Cognitive Science Quarterly (2000) ***1***, ..-..

Commitments: A Game-Theoretic and Logical Perspective[1]

Vincent Buskens
Utrecht University[2]

Lambèr Royakkers
Eindhoven University of Technology[3]

Many social interactions between agents demand the use of commitments to reach socially efficient or avoid socially inefficient outcomes. Commitments express the desires, goals, or intentions of the agents in an interaction. In this article, we distinguish between *unilateral* and *bilateral* commitments, and between whether or not an agent has to agree with a commitment made by the other agent before the commitment becomes effective. Using a game-theoretic model, we will show that, depending on the incentive structure, different interactions require different types of commitments to reach socially efficient outcomes. Based on these results, we discuss whether existing (or slightly adapted) logical formalizations are adequate for the description of certain types of commitments and which formalization is suitable for reaching a socially efficient outcome in a specific interaction. We claim that a logical formalization of commitment aiming at a socially efficient outcome should be based on assumptions about the type of interaction and the suitable type of commitment. A more general conclusion of this article is that game-theoretic arguments can help to provide specifications for logical formalizations of systems of more agents if one has an idea about the incentive structure of the interaction.

Keywords: commitment, intention, game theory, logic.

Introduction

Many social interactions between two (or more) agents demand for various reasons the use of commitments to reach socially efficient or avoid socially inefficient outcomes. As Castelfranchi (1995) states it: "*Commitment* is seen as the glue of the group, of collective activity: it links the agent with the joint

[1] Supported by a grant from the Netherlands Organization for Scientific Research (NWO) and a grant from the Niels Stensen Foundation.

[2] Department of Sociology, P. O. Box 80140, 3508 TC, Utrecht, The Netherlands, V.Buskens@fss.uu.nl.

[3] Faculty of Technology Management, P. O. Box 513, 5600 MB, Eindhoven, The Netherlands, L.M.M.Royakkers@tm.tue.nl.

Received 18.03.02

goal and the common solution, it links the members' actions with the collective plan, it links the members with each other."

We will start with an example. Assume you want to write an article together with a colleague. You are both convinced that joining forces will produce a better product than writing two articles separately. However, you as well as your colleague cannot be sure that the other will actually invest his fair share in this joint project (cooperate). Still, if both of you work hard, you will both be satisfied. You realize that if the colleague sits back (defects) while you do the job, he is even better off and you would have preferred to write an article alone. Clearly, your colleague also fears that you sit back and profit from his effort.

		Agent 2	
		Defect	Cooperate
Agent 1	Defect	2,2	4,1
	Cooperate	1,4	3,3

Figure 1: Strategic form of the Prisoner's Dilemma Game

The 'game' described above (without commitments) is called a Prisoner's Dilemma Game (cf. Luce & Raiffa, 1957). In strategic form,[4] the game is shown in figure 1. The two values in the cells of the matrix indicate the payoffs for agent 1 and agent 2, respectively, related to a combination of actions of the two agents, which are strictly ordered (from 1 to 4) and which do not represent exact payoffs but only ordinal utilities. The expected action in this game is 'defect' by both agents, because independent of the action of the other agent, each agent is better off by defecting. Consequently, both agents receive 2 instead of 3, although they could obtain 3 if they both would cooperate. Thus, the expected outcome (2,2) is socially inefficient. However, by committing to cooperation, e.g., by mutually informing the responsible professor who can incur sanctions on the researcher who does not work on the joint paper, cooperation becomes the best option for both agents. Hence, a mutual commitment leads to a better outcome for both agents in this situation.

If we want to represent such a simple interaction in a logical system, only the possible actions are described. Commitment is then introduced as an elementary proposition (cf. Shoham, 1993). This implies that the commitment is a fact that does or does not occur. More sophisticated theories (Dunin-Keplicz & Verbrugge, 1996, 1999, 2001; Meyer, Van der

[4] For all basic game-theoretic terminology and aspects we refer the reader to Rasmusen (1994).

Hoek & Van Linder, 1999) describe a formalization of motivational attitudes such as *intentions*, *goals*, and *wishes* to explain why agents behave the way they do. However, within the logical systems there is nothing that drives the motivational attitudes. It is only stated that if certain attitudes are present, commitments are used without explicit reasoning why and when a certain attitude leads to a commitment. Our main criticism of these logical systems is that they, strictly speaking, do *not* explain but only describe actions by agents, probably including the use of commitments. Logical systems fail to distinguish between which conditions provide incentives for agents to use commitments and which conditions do not provide such incentives. Moreover, logical systems cannot distinguish which commitment is or is not credible in a given interaction. The reason is that logical systems generally neglect the incentives related to various combinations of actions and the strategic interdependence between different agents. In game theory, however, motivational attitudes are represented by the payoffs agents receive at the end of an interaction, based on their combination of actions. By distinguishing different commitment types using game theory, we are able to specify conditions for which certain logical formalizations of commitments are appropriate using the goals and intentions of the agents.

The situation discussed above is only one example of a situation in which a commitment can change the expected outcome for an interaction between two agents. Likewise, the usefulness of commitment systems can be investigated for many social and legal interactions. For now, we will give an informal description of what we mean by a commitment in this article. Later we will become more precise and we will show that there are various types of commitments.

> **Definition:** A *commitment* is an action by one agent before an interaction with other agents that signals to the other agents the intention to perform a particular action later on in the interaction.

This description does not exclude that more agents can commit to an action simultaneously, and that it is not necessary that the other agents are convinced by the signal sent by the committing agent. However, we restrict ourselves in this article to commitments that ensure that the agent who commits to a certain action will execute this action (*binding commitments*). Still, this definition is very broad. The definition will be further specified in a more formal manner in the following section using game-theoretic notions. We will show that the broad definition provided here includes several types of commitments that can have different consequences depending on the goals and intentions of the agents in the situations in which the commitment is used. Later, we will extend logical systems with the formalizations that correspond with the game-theoretic analysis.

The following section demonstrates how commitments are introduced in game-theoretic analyses and illustrates how different commitment regimes can lead to different outcomes in interactions among agents. The subsequent section introduces the use of commitments in logical systems and translates the outcomes of the game-theoretic analysis in logical formalizations. Finally, we conclude and summarize.

Commitments in a Game-Theoretic Setting

Defining Commitments

The example in the introduction has shown that adding incentives to the different actions enables to explain why a commitment is effective in an interaction. In such a 'game', the *goal* or *wish* of the agents is to maximize their 'utility' at the end of the game. We only consider rational players (agents). In correspondence with this, we assume that agents use actions that correspond with (Nash) equilibrium strategies. A commitment of an agent signals the *intention* to perform a specified action. A sanction will be imposed on this agent if he deviates from the commitment. In the sequel we will show that a commitment can have two functions. First, the commitment can help the agents to coordinate on one of the equilibria if there exist multiple equilibria. Second, agents can escape an equilibrium outcome that is less attractive for one or both of the agents compared to one of the non-equilibrium outcomes, as is shown in the example in the introduction. The commitment can only be successful if the value of the sanction applied after an agent deviates from the action specified in the commitment is large enough. As indicated before, we assume that this value indeed is large enough such that an agent will always perform the action intended with the commitment (binding commitment).

Within a game-theoretic setting, we can define a commitment as an additional move/action in a game (Schelling, 1960; Snijders, 1996; Snijders & Buskens, 2001).[5] The two agents in a game as illustrated in the introduction have to decide whether or not they commit to one of the two possible actions before they play the underlying game. The payoffs related to the game will not change if none of the agents commits to one of these actions (cooperation or defection in the example). If one or both agents commit to one of the actions, the payoffs will change depending on who has committed and the rules that are related to the commitment regime that is chosen in a particular situation. Mostly, the execution of the commitment and imposing the related sanction for the agent who deviates from the committed action has to be

[5] The term 'move' in game theory corresponds with the term 'action' used in logical approaches. In this article, we will use the term 'action'.

done by a third party such as a witness or a normative authority. In this article, we will neglect the role of such a third party.[6]

Besides explaining the use and effectiveness of commitments, game theory can help to distinguish between different commitment regimes. Before discussing different games, we discuss four types of commitment(s) that can be distinguished under different commitment regimes. Such a regime determines which agent is allowed to commit and if both agents are allowed to commit whether this has to be done unilateral or bilateral.

Another aspect that has to be determined in the regime is whether the commitment of one agent needs the consent of the other agent to become effective. For example, a car driver will stop for somebody who started crossing the road although the car driver would have preferred to continue driving while the other person waited at the sidewalk. In this example, starting to cross the road is the commitment signaling the intention of the pedestrian to go first without the consent of the car driver. We will see in the examples below that, depending on the interaction, commitments that can be made by one agent without the consent of the other agent can lead to outcomes that are worse for the last agent than the outcome if no commitment would have been made.

A third aspect of the regime can prescribe that a commitment of one agent becomes effective only under that condition that the other agent commits to a specified other action. We describe now four (basic) types of commitments that can occur under the various regimes and that can have different implications in the example we discuss thereafter.

- *Unilateral commitment without agreement*: one agent expresses that he intends to perform an action and this commitment becomes effective without the need of the other agent's agreement;
- *Unilateral commitment with agreement*: one agent expresses that he intends to perform an action and this commitment becomes effective if and only if the other agent agrees with this commitment;
- *Bilateral commitment without agreement*: both agents express that they intend to perform an action and these commitments become effective without the need of the other agent's agreement;[7]
- *Bilateral commitment with agreement*: both agents express that they intend to perform an action under the condition that the other agent commits to perform another action and the commitments become effective if both agents agree on the combination of commitments.

[6] The third agent has a very crucial role in normative contexts (norms efficacy) (cf. Conte & Castelfranchi, 1995) and in contractual contexts implicating free riders and cheaters. Note that the professor in our example in the introduction can be considered as the third party.

[7] This commitment system can never change the outcome of the game because if the action of both agents is determined as soon as they have placed the commitment, placing the commitment is just equivalent to playing the underlying game itself.

		Agent 2	
		Left	Right
Agent 1	Top	*a1, a2*	*b1, b2*
	Bottom	*c1, c2*	*d1, d2*

Figure 2: A 2 × 2 game with general payoffs

Note that the commitment regimes apply to all situations in which two agents interact and the outcome depends on the actions of both agents. One could say that the commitments are 'social' commitments because the commitment of one agent has implications for the other agent as well. However, the crossing-the-road example shows that the implications do not need to be positive for both agents as is required in Castelfranchi's (1995; see also Singh, 1999) description of a social commitment. To prevent misunderstanding we will not use the term social commitment.

Figure 2 represents a 2 × 2 game (a game with 2 agents each choosing between two possible actions) with general payoffs for both agents. Agent 1 chooses between Top and Bottom, while agent 2 chooses between Left and Right. Both agents obtain the payoffs that belong to their combination of choices. E.g., if agent 1 chooses Top and agent 2 chooses Left, agent 1 receives *a1* and agent 2 receives *a2*.

Depending on the commitment regime, the payoffs in the 2 × 2 game will change if one agent or both agents commit to one of their strategies. For example, if agent 1 has to possibility for a unilateral commitment without agreement and he commits to playing Top, the payoffs will be changed as shown in matrix (1) in figure 3. The payoff for agent 1 is decreased with some value, say C_1, by stipulating a sanction if he deviates from the action specified by the commitment.[8] Since we have restricted our analysis to binding commitments, the decreases in payoff is such that agent 1 does not have an incentive to deviate from the action specified by the commitment, i.e., $a1 > c1\text{-}C_1$ and $b1 > d1 – C_1$. Because this is a unilateral commitment without agreement, the new payoff structure will be used even if agent 2 does not like this. However, if the commitment regime prescribes that agent 2 has to agree on the commitment made by agent 1, the new payoff structure will only be used if agent 2 indeed agrees. Otherwise, the original payoff structure will remain. Similar payoff changes occur if agent 2 has the possibility of a unilateral commitment. In case of a bilateral commitment regime, both agents have to decide whether or not they want to commit to

[8] Other payoff changes can also be specified, e.g., that the sanction for deviating from a commitment by one agent is transferred to the other agent rather than only subtracted from the deviating agent's payoff.

	Left	Right
Top	4,4	3,3
Bottom	1-C_1,1	2-C_1,2

(1)

	Left	Right
Top	2-C_1,4-C_2	4-C_1,1
Bottom	3,2-C_2	1,3

(2)

Figure 3: Payoffs 2 × 2 games with commitments

one of their possible actions. Assume that agent 1 commits to Bottom and agent 2 commits to Right. Then, in case of a bilateral commitment without agreement, the new payoff structure as shown in matrix (2) of figure 3 becomes effective unconditionally. It is possible that only one of the agents commits, which implies that only for this agent the payoffs are changed. If agreement is necessary for the bilateral commitment, the changes in the payoffs become effective only if both agents agree with the commit made by the other agent. Otherwise, the game is played with the original payoffs. We want to stress that commitments do not put restrictions on which actions the agents *can* perform. Agent 1 still can play Bottom or Top and agent 2 still can play Left or Right. The payoffs, however, are changed as a result of the commitment such that both agents do not have an incentive to deviate from the bilateral commitment.

An Overview of Games with Ordered Payoffs

As an illustration, we consider 2 × 2 games with preferences over the four possible outcomes that are strictly ordered for both agents. Because only the ordering of the payoffs is important for the analyses, they can be labeled as 1, 2, 3, and 4. Rapoport, Guyer and Gordon (1976) show that there exist 78 distinct 2 × 2 games with strictly ordered payoffs.[9] Each of the four outcomes represents a possible goal state for the agents. The goal states for the two agents do not need to coincide.

Now, we classify the 78 games in eight groups. The games are classified such that for all games within a group the same arguments hold if commitment regimes are considered. Figure 4 presents the matrices for one representative of each group.[10] In these games, agent 1 chooses between Top and Bottom, while agent 2 chooses between Left and Right.

[9] Two games are considered the same if the one can be constructed from the other by changing rows, columns, or person labels.

[10] Readers interested in the precise classification of all the games can contact the authors for an overview.

	Left	Right
Top	4,4	3,3
Bottom	1,1	2,2

(1)

	Left	Right
Top	2,4	4,1
Bottom	3,2	1,3

(2)

	Left	Right
Top	3,3	1,4
Bottom	4,1	2,2

(3)

	Left	Right
Top	2,4	4,1
Bottom	1,2	3,3

(4)

	Left	Right
Top	2,3	4,1
Bottom	1,2	3,4

(5)

	Left	Right
Top	3,4	2,1
Bottom	1,2	4,3

(6)

	Left	Right
Top	2,4	3,1
Bottom	1,2	4,3

(7)

	Left	Right
Top	3,3	2,4
Bottom	4,2	1,1

(8)

Figure 4: Representative examples of 2 x 2 games with strictly ordered outcomes

Examples (1) and (2) illustrate two situations in which both agents do not want or need to commit to any of the two actions. Example (1) represents a group of 58 games in which at least one of the two agents has a dominant strategy. An agent has a dominant strategy if there is one action the agent can perform that gives him a higher payoff for each of the actions the other agent can perform. The other agent optimizes his payoff given the dominant strategy of the first agent, and both agents cannot do better using a commitment for some other strategy. In the example, this implies that both agents obtain 4. Clearly, none of them can do better whatever commitment regime would be chosen.

Example (2) represents four games in which none of the agents has a dominant strategy and there exists only one (mixed) equilibrium in which the agents randomly choose between the two options. Their expected payoffs lie between 2 and 3.[11] If one agent would commit unilaterally, he would never obtain more than 2. E.g., if agent 1 commits to Top, agent 2 plays Left, and if agent 1 commits to Bottom, agent 2 plays Right. If both agents would commit, each of the agents is only willing to do that under the condition that the other agent commits to the action to which this last agent is not willing to commit. Consequently, there is no feasible commitment regime from which either one or both agents could profit.

Example (3) is the Prisoner's Dilemma game. This is the very special game that is also illustrated in the introduction. In this game, the game-theoretic solution predicts that both agents obtain 2, while they both would prefer to obtain 3. However, this would imply that both agents have to deviate from their dominant strategy. The only commitment regime that can work in this game is the bilateral commitment with agreement, implying that both agents should commit to not playing the dominant strategy and if one agent does not commit, a commitment of the other agent will not materialize. The reason for this is that the agents do not want to commit unilaterally to Top or Left, respectively, because the other agent then certainly plays the dominant strategy leaving the first agent with the worst outcome possible. The bilateral commitment without agreement does not work, because both agents are not willing to commit to the dominated strategy – which leads to the outcome (3,3) – if this commitment remains effective also if the other agent does not commit to his dominated strategy. In this last situation, both agents cannot be sure that they will not be exploited by the other agent and end up with a payoff 1, while the other agent receives 4.

Example (4) is also a unique game. In this game, agent 1 has a dominant strategy, because whatever agent 2 does, he is always better off playing Top. This implies that the agent 2 will choose Left, which leaves himself with 4, but agent 1 only with 2. Since, agent 2 reaches his most preferred outcome without commitments, a commitment regime that ask for his consent will never lead to an actual commitment unless the commitments do not change the outcome of the game. However, agent 1 wants to commit to playing Bottom, which would result in a payoff 3 for both agents if he can commit unilaterally. Consequently, the only commitment regime that can change the outcome of this game is a regime in which agent 1 can commit without the agreement of agent 2. Although this commitment does not correspond with Castelfranchi's (1995) description of a social commitment, because the

[11] Randomization indicates that an agent chooses with some probability *p* one action and with probability *1-p* the other action using some kind of randomization device, e.g., flipping a coin. Expected outcomes can be calculated if we know the probabilities and assume cardinal payoffs for a moment. If both agent use a probability ½ the expected payoff for both agents is (1+2+3+4) / 4 = 2.5. It is easy to show that the expected payoff should lay between 2 and 3 in the general case.

commitment does not lead to the goal of agent 2, such regime might be considered socially desirable because it ensures a more equal distribution of the payoffs among the agents.

Example (5) represents a group of eight games, in which both agents agree that one agent should commit. Without commitment agent 1 obtains 2, while agent 2 obtains 3. However, if agent 1 commits to Bottom, they receive 3 and 4, respectively. Agent 2 cannot commit to Right if he is not sure that agent 1 commits to Bottom, since that could still lead to the worst outcome for him. Therefore, unilateral commitment by agent 1 (with or without agreement of agent 2) will lead here to a better outcome for both agents compared to the game without commitments. Clearly, also a bilateral commitment with agreement in which agent 2 commits to Right in combination with the commitment to Bottom of agent 1 will work.

Example (6) represents three games, which are also called 'coordination' games. In these games, there are more equilibria. One equilibrium involves randomization between the actions by both agents. This leads to payoffs between 2 and 3 for both agents. Therefore, both agents want to coordinate on one of the equilibria without randomization, i.e., Top and Left, or Bottom and Right. However, without a commitment they do not have a clue about the other agent's choice. Coordination on one of the preferred equilibria is possible if one of the agents can commit unilaterally. Agreement of the other agent is not necessary, but he will agree because he is worse off in the situation without commitment. The agent who unilaterally commits first is best off because he can commit such that he will obtain 4 and the other agent 3. This is sometimes called a first-mover advantage. Bilateral commitments (with simultaneous decisions on committing to one of the actions) are problematic here, since this requires solving the same coordination problem while choosing on commitments rather than on actions.

Example (7), representing a group of only two games, looks very much the same as example (6). The only difference is that agent 1 prefers to play the game without a commitment, rather than that agent 2 commits to playing Left, while this is the best solution for agent 2. On the other hand, both agents prefer to play the game while agent 1 commits to playing Bottom over playing the game without a commitment. Consequently, a unilateral regime without agreement will work only if agent 1 can commit. However, using a unilateral commitment in which agent 2 can commit, the agreement of agent 1 is necessary to reach the outcome (4,3). Also using a bilateral commitment with agreement, the agents will reach this outcome, because agent 2 realizes that agent 1 will not agree with agent 2 committing to Left.

Finally, example (8) is a unique example in which different commitment regimes lead to three different solutions. If the agents can commit unilaterally without agreement, agent 1 commits to playing Bottom, while agent 2 commits to playing Right. The one who is allowed to commit obtains 4, while the other who has to follow obtains 2. The other agent will never agree upon a unilateral commitment with agreement, because the expected payoffs for both agents without commitment lie again between 2 and 3 using randomization. However, if they both can commit conditional on whether the other agent commits to their preferred action, they will agree on committing to play Top and Left, both obtaining 3, which is better than playing without a commitment.

Now we will indicate how we can use this game-theoretic analysis in formalizing logical systems and show that specifying the incentive structure that lies beneath a certain formal specification including the related commitment regime can increase the usefulness of logical systems. As long as the incentive structure beneath the interaction between two agents is unknown, it is unclear whether the specified commitment regime will work and will be efficient. In the following section, we will discuss some ideas about how these game-theoretic results can be integrated in logical formalizations.

Commitments in a Logical System

Existing formalizations

The formalization of the notion of commitment is a topic of continuing interest in AI for several reasons. In organization theories of Distributed Artificial Intelligence (DAI), negotiation systems, and cooperative software agents, it is emphasized that 'commitment' is a basic ingredient to analyze a collective activity or the structure of the organization (cf. Gasser, 1991). In Belief-Desire-Intention systems (BDI), important contributions have been made on *motivational attitudes* such as commitments and obligations to specify, analyze, and reason about the behavior of rational agents. BDI-agents are characterized by a 'mental state' described in terms of *beliefs* (viewed as informational attitudes), corresponding to the information the agent has about the environment; *desires* (viewed as its goals), representing options available to the agent; and *intentions* (viewed as motivational attitudes) representing the chosen options. Consequently, by formalizing commitment in a logical way or to propose a descriptive ontology for commitment (and other motivational attitudes), it is possible (1) to reason about commitments to achieve tasks; (2) to gain some insights in the fundamental notions of motivational attitudes; and (3) to analyze collective activity.

Most formal approaches formalizing commitments focus on *internal* commitments (e.g., Cohen & Levesque, 1990; Dignum, Meyer, Wieringa & Kuiper, 1996; Meyer et al., 1999; Rao & Georgeff, 1991): a relation between an agent and a task. An agent who is committed to a task has promised himself to achieve the task.[12] We assume that to commit to an action necessarily implies committing to some result of that action. Conversely, to commit to a goal always implies the commitment of at least one action that produces such a goal as result. Whether or not this goal is reached obviously depends as well on the action of the other agent. Thus, we consider the action/goal pair $\tau = (\alpha, g)$ as the real object of commitment, which we call 'task'. By means of τ, we will refer to the action α, to its intended goal g, or to both (cf. Castelfranchi & Falcone, 1998). Rao and Georgeff (1991) consider beliefs, goals and intentions to be primitive, and define a notion of internal commitment in terms of these by treating intentions as a commitment to the achievement of current tasks.[13] However, these notions (and their semantics) are very fruitful as a basis for the formalization of social motivational attitudes. Rao and Georgeff provide the following two axioms to capture the interrelationships among an agent's beliefs, goals, and intentions:

1. The axiom of belief-goal compatibility:

   $$\mathrm{GOAL}(i, i{:}\tau) \rightarrow \mathrm{BEL}(i, i{:}\tau),$$

   which states that if an agent has a goal, he also believes it. The formula $(i{:}\tau)$ stands for the proposition that agent i achieves τ (or that agent i sees to it that τ will be accomplished).

2. The axiom of goal-intention compatibility:

   $$\mathrm{INT}(i, i{:}\tau) \rightarrow \mathrm{GOAL}(i, i{:}\tau),$$

   which states that if an agent intends to achieve a task, he also has the goal to achieve that task.

The two above-mentioned formulas imply that if an agent intends to achieve task τ, he also believes it. Based on the above-mentioned notions, Dunin-Keplicz and Verbrugge (1999, 2001) give the following definition of (social) commitment inspired by Castelfranchi (1995):

$$\mathrm{COMM}(i, j, i{:}\tau) := \mathrm{INT}(i, i{:}\tau) \wedge \mathrm{GOAL}(j, i{:}\tau) \wedge \text{C-BEL}_{\{i,j\}}(\mathrm{INT}(i, i{:}\tau) \wedge \mathrm{GOAL}(j, i{:}\tau)).$$

If agent i is committed to agent j to achieve something, then i should have the intention to achieve that and j is interested in i fulfilling i 's intention. This condition can be seen as a goal adoption: the achievement of the task is a goal of j. Since we restrict ourselves to rational agents, we state that an agent j agrees with the commitment entered by agent i to achieve some task

[12] An agent is subject to an internal commitment if and only if she is the sole author of a commitment, and has the authority unilaterally to rescind it (Gilbert, 1999).

[13] For formal semantics of these primitive notions we refer to Rao and Georgeff (1991).

if and only if the achievement of the task by *i* is a goal of *j*.[14] So goal adoption implies agreement with or acceptance of the commitment.[15] Commitments require that an agent *j* to whom an agent *i* is committed is aware of *i* 's intention. The collective belief operator indicates that the agents have mutual knowledge / belief about the intention of *i* and the goal of *j*. In daily life, this is done by expressions in conditions of common knowledge (whether or not it is 'out in the open') as far as the two agents are concerned (cf. Gilbert, 1992), and in a business transaction, e.g., by a contract.

The definition of commitment gives rise to some remarks. According to Meyer et al. (1999), it is doubtful whether notions as goals and intentions are primitive, since motivational processes are stemming from internal drives, and are experienced by humans as *conscious desires*. *Wishes* (or *desires*) constitute the primitive motivational attitude that models what an agent likes to be the case, and therefore they take wishes as primitive, and define goals by means of these. Related to this, we suggest the following formalization for a goal in the definition of commitment as presented above:

$$\mathrm{GOAL}(j, i{:}\tau) := \mathrm{WISH}(j, i{:}\tau) \wedge \neg(i{:}\tau) \wedge \Diamond(i{:}\tau) \wedge \mathrm{SELECT}(j, i{:}\tau),$$

meaning that a goal of *j* is defined a selected wish of *j* ($\mathrm{WISH}(j, i{:}\tau) \wedge \mathrm{SELECT}(j, i{:}\tau)$) that is unfulfilled ($\neg(i{:}\tau)$), and can be implemented by *i* ($\Diamond(i{:}\tau)$): "it does not make sense for an agent to try and fulfill a wish that already has been fulfilled or for which fulfillment is not a practical possibility for that agent" (Meyer et al. 1999, p. 13). With the help of the notion of goal, they define (possible) intentions to achieve all the tasks that are correct and feasible with respect to some of their goals:

$$\mathrm{INT}(i, i{:}\tau) := \mathrm{CAN}(i, i{:}\tau) \wedge \mathrm{K}(i, \mathrm{GOAL}(i, i{:}\tau)),$$

where $\mathrm{CAN}(i, i{:}\tau)$ states that for agent *i* it is a practical possibility to achieve τ and K is the Kripke knowledge operator. In the sequel, we use for convenience the terms goal and intention as primitives to formalize commitments.

According to the formal definition of commitment, which refers to a relation between two agents and a task: the commitment of one agent to another, it is necessary that the agent who does not commit has the goal that the other agent achieves the intended task. However, a unilateral

[14] In the game-theoretic sense, this implies that the payoff for *j* is larger if *i* performs the action specified in the commitment rather than the other action.

[15] Fasli (2001) defines commitment replacing the goal adoption ($\mathrm{GOAL}(j, i{:}\tau)$) by a relativised obligation ($\mathrm{O}(i, j, i{:}\tau)$) implying that *i* has an obligation toward *j* to achieve task τ. Because we want to distinguish between commitments with and without agreement, the definition using $\mathrm{GOAL}(j, i{:}\tau)$ is closer related to our focus.

commitment *without* agreement does not need to lead to the goal state of the agent who does not commit, as we have seen in the crossing-the-road example. Consequently, the definition of commitment is too strong for using it *a priori* in all situations. Actually, the definition provided above only formalizes the unilateral commitment with agreement. As the game-theoretical analysis has shown, this commitment regime is relevant only in a limited number of possible interactions. In the following section, we adapt Dunin-Keplicz and Verbrugge's definition for the other three commitment regimes as well.

Further Formalizations Based on the Game-Theoretic Analysis

From the game-theoretic analysis follows that the four commitment regimes mentioned before represent crucial distinctions and cause different outcomes in various interactions among agents. Now, we will provide logical formalizations for these regimes.

- Unilateral commitment without agreement:

$$\text{COMM}_{1w}(i, j, i{:}\tau) := \text{INT}(i, i{:}\tau) \wedge \text{C-BEL}_{\{i,j\}}(\text{INT}(i, i{:}\tau)).$$

  In this commitment regime, agent *j* cannot withhold this commitment. The definition excludes that there has to be an *agreement* between the agents about whether or not the commitment can be made. In the crossing-the-road example, we have shown such a commitment. Since the agreement is missing, there is no requirement that the commitment contributes to a goal of the agent who is not committing. This implies that this commitment becomes effective independent of whether or not agent 2 prefers this commitment.

- Unilateral commitment with agreement:

$$\text{COMM}_{1a}(i, j, i{:}\tau) := \text{INT}(i, i{:}\tau) \wedge \text{GOAL}(j, i{:}\tau) \wedge \text{C-BEL}_{\{i,j\}}(\text{INT}(i, i{:}\tau) \wedge \text{GOAL}(j, i{:}\tau)).$$

  In this commitment regime, agent *j* has to agree upon the commitment of agent *i*. The commitment is only effective under the condition that agent *j* agrees upon the commitment. This requires (in the game-theoretic sense) that the resulting payoff for agent *j*, if this commitment is effective, is larger than the expected payoff without a commitment. In other words, the commitment has to lead to a goal of agent *j*. Unilateral commitment with agreement is a subclass of the unilateral commitment without agreement. We do not consider this as a problem. Distinguishing the unilateral commitment without agreement is done to indicate that, on the one hand, there might be situations in which one agent can force an outcome on the other agent (see examples (4) and (8) in figure 4), and, on the other hand, that some problems such as coordination problems can be solved without explicitly requiring that the other agent agrees with the commitment (see examples (5), (6), and (7) in figure 4). The two commitment regimes can be made logically exclusive by requiring that the intention of the committing agent should *not* be a goal of the second agent for unilateral commitment without agreement. However, it seems substantively nonsensical to define a commitment regime that requires one agent to commit to an action related to an outcome that is explicitly not the goal of the other agent.

- Bilateral commitment without agreement:

$$\mathrm{COMM}_{2\mathrm{w}}(i, j, i{:}\tau_1, j{:}\tau_2) := \mathrm{COMM}_{1\mathrm{w}}(i, j, i{:}\tau_1) \wedge \mathrm{COMM}_{1\mathrm{w}}(j, i, j{:}\tau_2).$$

This commitment is composed of the two unilateral commitments without agreement. For example, an agent *i* commits to agent *j* to make dinner, and agent *j* commits to agent *i* to do the laundry (agent *j* makes a 'counter commitment'). If agent *i* does not commit, this does not affect the status of the *j* 's commitment as a standing commitment (Gilbert, 1999), which differs for the bilateral commitment with agreement.

- Bilateral commitment with agreement:

$$\mathrm{COMM}_{2\mathrm{a}}(i, j, i{:}\tau_1, j{:}\tau_2) := \mathrm{COMM}_{1\mathrm{a}}(i, j, i{:}\tau_1) \wedge \mathrm{COMM}_{1\mathrm{a}}(j, i, j{:}\tau_2)$$

meaning that 'on the condition that agent *j* commits to agent *i* to achieve τ_2, *i* commits to *j* to achieve τ_1.' Consequently, if *j* does not commit, this affects the status of *i* 's commitment: *i* 's commitment stands no longer. Due to the symmetry in the formulation, it also holds that agent *j* 's commitment is only effective under the condition that agent *i* made the related commitment.[16] Clearly, this commitment can only be realized if the agents have the same goal.

We will now reconsider the eight examples of figure 4 to use them for specifying logical formalizations based on the game-theoretic analysis. For examples (1) and (2), it is impossible to formalize a commitment that affects the behavior of the agents. Any commitment the agents want to make leads to the same behavior as they would execute if there was no commitment.

In example (3) the social efficient outcome (3,3) will be reached by a bilateral commitment with agreement: agents 1 and 2 have to commit to Top and Left, respectively. So to reach the socially efficient outcome (3,3), the commitment regime should be formalized as:

[16] Although only assuming a one-sided condition of *i* 's commitment on *j* 's commitment might seem to be a 'weaker' form of commitment, the fact that the commitments will only be effective for one specified combination of commitments ensures that this symmetric formulation is equivalent to a one-sided conditional formulation for the types of games considered in this article.

$$\mathrm{COMM}_{2a}(1, 2, 1{:}\mathrm{Top}, 2{:}\mathrm{Left}).$$

Example (4) shows that both agents do not have the same goal state. The outcome (3,3) is the goal state of agent 1 while (2,4) is the goal state of agent 2. Moreover, without commitment the outcome will be (2,4). Consequently, agent 1 wants to commit to play Bottom. Because this is not the goal state of agent 2, this can only be reached if agent 1 can commit unilaterally without the need of agreement. This can be formalized as:

$$\mathrm{COMM}_{1w}(1, 2, 1{:}\mathrm{Bottom}).$$

The definition of unilateral commitment with agreement is a suitable formalization for a commitment that leads to a socially efficient outcome in example (5). Without a commitment the outcome would be (2,3). However, if agent 1 commits to Bottom, the outcome will be (3,4). Clearly, this commitment leads to a goal of agent 2:

$$\mathrm{COMM}_{1a}(1, 2, 1{:}\mathrm{Bottom}).$$

Also in example (6), the definition of the unilateral commitment with agreement is a suitable formalization for a commitment that leads to a socially efficient outcome. There is a complication because both agents might commit, but they should not commit simultaneously (thus not by a bilateral commitment without agreement). Therefore, a suitable commitment regime should prescribe which agent is allowed to commit. Both agents want to commit because the committed agent receives 4, while the other agent receives 3. The regime can be formalized by the convention:

$$\mathrm{COMM}_{1a}(1, 2, 1{:}\mathrm{Bottom}) \vee \mathrm{COMM}_{1a}(2, 1, 2{:}\mathrm{Left}) \wedge \neg\mathrm{COMM}_{2w}(1, 2, 1{:}\mathrm{Bottom}, 2{:}\mathrm{Left}).$$

In example (7), a commitment of one agent is again necessary to coordinate on one of the equilibria. The best option for agent 2 is to commit unilaterally to Left: $\mathrm{COMM}_{1w}(2, 1, 2{:}\mathrm{Left})$. The best option for agent 1 is to commit unilaterally to Bottom: $\mathrm{COMM}_{1w}(1, 2, 1{:}\mathrm{Bottom})$. However, if the commitment needs to be agreed upon by the other agent, agent 1 will not accept the commitment of agent 2, because he is better off without a commitment. On the contrary, agent 2 will accept agent 1’s commitment because the related outcome is still better than playing without a commitment. This analysis suggests that $\mathrm{COMM}_{1a}(1, 2, 1{:}\mathrm{Bottom})$ is the preferred formalization of a commitment in this situation.

Again both agents would want to commit unilaterally (without agreement) in example (8), which is the reason that this example differs from example (3). However, each agent prefers to play the game without a commitment rather than with a unilateral commitment of the other agent. The socially efficient outcome (3,3) can only be reached with a bilateral commitment with agreement.

What we learn from this classification of simple 2 × 2 games is that the formal definition of commitment provided at the beginning of this section leaves too many essential dimensions of a commitment unspecified. The decision of an agent who has a possibility to commit might depend on whether or not the other agent has to agree with the commitment. It might be crucial whether one or both agents have an option to commit to an action and in which order the agents obtain the opportunity to commit. In game-theoretic terms, these options can be formalized by adding moves to the game that implement the possibilities for the agents to commit and, eventually, to accept the commitment of the other agent. Using game-theoretic reasoning, solutions of these extended games can be calculated, which provides predictions about whether or not commitments will be used and what the consequences of these commitments are depending on the chosen commitment regime. As a result, insights are obtained about whether a commitment regime is socially efficient or favors one of the two agents.

Conclusion

In this article, we have shown that game-theoretic reasoning provides new insight with respect to commitments for logical systems. Specifically, our analysis demonstrates that the distinction between unilateral and bilateral commitments has been obscured in existing logical formalizations, but this distinction is important for the effectiveness of commitments in certain situations. The possibility of making a commitment without the necessity of the other agent's agreement is also neglected in existing logical formalizations. To gain insight into the different types of commitment we introduced different commitment regimes. The game-theoretic analysis showed that the different regimes could lead to different outcomes in interactions between two agents. Consequently, the usefulness of logical systems formalizing commitment would be increased by specifying the incentive structure that lies beneath a certain formal specification and adapting the commitment regime such that an underlying 'social goal' (social efficiency or equal distribution) can be realized. As long as the incentive structure beneath the interaction between two agents is unknown (i.e., without explicit reasoning why and when a certain motivational attitude leads to a commitment), it is unclear whether the specified commitment regime will work and will lead to an efficient outcome. We have shown that the four regimes can smoothly be formalized in a logical framework based on existing primitive notions as intentions and goals, in which we can distinguish whether a commitment is credible or not in a given interaction.

So far, we have limited the game-theoretic analysis to a small sample of games. In further research, the analysis can be extended to 2 × 2 games without strictly ordered payoffs, or groups of agents who have to accomplish a task. This could lead to extensions related to more general notions such as a collective commitment. Such a commitment is defined as the internal commitment of a group (a collective agent) given a selected plan (in the line of the work of Bratman, Israel & Pollack, 1988) and the underlying structure of the group with respect to the achievement of the collective task (cf. Grosz & Kraus 1996; Dunin-Keplicz & Verbrugge, 1999, 2001; Wooldridge & Jennings, 1999; Royakkers & Dignum, 2000). All these papers neglect the incentives of the different agents to perform certain actions and, e.g., possibilities for conflicting interests that follow from these incentives. Consequently, all these formalizations might be extended with explicit reasoning about the incentive structure of the problems under research, which probably will lead to a richer set of formalizations of commitments.

An extension to games in which agents perform sequentially including actions for the commitment decisions seems interesting, which gives the possibility to analyze one-sided conditional commitments. Sandholm and Lesser (2001) provide detailed game-theoretic analyses for such extensions. Their ‘leveled’ commitments are also not necessarily binding, and actions of agents are not necessarily observable. Broersen, Dastani and Van der Torre (2000) started logical formalizations of such commitments using dynamic deontic logic. In this context, more normative issues can be considered, e.g., what happens if an agent drops his commitment? Whether the committed agent has to fulfill the ‘obligation’ implied in the commitment will depend on normative restrictions defined in the formalization. These restrictions can also be included in the game-theoretic analysis by making the sanction for deviating from the commitment dependent on the norms present in a social context. In a context in which norms are very strong, the sanctions will be larger than in a context in which norms are less strong.

Finally, we want to comment on the rationality assumption that we make in this article. Researchers have been arguing that game-theoretic models are “idealizations of the way agents would operate, and ignore the practicalities of *computing* an appropriate action to perform” (Wooldridge & Jennings, 1999, p. 566). Also Castelfranchi and Conte (1998) indicate a number of limitations of a game-theoretic approach, although they indicate several merits of game theory as well. We want to stress that for the point we make in this article, we only need game theory as a analytic tool to distinguish between some typical social situations formulated in terms of games. We only assume that agents in social situations perceive these situations as indicated in these games, and that the agents are able to behave sensible given the structure of these situations. This implies that the agents assign some utility to the different outcomes resulting from a pair of actions for themselves and for the other agent. We do not make any assumption about where the utility of the actors depends on or how it is formed. Just on this basis, we have shown that distinguishing the different commitment regimes developed in this article is worthwhile.

References

Bratman, M. E., Israel, D., & Pollack, M. E. (1988). Plans and resource-bounded practical reasoning. *Computational Intelligence, 4,* 349-355.

Broersen, J., Dastani, M., & Van der Torre, L. (2000). Leveled Commitment and Trust in Negotiation. In *Proceedings of the Autonomous Agents 2000 Workshop on Deception, Fraud and Trust in Agent Societies*, Barcelona, 2000.

Castelfranchi, C. (1995). Commitments: From individual intentions to groups and organizations. In V. R. Lesser (Ed.), *Proceedings First International Conference on Multi-Agent Systems* (pp. 41-48). San Francisco, CA: AAAI-Press and MIT Press.

Castelfranchi, C., & Conte, R. (1998). Limits of economic and strategic rationality for agents and MA systems. *Robotics and Autonomous Systems, 24,* 127-139.

Castelfranchi, C. & Falcone, R. (1998). Towards a theory of delegation for agent-based systems. *Robotics and Autonomous Systems, 24,* 141-157.

Cohen, P. R, & Levesque, H. J. (1990). Intention is choice with commitment. *Artificial Intelligence, 42,* 213-261.

Conte, R., & Castelfranchi, C. (1995). *Cognitive and social action*. London: UCL Press.

Dignum, F., Meyer, J.-J. Ch., Wieringa, R. J., & Kuiper, R. (1996). A modal approach to intentions, commitments and obligations: Intentions plus commitment yields obligations. In: M.A. Brown & J. Carmo (Eds.), *Deontic logic, agency and normative systems* (pp. 80-97). Berlin: Springer.

Dunin-Keplicz, B., & Verbrugge, R. (1996). Collective commitments. In M. Tokora (Ed.), *Proceedings Second International Conference on Multi-Agent Systems* (pp. 56-63). San Francisco, CA: AAAI-Press.

Dunin-Keplicz, B., & Verbrugge, R. (1999). Collective motivational attitudes in cooperative problem solving. In V. Gorodetsky (Ed.), *Proceedings of the First International Workshop of Central and Eastern Europe of Multi-Agent Systems (CEEMAS '99)* (pp. 22-41). St. Petersburg.

Dunin-Keplicz, B., & Verbrugge, R. (2001). The role of dialogue in collective problem solving. In E. Davis, J. McCarthy, L. Morgenstern & R. Reiter (Eds.), *Proceedings of the Fifth International Symposium on the Logical Formulations of Commonsense Reasoning (Commonsense 2001)* (pp. 89-104). New York.

Fasli, M. (2001). On obligations, relativised obligations, and bilateral commitments. In E. Stroulia & S. Matwin (Eds.), *Proceedings of the Canadian Artificial Intelligence Conference* (pp. 287-296). Berlin: Springer-Verlag.

Gasser, L. (1991). Social conceptions of knowledge and action: DAI foundations and open systems semantics. *Artificial Intelligence, 47,* 107-138.

Gilbert, M. (1992). *On social facts.* Princeton, NJ: Princeton University Press.

Gilbert, M. (1999). Considerations on obligation. *Utilitas, 11,* 143-163.

Grosz, B. J., & Kraus, S. (1996). Collaborative plans for complex group action. *Artificial Intelligence, 86*, 269-357.

Luce, R.D., & Raiffa, H. (1957). *Games and Decisions*. New York: Wiley.

Meyer, J.-J. Ch., Van der Hoek, W., & Van Linder, B. (1999). A logical approach to the dynamics of commitments. *Artificial Intelligence, 113*, 1-40.

Rao, A., & Georgeff, M. (1991). Modelling rational agents within a BDI-architecture. In R. Fikes & E. Sandewall (Eds.), *Proceedings of the Second Conference on Knowledge Representation and Reasoning* (pp.473-484). San Mateo, CA: Morgan Kaufman.

Rapoport, A., Guyer, M. J., & Gordon, D. G. (1976). *The 2 ´ 2 game*. Ann Arbor, MA: University of Michigan Press.

Rasmusen, E. (1994). *Games and information: An introduction to game theory* (2nd). Oxford: Blackwell.

Royakkers, L. M. M., & Dignum, F. (2000). Organizations and collective obligations. In M. Ibrahim, J. Küng & N. Revell (Eds.), *Database and expert systems applications* (pp. 302-311). London: Springer.

Sandholm, T. W., & Lesser, V. R. (2001). Leveled Commitment Contracts and Strategic Breach. *Games and Economic Behavior, 35*, 212-270.

Schelling, T. C. (1960). *The strategy of conflict*. Cambridge MA: Harvard University Press.

Shoham, Y. (1993). Agent-oriented programming. *Artificial Intelligence, 60*, 51-92.

Singh, M. P. (1999). An ontology for commitments in multiagent systems: Toward a unification of normative concepts. *Artificial Intelligence and Law, 7*, 97-113.

Snijders, C. (1996). *Trust and commitments*. Amsterdam: Thesis Publishers.

Snijders, C., & Buskens, V. (2001). How to convince someone that you can be trusted? The role of 'hostages'. *Journal of Mathematical Sociology, 25*, 355-384.

Wooldridge, M., & Jennings, N.R. (1999). The cooperative problem-solving process. *Journal of Logic Computation, 9*, 563-592.