\beta}}, r_{\mathrm{II}_{\beta\alpha}} : T \rightarrow \mathbb{IR}_L$, $s_{\mathrm{II}_{\alpha\beta}}, s_{\mathrm{I}_{\beta\alpha}} : T \rightarrow \mathbb{IR}_L$, $t_{\mathrm{I}_{\beta\alpha}}, t_{\mathrm{II}_{\beta\alpha}} : T \rightarrow \mathbb{IR}_L$ and $v_{\mathrm{I}_{\alpha\beta}}, v_{\mathrm{II}_{\alpha\beta}} : T \rightarrow \mathbb{IR}_L$ are partial functions that evaluate, at each time instant $t \in T$, the investment, satisfaction, debt and credit values[4], respectively, involved in the exchange. Denote

[4] The values are undefined if no service is done at all at a given moment $t \in T$.

$r_{\mathrm{I}_{\alpha\beta}}(t) = r^t_{\mathrm{I}_{\alpha\beta}}$, $r_{\mathrm{II}_{\beta\alpha}}(t) = r^t_{\mathrm{II}_{\beta\alpha}}$, $s_{\mathrm{II}_{\alpha\beta}}(t) = s^t_{\mathrm{II}_{\alpha\beta}}$, $s_{\mathrm{I}_{\beta\alpha}}(t) = s^t_{\mathrm{I}_{\beta\alpha}}$, $t_{\mathrm{I}_{\beta\alpha}}(t) = t^t_{\mathrm{I}_{\beta\alpha}}$, $t_{\mathrm{II}_{\beta\alpha}}(t) = t^t_{\mathrm{II}_{\beta\alpha}}$, $v_{\mathrm{I}_{\alpha\beta}}(t) = v^t_{\mathrm{I}_{\alpha\beta}}$ and $v_{\mathrm{II}_{\alpha\beta}}(t) = v^t_{\mathrm{II}_{\alpha\beta}}$. A *configuration of exchange values* is specified by one of the tuples $(r^t_{\mathrm{I}_{\alpha\beta}}, s^t_{\mathrm{I}_{\beta\alpha}}, t^t_{\mathrm{I}_{\beta\alpha}}, v^t_{\mathrm{I}_{\alpha\beta}})$ or $(v^t_{\mathrm{II}_{\alpha\beta}}, t^t_{\mathrm{II}_{\beta\alpha}}, r^t_{\mathrm{II}_{\beta\alpha}}, s^t_{\mathrm{II}_{\alpha\beta}})$. The sets of configurations of exchange values from α to β, for stages I and II, are denoted by $\mathrm{EV}^{\mathrm{I}}_{\mathbf{IR}_{\alpha\beta}}$ and $\mathrm{EV}^{\mathrm{II}}_{\mathbf{IR}_{\alpha\beta}}$, respectively.

Consider the functions $\mathrm{I}_{\alpha\beta} : T \to \mathrm{EV}^{\mathrm{I}}_{\mathbf{IR}_{\alpha\beta}}$ and $\mathrm{II}_{\alpha\beta} : T \to \mathrm{EV}^{\mathrm{II}}_{\mathbf{IR}_{\alpha\beta}}$, defined, respectively, by $\mathrm{I}_{\alpha\beta}(t) = \mathrm{I}^t_{\alpha\beta} = (r^t_{\mathrm{I}_{\alpha\beta}}, s^t_{\mathrm{I}_{\beta\alpha}}, t^t_{\mathrm{I}_{\beta\alpha}}, v^t_{\mathrm{I}_{\alpha\beta}})$ and $\mathrm{II}_{\alpha\beta}(t) = \mathrm{II}^t_{\alpha\beta} = (v^t_{\mathrm{II}_{\alpha\beta}}, t^t_{\mathrm{II}_{\beta\alpha}}, r^t_{\mathrm{II}_{\beta\alpha}}, s^t_{\mathrm{II}_{\alpha\beta}})$. A *stage of social exchange* from α to β is either a value $\mathrm{I}^t_{\alpha\beta}$, where $r^t_{\mathrm{I}_{\alpha\beta}}$ is defined, or $\mathrm{II}^t_{\alpha\beta}$, where $r^t_{\mathrm{II}_{\beta\alpha}}$ is defined.

A *social exchange process* between any two agents α and β, occurring during the time instants $T = t_1, \ldots, t_n$, is any finite sequence $\mathrm{s}^T_{\{\alpha,\beta\}} = e_{t_1}, \ldots, e_{t_n}$, $n \geq 2$, of exchange stages from α to β and from β to α, where there are $t, t' \in T$, $t \neq t'$, with *well defined investment values* $r^t_{\mathrm{I}_{\alpha\beta}}$ and $r^{t'}_{\mathrm{II}_{\beta\alpha}}$ (or $r^t_{\mathrm{I}_{\beta\alpha}}$ and $r^{t'}_{\mathrm{II}_{\alpha\beta}}$).

The *material results* $M_{\alpha\beta}$ and $M_{\beta\alpha}$ of a social exchange process, from the points of view of α and β, respectively, are given by the respective sum of the material values involved in the process. Considering $k^T_{\mathrm{I}_{\lambda\delta}} = \sum_{t\in T} k^t_{\mathrm{I}_{\lambda\delta}}$ and $k^T_{\mathrm{II}_{\lambda\delta}} = \sum_{t\in T} k^t_{\mathrm{II}_{\lambda\delta}}$, for all well defined $k^t_{\mathrm{I}_{\lambda\delta}}$ and $k^t_{\mathrm{II}_{\lambda\delta}}$, with $k = r, s$, then $M_{\alpha\beta} = r^T_{\mathrm{I}_{\alpha\beta}} + s^T_{\mathrm{II}_{\alpha\beta}} + r^T_{\mathrm{II}_{\alpha\beta}} + s^T_{\mathrm{I}_{\alpha\beta}}$ and $M_{\beta\alpha} = r^T_{\mathrm{I}_{\beta\alpha}} + s^T_{\mathrm{II}_{\beta\alpha}} + r^T_{\mathrm{II}_{\beta\alpha}} + s^T_{\mathrm{I}_{\beta\alpha}}$. The process is said to be in equilibrium if $M_{\alpha\beta} \in \Theta_\epsilon$ and $M_{\beta\alpha} \in \Theta_\epsilon$. If a material result of a social exchange process is not in equilibrium, then any ϵ-quasi-symmetric of $M_{\alpha\beta}$ ($M_{\beta\alpha}$) is called a *compensation* value from α's (β's) point of view.

4 Solving the Equilibration Problem Using QI–MDP

4.1 The Basics of an QI–MDP

We conceive that, in the context of a social exchange process between two agents, a third agent, called *equilibrium supervisor*, analyzes the exchange process and makes suggestions of exchanges to the two agents in order to keep the material results of exchanges in equilibrium. To achieve that purpose, the equilibrium supervisor models the exchanges between the two agents as a MDP, where the *states* of the model represent "possible material results of the overall exchanges" and the *optimal policies* represent "sequences of *actions* that the equilibrium supervisor recommends that the interacting agents execute".

Consider $\epsilon, L \in \mathbb{R}$ ($\epsilon \geq 0, L > 0$), $n \in \mathbb{N}$ ($n > 0$) and let $\hat{E} = \{E^{-n}, \ldots, E^n\}$ be the set of equivalence classes of intervals, defined, for $i = -n, \ldots, n$, as:

$$E^i = \begin{cases} \{X \in \mathbb{IR}_L \mid i\frac{L}{n} \leq mid(X) < (i+1)\frac{L}{n}\} & \text{if } -n \leq i < -1 \\ \{X \in \mathbb{IR}_L \mid -\frac{L}{n} \leq mid(X) < -\epsilon\} & \text{if } i = -1 \\ \{X \in \mathbb{IR}_L \mid -\epsilon \leq mid(X) \leq +\epsilon\} & \text{if } i = 0 \\ \{X \in \mathbb{IR}_L \mid \epsilon < mid(X) \leq \frac{L}{n}\} & \text{if } i = 1 \\ \{X \in \mathbb{IR}_L \mid (i-1)\frac{L}{n} < mid(X) \leq i\frac{L}{n}\} & \text{if } 1 < i \leq n. \end{cases} \tag{4.1}$$

The classes E^i are the supervisor representations of classes of unfavorable ($i < 0$), equilibrated ($i = 0$) and favorable ($i > 0$) material results of exchange balances.

Table 1. Specification of compensation intervals

State	Compensation Interval C^i	State	Compensation Interval C^i
$E^i_{-n\leq i<-1}$	$[-\left(\frac{2i+1}{2}\frac{L}{n}\right)-\epsilon, -\left(\frac{2i+1}{2}\frac{L}{n}\right)+\epsilon]$	$E^i_{1<i\leq n}$	$[\frac{(1-2i)}{2}\frac{L}{n}-\epsilon, \frac{(1-2i)}{2}\frac{L}{n}+\epsilon]$
E^{-1}	$[\frac{1}{2}\left(\frac{L}{n}+\epsilon\right)-\epsilon, \frac{1}{2}\left(\frac{L}{n}+\epsilon\right)+\epsilon]$	E^1	$[-\frac{1}{2}\left(\frac{L}{n}+\epsilon\right)-\epsilon,$ $-\frac{1}{2}\left(\frac{L}{n}+\epsilon\right)+\epsilon]$
E^0	$[0,0]$		

Whenever it is understood from the context, we shall denote by E^- (or E^+) any class $E^{i<0}$ (or $E^{i>0}$). The *accuracy* of the equilibrium supervisor is given by $\kappa_n = \frac{L}{n}$. ϵ is the admissible tolerance for the equilibrium point. The range of the midpoints of the intervals that belong to a class E^i is called the *representative* of the class E^i, denoted $[E^i]$. In this paper, whenever it is clear from the context, we shall identify a class E^i with its representative.

The states of the QI–MDP model are pairs of classes (E^i_α, E^j_β), representing the material results of the social exchange process from the point of view of the agents α and β. The pair of classes (E^0_α, E^0_β) is a *terminal* state, representing that the system is in equilibrium.

The *actions* considered in the model are state transitions $(X^i, X^j) : \hat{E}\times\hat{E} \rightarrow \hat{E}\times\hat{E}$, with $i, j = -n, \ldots, n$, defined by $(X^i, X^j)(E^i_\alpha, E^j_\beta) = (E^{i'}_\alpha, E^{j'}_\beta)$ if $mid(\ [E^i_\alpha] + X^i) \in E^{i'}_\alpha$ and $mid([E^j_\beta] + X^j) \in E^{j'}_\beta$, which occur by the addition, to the representatives of the classes E^i_α and E^j_β, of intervals X^i and X^j that should be of the following types: (i) the *absolute ϵ-null value* $0_\epsilon = [-\epsilon, +\epsilon]$; (ii) a *compensation interval*, which is the least quasi-symmetric, denoted by C^i, of a class representative E^i; (iii) a *go-forward-k-step interval*, which is an interval, denoted by F^i_k, that transforms a class E^i into $E^{(i+k)\neq 0}$, with $i \neq L$; (iv) a *go-backward-k-step interval*, which is an interval, denoted by B^i_{-k}, that transforms a class E^i into $E^{(i-k)\neq 0}$, with $i \neq -L$.

The set $\mathcal{C}$ of compensation intervals is shown in Table 1. The set $\mathcal{F}$ of go-forward intervals and their respective effects are partially presented in Table 2. The set of go-backward intervals, denoted by $\mathcal{B}$, can be specified analogously.

For example, for a state of type

$$(E^i_\alpha, E^j_\beta)_{-n\leq i<-1, 1<j\leq n} \equiv ([i\frac{L}{n}, (i+1)\frac{L}{n}], [(j-1)\frac{L}{n}, j\frac{L}{n}]),$$

the *compensation–compensation* action and the *go-backward*$_{-3}$–*go-forward*$_{+2}$ actions are given by (A1) $(C^i, C^j) = ([-\frac{2i+1}{2}\frac{L}{n} - \epsilon, -\frac{2i+1}{2}\frac{L}{n} + \epsilon], [\frac{(1-2j)}{2}\frac{L}{n} - \epsilon, \frac{(1-2j)}{2}\frac{L}{n} + \epsilon])$ and (A2) $(B^i_{-3}, F^j_{+2}) = ([-3\frac{L}{n} - \epsilon, -3\frac{L}{n} + \epsilon], [2\frac{L}{n} - \epsilon, 2\frac{L}{n} + \epsilon])$, respectively, resulting in the state transitions: $(E^i_\alpha, E^j_\beta)_{-n\leq i<-1, 1<j\leq n} \overset{(A1)}{\mapsto} (E^0_\alpha, E^0_\beta)$ and $(E^i_\alpha, E^j_\beta)_{-n\leq i<-1, 1<j\leq n} \overset{(A2)}{\mapsto} (E^{(i-3)}_\alpha, E^{(j+2)}_\beta)$.

The equilibrium supervisor has to find, for each state E, the action that shall achieve the terminal state or, at least, another state from where the terminal state can be achieved,

Table 2. Specification of some go-forward intervals and their respective effects

State	Go-forward interval F^i_{+k}	Effect
$E^i_{-n\leq i<-1}$	$\left[k\frac{L}{n}-\epsilon, k\frac{L}{n}+\epsilon\right]_{1-i\leq k\leq n-i-1}$	$E^i \mapsto E^{i+k}, 1 < i+k \leq n$
E^{-1}	$\left[k\frac{L}{n}-\epsilon, k\frac{L}{n}+2\epsilon\right]_{2<k\leq n}$	$E^{-1} \mapsto E^{-1+k}, 1 < -1+k \leq n$
E^0	$\left[k\frac{L}{n}, (k+1)\,\frac{L}{n}\right]_{0<k\leq n-1}$	$E^0 \mapsto E^{k+1}, 1 < k+1 \leq n$
E^1	$\left[k\frac{L}{n}-2\epsilon, k\frac{L}{n}+\epsilon\right]_{0<k\leq n-i}$	$E^1 \mapsto E^{1+k}, i < 1+k \leq n$
$E^i_{1<i\leq n}$	$\left[k\frac{L}{n}-\epsilon, k\frac{L}{n}+\epsilon\right]_{0<k\leq n-i}$	$E^i \mapsto E^{i+k}, i < i+k \leq n$

with the least number of steps. The choice of such actions is also regulated by the rules of the social exchanges, and, therefore, there are some state transitions that are not allowed. Based on a optimal policy, the equilibrium supervisor may be asked to recommend that the agents act optimally. An *optimal exchange recommendation* consists of a function that gives, for each actual material result (represented by a state of the model), a partially defined exchange stage that shall restore or establish the material equilibrium or, at least, give conditions that it be achieved in a least number of steps with least value uncertainty. The optimal exchange recommendation associates state transitions determined by the optimal policy with agents' social exchanges.

Although the interacting agents acknowledge the optimal recommendations from the equilibrium supervisor, they are autonomous in the sense that *they may not follow the recommendations exactly*. Thus, the system may achieve another state different from the one expected by the supervisor and, therefore, there may be a great deal of uncertainty about the effects of the agents actions. Even if the agents follow a recommendation exactly, we will show that the effect may not be the expected by the supervisor, since it depends on the ratio $\frac{\kappa_n}{\epsilon}$, where $\kappa_n = \frac{L}{n}$ is the equilibrium supervisor accuracy and ϵ $(0 \leq \epsilon < \kappa_n)$ is the admissible tolerance. On the other hand, we assume that there is never any uncertainty about the current state of the system, that is, the equilibrium supervisor always has access to the current configuration of exchange values and has complete and perfect abilities to evaluate the current material balance.

Definition 1. *A* Qualitative Interval Markov Decision Process *(QI–MDP) for keeping social exchanges in equilibrium is a tuple* $\langle E, A, F, R\rangle^{L,n}_{\epsilon}$*, where*[5]*:*

- The set of the states is the set of pairs of equivalence classes of intervals $E = E_\alpha \times E_\beta$*, with* $E_\lambda = \{E^i \mid i = -n, \ldots, -1, 0, 1, \ldots, n\}$ *defined in (4.1).*

- $A = (\mathcal{C} \cup \mathcal{F} \cup \mathcal{B} \cup \{[-\epsilon, +\epsilon]\}) \times (\mathcal{C} \cup \mathcal{F} \cup \mathcal{B} \cup \{[-\epsilon, +\epsilon]\})$ *is the set of possible actions, where* $\mathcal{C}$*,* $\mathcal{F}$ *and* $\mathcal{B}$ *are the sets of compensation, go-forward and go-backward intervals, respectively.*

- $F : E \times A \rightarrow \Pi(E)$ *is the state-transition function, that gives for each state and each action, a probability distribution over the set of states;*

- $R : (E \times A) \rightarrow \mathbb{R}$ *is the reward function, giving the expected immediate reward gained by choosing an action* a *when the current state is* e*.*

[5] In this model, the next state and the expected reward depend only on the previous state and the action taken, satisfying the so-called *Markov property*.

4.2 The Optimal Policy and the Reward Function

The reward function plays an important role when the equilibrium supervisor is choosing the action that will generate a recommendation of agents interaction, in each state. The supervisor aims to maximize the utility of sequences of actions, evaluated according to the reward function.

A sample reward function $R : (E \times A) \rightarrow \mathbb{R}$ that conforms to the idea of supporting a recommendation function that is able to direct agents into social equilibrium is partially sketched in Table 3. This particular function illustrates various requirements that should be satisfied by all reward functions of the model. Observe, for instance, that if the current state is of the type (E^-, E^+), then the best action to be chosen is a *compensation-compensation* action (C, C), which results in a state transition $(E^-, E^+) \mapsto (E^0, E^0)$. Any other choice will make the agents either take a long way to the equilibrium or get away from it.

On the other hand, if the current state is of type (E^-, E^-), then a *compensation-compensation* action (C, C) would generate a recommendation of agent exchanges of *satisfaction-satisfaction* type, which is impossible according to the model of social interactions [7], since it is impossible for an agent to get a satisfaction value from no service at all. The reward function R states that (C, C) is a very bad action to be chosen in such situation.

Any optimal policy $\pi^* : E \rightarrow A$ solving the social equilibrium problem should satisfy the set of requirements expressed by the schema partially sketched in Table 4 [6] The *optimal recommendation* associated to an optimal policy π^* is a function ρ_{π^*} that gives, for each state (E^i_α, E^j_β) and optimal action $\pi^*(E^i_\alpha, E^j_\beta) = (X^i, Y^j)$, a partial definition of a recommended exchange stage, consisting of pairs $((r_{\alpha\beta}, X^i), (s_{\beta\alpha}, Y^j))$ or $((r_{\beta\alpha}, Y^j), (s_{\alpha\beta}, X^i))$, where $(r_{\lambda\delta}, W)$ means the realization, by the agent λ, of a service with investment value $W < 0$, and $(s_{\delta\lambda}, W')$ means δ's satisfaction with interval value W', for receiving the service. The optimal recommendation ρ_{π^*} is also partially sketched in Table 4.

Table 3. Partial schema of the reward function R

R	(C, C)	$(0_\epsilon, C)$	$(C, 0_\epsilon)$	(B_{-1}, F_1)	(B_{-3}, F_3)	(F_1, B_{-1})	(C, B_{-1})	(F_1, C)
(E^-, E^+)	30	20	-30	-5	-10	3	20	20
(E^+, E^+)	30	20	20	0	0	0	18	20
(E^-, E^-)	-30	-30	-30	30	0	30	28	-30

5 Discussion

In the following, consider that the *agents always follow the recommendations given by the equilibrium supervisor*. We show that, even in this favorable case, the decision process is a non-trivial one, due the qualitative nature of exchange values. The results concern the reachability of the terminal state show that under some conditions, it is

[6] Notice that it is a non deterministic policy.

Table 4. Partial schemata of the optimal policy π^* and associated optimal recommendation ρ_{π^*}

State	Optimal policy	Recommendation
$(E^i, E^j)^{-n \leq i < -1}_{1 < j \leq n}$	$(C^i > 0, C^j < 0)$	$((r_{\beta\alpha}, C^j), (s_{\alpha\beta}, C^i))$
$(E^i, E^j)_{1 < i,j \leq n}$	$(C^i < 0, C^j < 0)$	$((r_{\alpha\beta}, C^i), (s_{\beta\alpha}, C^j))$ or $((r_{\beta\alpha}, C^j), (s_{\alpha\beta}, C^i))$
$(E^0, E^j)_{1 < j \leq n}$	$(0_\epsilon, C^j < 0)$	$((r_{\beta\alpha}, C^j), (s_{\alpha\beta}, 0_\epsilon))$
$(E^0, E^i)_{-n \leq i < -1}$	$(B^0_{-1} < 0, F^i_{+(-i+1)} > 0)$	$((r_{\alpha\beta}, B^0_{-1}), (s_{\beta\alpha}, F^i_{+(-i+1)}))$
$(E^1, E^i)_{-n \leq i < -1}$	$(B^1_{-1} < 0, C^i > 0)$	$((r_{\alpha\beta}, B^1_{-1}), (s_{\beta\alpha}, C^i))$
(E^{-1}, E^1)	$(F^{-1}_{+1} > 0, B^1_{-1} < 0)$	$((r_{\beta\alpha}, B^1_{-1}), (s_{\alpha\beta}, F^{-1}_{+1}))$
(E^1, E^{-1})	$(B^1_{-1} < 0, F^{-1}_{+1} > 0)$	$((r_{\alpha\beta}, B^1_{-1}), (s_{\beta\alpha}, F^{-1}_{+1}))$
$(E^i, E^1)_{-n \leq i < -1}$	$(C^i > 0, B^1_{-1} < 0)$	$((r_{\beta\alpha}, B^1_{-1}), (s_{\alpha\beta}, C^i))$
(E^{-1}, E^0)	$(F^{-1}_{+1} > 0, B^0_{-1} < 0)$	$((r_{\beta\alpha}, B^0_{-1}), (s_{\alpha\beta}, F^{-1}_{+1}))$
(E^0, E^{-1})	$(B^0_{-1} < 0, F^{-1}_{+1} > 0)$	$((r_{\alpha\beta}, B^0_{-1}), (s_{\beta\alpha}, F^{-1}_{+1}))$
$(E^i, E^j)_{-n \leq i,j < -1}$	$(F^i_{+(-i+1)} > 0, B^j_{-1} < 0)$ or $(B^j_{-1} < 0, F^i_{+(-i+1)} > 0)$	$((r_{\beta\alpha}, B^j_{-1}), (s_{\alpha\beta}, F^i_{+(-i+1)})$ or $((r_{\alpha\beta}, B^j_{-1}), (s_{\beta\alpha}, F^i_{+(-i+1)})$

always possible to have the system equilibrated in at most four steps. Let $M^\tau_{\alpha\beta}$ and $M^\tau_{\beta\alpha}$ be the material results of an exchange process, according to the points of view of the agents α and β, respectively, at step τ.

Proposition 2. *If $M^0_{\alpha\beta} \in E^{-1}$ and $M^0_{\beta\alpha} \in E^1$, then the system achieves the equilibrium in one step if and only if $1 < \frac{\kappa_n}{\epsilon} \leq 3$, $\epsilon > 0$.*

Proof. ($\Rightarrow$) If the system is at the state (E^{-1}, E^1), then, for the β's material result, it holds that $\epsilon < mid(M^0_{\beta\alpha}) \leq \frac{L}{n}$, and the optimal recommendation (Table 4, $row7$) is based on the optimal action $(C, C) = \left[-\frac{1}{2}\left(\frac{L}{n} + \epsilon\right), -\frac{1}{2}\left(\frac{L}{n} + \epsilon\right)\right]$. It follows that: $\epsilon - \frac{1}{2}\left(\frac{L}{n} + \epsilon\right) < mid(M^0_{\beta\alpha}) - \frac{1}{2}\left(\frac{L}{n} + \epsilon\right) \leq \frac{L}{n} - \frac{1}{2}\left(\frac{L}{n} + \epsilon\right) \Rightarrow \frac{1}{2}\left(-\frac{L}{n} + \epsilon\right) < mid(M^1_{\beta\alpha}) \leq \frac{1}{2}\left(\frac{L}{n} - \epsilon\right) \Rightarrow \frac{1}{2}(-h\epsilon + \epsilon) < mid(M^1_{\beta\alpha}) \leq \frac{1}{2}(h\epsilon - \epsilon)$, where $\frac{L}{n} = h\epsilon$, with $h > 1$. If the system achieves the equilibrium in the step 1, then it holds that $\frac{1}{2}(h\epsilon - \epsilon) \leq \epsilon$. It follows that $1 < h \leq 3$, and therefore, $1 < \frac{\kappa_n}{\epsilon} \leq 3$, since $\kappa_n = \frac{L}{n}$. The proofs for α's material result and of ($\Leftarrow$) are analogous. □

Proposition 3. *(i) If $M^0_{\alpha\beta} \in E^i$, with $1 < i \leq n$, then it is possible to get $M^\tau_{\alpha\beta} \in E^0_\alpha$ in at most $\tau = 2$ steps if and only if $1 < \frac{\kappa_n}{\epsilon} \leq 3$; (ii) If $M^0_{\beta\alpha} \in E^i$, with $-n \leq i < -1$, then it is possible to get $M^\tau_{\beta\alpha} \in E^0_\beta$ in at most $\tau = 2$ steps if and only if $1 < \frac{\kappa_n}{\epsilon} \leq 3$.*

Proof. (i)($\Rightarrow$) If $(i-1)\frac{L}{n} \leq mid(M^0_{\alpha\beta}) < i\frac{L}{n}$ and the optimal recommendation (Table 4, $row2$) is based on the optimal action $C = \left[\frac{(1-2i)}{2}\frac{L}{n}, \frac{(1-2i)}{2}\frac{L}{n}\right]$, then $(i-1)\frac{L}{n} + \frac{(1-2i)}{2}\frac{L}{n} < mid(M^0_{\beta\alpha}) + \frac{(1-2i)}{2}\frac{L}{n} \leq i\frac{L}{n} + \frac{(1-2i)}{2}\frac{L}{n}$, that is, $-\frac{1}{2}\frac{L}{n} < mid(M^1_{\beta\alpha}) \leq \frac{1}{2}$. It holds that $M^1_{\beta\alpha} \in E^1_\alpha$. From Prop. 2, it follows that with more one step we can get the desired result. The proofs of (i)($\Leftarrow$) and (ii) are analogous. □

From Prop. 3 it follows that an individual transition from a material result that belongs to a class E^i, with $1 < i \leq n$ or $-n \leq i < -1$, to the equilibrium can be done in at most two steps ($E^i \mapsto E^1($ or $E^{-1}) \mapsto E^0$). However, in any interaction between two agents, combined transitions departing from a state (E^i, E^j) or (E^j, E^i), with $1 < i \leq n$ and $-n \leq j < -1$, may result in a state different from (E^1, E^{-1}), (E^{-1}, E^1) or (E^0, E^0). We may have, for example, (E^{-1}, E^0), and, in this case, it will not be possible to get the equilibrium in one more step, since any *compensation* or *go-forward* action for α is not allowed without a correspondent β's service. The solution given by the optimal policy is then to have a transition to (E^1, E^{-1}) and then, finally, to reach (E^0, E^0). Thus, the overall process takes three steps.

The worst case is when the interaction presents material results that belong to the state (E^i, E^j), with $-n \leq i, j < -1$, since two simultaneous positive compensation actions (that would require a recommendation of satisfaction values for the two agents without any service at all) are not allowed. In this case, the optimal recommendation (Table 4) leads the agents to get the material equilibrium in at most four steps, by one of the following transitions: $(E^i, E^j)_{-n \leq i,j < -1} \overset{row12}{\mapsto} (E^1, E^j)_{-n \leq j < -1} \overset{row6}{\mapsto} (E^0, E^{-1}) \overset{row11}{\mapsto} (E^{-1}, E^1) \overset{row7}{\mapsto} (E^0, E^0)$, or $(E^i, E^j)_{-n \leq i,j < -1} \overset{row13}{\mapsto} (E^j, E^1)_{-n \leq j < -1} \overset{row9}{\mapsto} (E^{-1}, E^0) \overset{row10}{\mapsto} (E^1, E^{-1}) \overset{row8}{\mapsto} (E^0, E^0)$.

6 Conclusion

This paper introduced the QI–MDP version of the Markov Decision Process. The combination of interval-based modelling and qualitative approach to the comparison of values of the model made it well suited for solving the problem of keeping social exchanges in equilibrium. From the point of view of Jean Piaget's theory of social interactions, the QI–MDP means a sound way of making practical use of the INRC group of social exchanges that structure the social interactions and defines its equilibrium problem [1]. The QI–MDP model is general enough to be applied to other problems, besides the problem of keeping social interactions in equilibrium. It can also be applied to equilibrium problems of other kinds of systems, besides systems of social exchanges, if such systems have one single equilibrium state.

Future work will be concerned with the case of an equilibrium supervisor that is not able to determine the material balance of social exchange processes with complete reliability (i.e., it is not allowed to know all the exchange values of the two agents). In this case, a partially observable Markov decision process (POMDP) shall be considered (see, p.ex., [3]), since the equilibrium supervisor shall be able to make external observations (also probabilistic) to help him to decide about the recommendations.

References

1. A.C.R. Costa and G.P. Dimuro. The Case for Using Exchange Values in the Modelling of Collaborative Learning Interactions. In J. Mostow and P. Tedesco, eds, *Proceedings of Workshop 9 in the 7th International Conference on Intelligent Tutoring Systems, ITS 2004*, pages 19–24, Maceió, 2004.
2. M. d'Inverno and M. Luck. *Understanding Agent Systems.* Springer, Berlin, 2001.
3. L.P. Kaelbling, M.L. Littman, and A.R. Cassandra. Planning and Acting in Partially Observabe Stochastic Domains. *Artificial Intelligence*, 101(1):99–134, 1998.

4. G.C. Homans. *Social Behavior - Its Elementary Forms*. Harcourt, Brace & World, New York, 1961.
5. R.A. Howard. *Dynamic Programming and Markov Processes*. MIT Press, Cambridge, 1960.
6. R.E. Moore. *Methods and Applications of Interval Analysis*, SIAM, Philad., 1979.
7. J. Piaget. *Socialogical Studies*. Routlege, London, 1995.
8. M.L. Puterman. *Markov Decision Processes – Discrete Stochastic Dynamic Programming*. Wiley, New York, 1994.
9. M.R. Rodrigues, A.C.R. Costa, and R. Bordini. A System of Exchange Values to Support Social Interactions in Artificial Societes. In *Proceeding of the Second International Conference on Autonomous Agnets and Multiagents Systems, AAMAS 2003*, pages 81–88, Melbourne, Australia, 2003. ACM.
10. M.R. Rodrigues and A.C.R. Costa. Using Qualitative Exchange Values to Improve the Modelling of Social Interactions. In D. Hales, B. Edmonds, E. Norling, and J. Rouchier, eds, *Procedings of 4th Workshop on Agent Based Simulations*, n. 2927 in Lecture Notes in Computer Science, pages 57–72, Melbourne, Australia, 2003.
11. S. Russel and P. Norvig. *Artificial Intelligence, a Modern Approach*. Prentice Hall, Reading, 2003.
12. D.J. White. *Markov Decision Processes*. Wiley, New York, 2002.
13. M. Wooldridge. *An Introduction to Multi-Agent Systems.* Wiley, New York, 2002.